Probability theory

A fast course

Lasse Leskelä
Aalto University

21 October 2024

Contents

Preface

This text is designed for students in mathematics, statistics, computer science, and other disciplines who have taken a first course in probability and basic courses in calculus. The goal is to walk the student to the deep end of probability theory in a fast pace, with minimal prerequisites, but still in a mathematically rigorous manner. Compared to classical textbooks, this text has more emphasis on probability kernels, density functions with respect to Lebesgue and counting measures, and metrics between probability measures; and less emphasis on filtrations of sigma-algebras and conditional expectations defined using sigma-algebras.

For corrections, remarks, and helpful discussions, I would like to thank Kalle Kytölä, Sari Rogovin, Kalle Alaluusua, Eeli Asikainen, Marcell Berta, Ruslan Ershov, Aarni Haapaniemi, Ilari Helander, Eero Härmä, Parag Ingle, Tuomas Juuranto, Konsta Kemppainen, Juho Korkeala, Niko Miller, Vilma Moilanen, Martin Mäkipää, Aaro Niini, Taneli Pääkkö, Verneri Seppänen, Konsta Tiilikainen, and Quan Tran. Your comments have much improved the presentation. The text still under construction and likely to be updated in the coming years. Further corrections and comments are much appreciated.

Espoo, 21 October 2024

Notations

Symbol	Meaning
$\mathbb{Z}$	Integers
$\mathbb{Z}_+$	Nonnegative integers
$\mathbb{N}$	Strictly positive integers
$\bar{\mathbb{Z}}_+$	$\mathbb{Z}_+ \cup \{\infty\}$
$\bar{\mathbb{Z}}$	$\mathbb{Z} \cup \{-\infty\} \cup \{\infty\}$
$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Nonnegative real numbers
$\bar{\mathbb{R}}_+, [0,\infty]$	$\mathbb{R}_+ \cup \{\infty\}$
$\bar{\mathbb{R}}$, $[-\infty,\infty]$	$\mathbb{R} \cup \{-\infty\} \cup \{\infty\}$
$x \in A$	x belongs to A
$\emptyset$	empty set
$A \subset B$	A is a subset of B
A^c	complement of A
$A \cap B$	intersection of A and B
$A \cup B$	union of A and B
$B \setminus A$	set difference $B \cap A^c$
$A \times B$	cartesian product of A and B
$f(A)$	image of A by f, $\{f(x)\colon x \in A\}$
$f^{-1}(A)$	preimage of A by f, $\{x\colon f(x) \in A\}$
$A + h$	$\{a + h\colon a \in A\}$
1_A	indicator function of A
$\{x\}$	singleton set only containing x
$x \wedge y$	minimum of x and y
$x \vee y$	maximum of x and y
$[n]$	$\{1, 2, \ldots, n\}$
$\mathcal{S}$	sigma-algebra on S
$\mathcal{B}(\mathbb{R})$	Borel sigma-algebra on $\mathbb{R}$
$\sigma(\mathcal{C})$	sigma-algebra generated by $\mathcal{C}$
π_i	i-th coordinate function

$\mathcal{S}_1 \otimes \mathcal{S}_2$	product sigma-algebra
$\#$	counting measure
$\lambda, \lambda(dx), dx$	Lebesgue measure
λ_d	d-dimensional Lebesgue measure
δ_x	Dirac measure at x
$\mathrm{Ber}(p)$	Bernoulli distribution with parameter p
$\mathrm{Poi}(a)$	Poisson distribution with parameter a
$\mathrm{Nor}(0,1)$	standard normal distribution
$\mathbb{P}$	probability measure
$\mathbb{E}$	expectation
$\mu(f), \int f\,d\mu, \int f(x)\mu(dx)$	integral of f against μ
$L^1(\mu)$	integrable functions against μ
$L^p(\mu)$	power-integrable functions of order p against μ
$\mathcal{P}(S)$	probability measures on S
$\mathcal{P}_p(\mathbb{R})$	probability measures on $\mathbb{R}$ with finite p-th moments
$\mu \otimes \nu$	product of measures μ and ν
$\mu \otimes K$	product of measure μ and kernel K
μK	pushforward of measure μ by kernel K
d_{tv}	total variation distance
W_p	Wasserstein distance of order p

Chapter 1

Probabilities and measures

> *Die Wahrscheinlichkeitstheorie als mathematische Disziplin soll und kann genau in demselben Sinne axiomatisiert werden wie die Geometrie oder die Algebra.*
>
> —*Andrey Kolmogorov*

The theory of probability is written in the language of measure theory in which *sets* represent events, and *measures* assign numbers to sets corresponding to probabilities of events. Uncountably infinite spaces contain sets for which we cannot assign probabilities in a meaningful manner. This is why we restrict to set families that are small enough to rule out unnecessary pathologies, and large enough to be closed under set operations corresponding to logical connectives of events. Such families are called *sigma-algebras*.

Key concepts: Measure, sigma-algebra, indicator function, monotone set limit, probability mass function, Dirac measure

Learning outcomes:

- Get familiar with abstract sums $\sum_{x \in A} f(x)$ over finite and countably infinite sets.
- Get introduced to working with basic arithmetic operations (sum, product) and analytic concepts (limits, sums of series) on $[0, \infty]$.
- Learn to construct discrete probability measures using probability mass functions and Dirac measures.

Prerequisites: Set union, set intersection, countable set, infinite sum

1.1 Set operations

A *set* S is an unordered collection of objects called the members of S. We denote $x \in S$ if x is a member of S. We denote $A \subset B$ and say that A is a *subset* of B if every member of A is also a member of B. The *empty set* is denoted by $\emptyset$. The *indicator function* of a set A is defined by

$$1_A(x) \;=\; \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases} \tag{1.1}$$

The indicator function 1_A uniquely characterises the set A. Indicator functions provide an important bridge between sets and functions that is constantly used in probability theory.

For subsets A and B of S, we define the set operations *intersection*, *union*, *difference*, and *complement* by

$$\begin{aligned} A \cap B &= \{x \in S \colon x \in A \text{ and } x \in B\}, \\ A \cup B &= \{x \in S \colon x \in A \text{ or } x \in B\}, \\ B \setminus A &= \{x \in S \colon x \in B \text{ and } x \notin A\}, \\ A^c &= \{x \in S \colon x \notin A\}. \end{aligned}$$

Exercise 1.14 helps to understand the interplay between indicator functions and intersections and unions.

A *set family* on S is a set of subsets of a *ground set* S. A set family is often denoted by $\{A_i \colon i \in I\}$ with the understanding that we associate to each $i \in I$ a set $A_i \subset S$. In this case we say that the set family is *indexed* by I. The members of $\{A_i \colon i \in I\}$ are mutually *disjoint* if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

A set family is a set of sets. The members of a disjoint set family do not overlap each other.

A set is *countable* if it can be enumerated as $A = \{x_1, x_2, \dots\}$ using a finite or infinite ordered list $x_1, x_2, \dots$ In the former case the set is called *finite*, and in the latter case *countably infinite*. The intersection and the union of a set family $\{A_i \colon i \in I\}$ are denoted by

$$\begin{aligned} \bigcap_{i \in I} A_i &= \{x \in S \colon x \in A_i \text{ for all } i \in I\}, \\ \bigcup_{i \in I} A_i &= \{x \in S \colon x \in A_i \text{ for some } i \in I\}. \end{aligned}$$

A set family $\{A_i \colon i \in I\}$ is countable if its index set I is countable. We write $\cap_{i\in I} = \cap_{i=1}^{\infty}$ and $\cup_{i\in I} = \cup_{i=1}^{\infty}$ when the index set equals the set of positive integers. The following results (Exercise 1.13) are known as *De Morgan's laws*[1]:

$$\left(\bigcap_{i\in I} A_i \right)^c = \bigcup_{i\in I} A_i^c, \tag{1.2}$$

$$\left(\bigcup_{i\in I} A_i \right)^c = \bigcap_{i\in I} A_i^c. \tag{1.3}$$

De Morgan's laws are useful tools that allow us to switch from sets to their complements when analysing intersection and unions.

1.2 Sigma-algebras

A *sigma-algebra* is a set family $\mathcal{S}$ on S that contains $\emptyset, S$, and is closed under complement, countable union, and countable intersection:

(i) $A \in \mathcal{S} \implies A^c \in \mathcal{S}$.

(ii) $A_1, A_2, \ldots \in \mathcal{S} \implies A_1 \cup A_2 \cup \cdots \in \mathcal{S}$.

(iii) $A_1, A_2, \ldots \in \mathcal{S} \implies A_1 \cap A_2 \cap \cdots \in \mathcal{S}$.

The members of a sigma-algebra are called *measurable sets*. A set equipped with a sigma-algebra is denoted $(S, \mathcal{S})$ and called a *measurable space*. Two fundamental examples of sigma-algebras are given below. Exercise 1.15 provides further examples and non-examples.

Example 1.1. The *trivial sigma-algebra* on S is the set family $\{\emptyset, S\}$. This is the smallest possible sigma-algebra that one can construct on a given set.

Example 1.2. The *discrete sigma-algebra* on S, denoted 2^S, is the set family consisting of all subsets of S. This is the largest possible sigma-algebra that one can construct on a given set. A *discrete measurable space* is a pair $(S, 2^S)$ in which S is a countable set.

[1] Named after a British mathematician Augustus De Morgan (1806–1871).

The discrete sigma-algebra provides a natural framework for a countable space, but is too large to be useful for uncountably infinite spaces.

Set families are commonly encountered in many areas of mathematics. A *topological space* is a set equipped with a family of sets called *open sets*. The set family of a topological space is closed under *finite* intersections and *arbitrary* unions. A *hypergraph* is a (usually finite) set equipped with a structure-free family of sets called *hyperedges*. Table 1.1 summarises these concepts.

$(S, \mathcal{S})$	Measurable space	Topological space	Hypergraph
S called	Ground set	Space	Node set
$\mathcal{S}$ called	Sigma-algebra	Topology	Hyperedge set
$x \in S$ called	Point, outcome	Point	Node, vertex
$A \in \mathcal{S}$ called	Measurable set	Open set	Hyperedge
$\mathcal{S}$ contains $\emptyset$	Yes	Yes	No
$\mathcal{S}$ contains S	Yes	Yes	No
$\mathcal{S}$ closed under $(\ )^c$	Yes	No	No
$\mathcal{S}$ closed under $\cap$	Countable	Finite	No
$\mathcal{S}$ closed under $\cup$	Countable	Any	No

Table 1.1: Comparison of set-theoretical structures.

1.3 Measures

A *measure* on a measurable space $(S, \mathcal{S})$ is a map $\mu \colon \mathcal{S} \to [0, \infty]$ which satisfies $\mu(\emptyset) = 0$ and is *countably disjointly additive*[2] in the sense that[3]

$$\mu\Big(\bigcup_{n \geq 1} A_n\Big) = \sum_{n \geq 1} \mu(A_n) \tag{1.4}$$

for any finite or countably infinite list of mutually disjoint sets $A_1, A_2, \dots \in \mathcal{S}$. A measure is *finite* if $\mu(S) < \infty$. A *probability measure* is a measure with $\mu(S) = 1$. When the sigma-algebra is assumed clear from the context, we say that μ is a measure on S.

[2] aka *sigma-additive*

[3] On the right side of (1.4) we work with the extended half line with the conventions that $x + \infty = \infty$ and $0 \cdot x = 0$ for all $x \in [0, \infty]$. Appendix A provides more details.

Measures are functions that map sets into $[0, \infty]$. A measure can be considered as a method to quantify the size of a set so that the size of the empty set is zero, and the sizes of disjoint sets add up.

Example 1.3. The *Dirac measure* at a point $x \in S$ in a measurable space $(S, \mathcal{S})$ is a function $\delta_x \colon \mathcal{S} \to [0, \infty]$ defined by

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

To see that δ_x is countably disjointly additive, assume that $A_1, A_2, \dots \in \mathcal{S}$ are mutually disjoint and consider the following cases:

(i) If $\delta_x(\cup_{n \ge 1} A_n) = 0$, then $x \notin \cup_{n \ge 1} A_n$, which means that x does not belong to any of the sets A_n. We find that $\delta_x(A_n) = 0$ for all n. Therefore, $\sum_{n \ge 1} \delta_x(A_n) = 0$, and (1.4) is valid.

(ii) If $\delta_x(\cup_{n \ge 1} A_n) = 1$, then $x \in \cup_{n \ge 1} A_n$ implies that $x \in A_k$ for some k. Because the sets $A_1, A_2, \dots$ are disjoint, we also see that $x \notin A_n$ for all $n \ne k$. Therefore, $\delta_x(A_k) = 1$ and $\delta_x(A_n) = 0$ for all $n \ne k$. Hence $\sum_{n \ge 1} \delta_x(A_n) = 1$, and (1.4) is again valid.

We have thus seen that δ_x is a countably disjointly additive set function. Because $\delta_x(\emptyset) = 0$, it follows that δ_x is a measure. Because $\delta_x(S) = 1$, we conclude that δ_x is a probability measure.

The Dirac measure is a simple concept when viewed as a set function, instead of trying to represent it as an ordinary function. Dirac measures serve as important building blocks in constructing other probability measures.

There are lots of further examples of probability measures—in fact *all* probability distributions are probability measures. Examples of infinite measures include the counting measure on the integer lattice $\mathbb{Z}$, and the Lebesgue measure on the real line $\mathbb{R}$. We will soon properly define these. In the meantime, Exercise 1.16 provides further examples and non-examples.

1.4 Monotone continuity of measures

The countable disjoint additivity, included as a defining feature of a measure, encodes important monotonicity and continuity properties that we will state and prove in this section.

Monotonicity corresponds to our intuition about the measure of a set describing its size: any set that is fully contained inside another set must have a smaller size than the other.

Proposition 1.4 (Monotonicity)**.** *Every measure μ is monotone in the sense that for all measurable sets: $A \subset B \implies \mu(A) \le \mu(B)$.*

Proof. Fix a measure μ on a measurable space $(S, \mathcal{S})$. Fix $A, B \in \mathcal{S}$ such that $A \subset B$. We may express $B = A \cup (B \setminus A)$ as a union of disjoint sets, where by writing $B \setminus A = B \cap A^c$ we see that also the latter set belongs to the sigma-algebra $\mathcal{S}$. The disjoint additivity (1.4) and nonnegativity of μ then imply that

$$\mu(B) \;=\; \mu(A \cup (B \setminus A)) \;=\; \mu(A) + \mu(B \setminus A) \;\ge\; \mu(A).$$

□

Measures also admit important continuity properties with respect to *monotone set limits* indicated by:

(i) $A_n \uparrow A$ if $A_1 \subset A_2 \subset \cdots$ and $\cup_{n \ge 1} A_n = A$.

(ii) $A_n \downarrow A$ if $A_1 \supset A_2 \supset \cdots$ and $\cap_{n \ge 1} A_n = A$.

We use similar notations for monotone limits in $[0, \infty]$, so that $x_n \uparrow x$ means that $x_1 \le x_2 \le \cdots$ and $\lim_{n \to \infty} x_n = x$. The following important result summarises the monotone continuity of measures. Exercise 1.17 helps to develop intuition on set limits.

Set limits share some features in common with number limits, but are quite different in certain aspects. For example, no ϵ nor δ are needed to define a set limit.

Proposition 1.5 (Monotone continuity)**.** *Every measure μ is continuous under monotone set limits in the sense that for any measurable sets:*

(i) $A_n \uparrow A \implies \mu(A_n) \uparrow \mu(A)$.

(ii) $A_n \downarrow A$ *and* $\mu(A_1) < \infty \implies \mu(A_n) \downarrow \mu(A)$.

Proof. (i) Consider measurable sets such that $A_n \uparrow A$. In analogy with the geological cross section of the Earth, we shall regard the set A_n as a body composed of layers defined by $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus A_2$, and so on. Then we see that $A_n = B_1 \cup \cdots \cup B_n$ where the sets $B_1, \dots, B_n \in \mathcal{S}$ are mutually disjoint and belong to $\mathcal{S}$. Disjoint additivity (1.4) then implies that

$$\mu(A_n) \;=\; \sum_{k=1}^{n} \mu(B_k).$$

The summands on the right side above are nonnegative, and therefore the right side converges monotonically to the infinite sum $\sum_{k=1}^{\infty} \mu(B_k)$ as $n \to \infty$. We conclude that

$$\mu(A_n) \;\uparrow\; \sum_{k=1}^{\infty} \mu(B_k). \tag{1.5}$$

Next, a moment's reflection reveals that $\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} B_k$. Then by applying (1.4) to the infinite union of disjoint sets $B_1, B_2, \dots$, we find that

$$\mu(A) \;=\; \mu\Big(\bigcup_{k=1}^{\infty} A_k\Big) \;=\; \mu\Big(\bigcup_{k=1}^{\infty} B_k\Big) \;=\; \sum_{k=1}^{\infty} \mu(B_k).$$

Claim (i) follows by combining the above equality with (1.5).

(ii) Consider measurable sets $A_n \downarrow A$ such that $\mu(A_1) < \infty$. Denote $B_n = A_1 \setminus A_n$ and $B = A_1 \setminus A$. By writing $A_1 = A_n \cup B_n = A \cup B$ and noting that both these unions are disjoint, we see that

$$\mu(A_1) \;=\; \mu(A_n) + \mu(B_n) \;=\; \mu(A) + \mu(B).$$

Because $\mu(A_1) < \infty$, we see that all terms in the above equality are finite, and therefore,

$$\begin{aligned} \mu(A_n) \;&=\; \mu(A_1) - \mu(B_n), \\ \mu(A) \;&=\; \mu(A_1) - \mu(B). \end{aligned}$$

Then we observe that $B_n \uparrow B$, so that $\mu(B_n) \to \mu(B)$ by (i). We conclude that

$$\begin{aligned} \lim_{n\to\infty} \mu(A_n) \;&=\; \lim_{n\to\infty} \Big(\mu(A_1) - \mu(B_n)\Big) \\ &=\; \mu(A_1) - \lim_{n\to\infty} \mu(B_n) \\ &=\; \mu(A_1) - \mu(B) \\ &=\; \mu(A). \end{aligned}$$

Monotonicity (Proposition 1.4) now implies that $\mu(A_n) \downarrow \mu(A)$. □

The following result is also known as the *union bound*.

Proposition 1.6 (Subadditivity)**.** *Every measure μ is countably subadditive in the sense that $\mu(\cup_{n\geq 1}A_n) \leq \sum_{n\geq 1}\mu(A_n)$ for any finite or countably infinite list of measurable sets.*

Proof. Monotonicity (Proposition 1.4) implies that for all measurable sets A and B,

$$\mu(A\cup B) \;=\; \mu(A\cup(B\setminus A)) \;=\; \mu(A)+\mu(B\setminus A) \;\leq\; \mu(A)+\mu(B).$$

By induction it follows that

$$\mu\Big(\bigcup_{k=1}^n A_k\Big) \;\leq\; \sum_{k=1}^n \mu(A_k). \tag{1.6}$$

Let us extend (1.6) into infinite lists of sets. Denote $C_n = \cup_{k=1}^n A_k$ and $C = \cup_{k=1}^\infty A_k$. Then $C_n \uparrow C$, and monotone continuity (Proposition 1.5) implies that $\mu(C_n)\uparrow\mu(C)$. On other hand, (1.6) implies that $\mu(C_n)\leq c$ for all n, where $c = \sum_{k=1}^\infty \mu(A_k)$. □

1.5 Sums of measures

The *sum* of measures μ and ν on $(S,\mathcal{S})$ is a set function $\mu+\nu$ defined by the formula

$$(\mu+\nu)(A) \;=\; \mu(A)+\nu(A), \tag{1.7}$$

and the *scalar multiplication* of a measure by a constant $c\in[0,\infty]$ is a set function $c\mu$ defined by

$$(c\mu)(A) \;=\; c\mu(A). \tag{1.8}$$

The following result implies that $\mu+\nu$ and $c\mu$ are measures. Therefore, the space of measures on $(S,\mathcal{S})$ is closed under linear combinations with nonnegative weights. Remarkably, the result is valid equally well for finite and countably infinite linear combinations.

New measures may be constructed from old measures by taking linear combinations. Any nonzero finite measure may be normalised to a probability measure by scalar multiplication.

Proposition 1.7 (Weighted sums of measures are measures)**.** *Let $\mu_1,\mu_2,\dots$ be measures on $(S,\mathcal{S})$, and let $c_1,c_2,\dots\in[0,\infty]$. Then*

(i) $\mu = \sum_{k=1}^\infty c_k\mu_k$ is a measure on $(S,\mathcal{S})$.

(ii) $\mu = \sum_{k=1}^\infty c_k \mu_k$ *is a probability measure if* $\mu_1, \mu_2, \dots$ *are probability measures and* $\sum_{k=1}^\infty c_k = 1$.

Proof. (i) Because $\mu(A) = \sum_{k=1}^\infty c_k \mu_k(A) \in [0,\infty]$ for all $A \in \mathcal{S}$, we see that μ is a set function from $\mathcal{S}$ onto $[0,\infty]$. Let us verify that μ is countably disjointly additive. Fix disjoint measurable sets $A_1, A_2, \dots$ Because μ_k is countably disjointly additive, we know that

$$\mu_k(\cup_{i=1}^\infty A_i) \ = \ \sum_{i=1}^\infty \mu_k(A_i).$$

Therefore,

$$\mu(\cup_{i=1}^\infty A_i) \ = \ \sum_{k=1}^\infty c_k \mu_k(\cup_{i=1}^\infty A_i) \ = \ \sum_{k=1}^\infty \sum_{i=1}^\infty c_k \mu_k(A_i).$$

Because the order of a double sum with terms in $[0,\infty]$ can be always be swapped (Lemma A.11), it follows that

$$\mu(\cup_{i=1}^\infty A_i) \ = \ \sum_{i=1}^\infty \sum_{k=1}^\infty c_k \mu_k(A_i) \ = \ \sum_{i=1}^\infty \mu(A_i).$$

(ii) Assume that $\mu_1, \mu_2, \dots$ are probability measures and $\sum_{k=1}^\infty c_k = 1$. Part (i) then confirms that μ is a measure. Because $\mu_k(S) = 1$ for all k, we find that

$$\mu(S) \ = \ \sum_{k=1}^\infty c_k \mu_k(S) \ = \ \sum_{k=1}^\infty c_k \ = \ 1.$$

Therefore, μ is a probability measure. □

1.6 Discrete probability spaces

Probability measures on countable spaces can be conveniently represented and handled using probability mass functions. A *probability mass function* on S is a function $f \colon S \to [0,1]$ such that $\sum_{x \in S} f(x) = 1$. Here we adopt the *abstract sum notation*

$$\sum_{x \in A} f(x) \ = \ \sum_{k \geq 1} f(x_k)$$

where $A = \{x_1, x_2, \dots\}$ is an arbitrary enumeration of a countable set A using a sequence of distinct elements. This convention is justified by the fact that finite and countably infinite sums of elements in $[0, \infty]$ are insensitive to the order of summation (see Lemma A.9).

The probability mass function of a probability measure P is defined by $f_P(x) = P(\{x\})$. The following result confirms that f_P is a proper probability mass function as defined above, and that all values of P can be computed using f_P.

Probability measures on discrete spaces can be equivalently represented as probability mass functions.

Proposition 1.8. *Let P be a probability measure on $(S, 2^S)$ in which S is countable. Then $f_P(x) = P(\{x\})$ is a probability mass function on S, and*

$$P(A) \ = \ \sum_{x \in A} f_P(x) \qquad \text{for all } A \subset S. \tag{1.9}$$

Proof. Fix a set $A \subset S$. Assume that A is countably infinite (the finite case is easy). We note that $f_P(x) = P(\{x\}) \geq 0$ for all x. Fix an enumeration $A = \{x_1, x_2, \dots\}$. Then the sets $A_i = \{x_i\}$ are disjoint and such that $\cup_{i=1}^\infty A_i = A$. Hence

$$P(A) \ = \ P(\cup_{i=1}^\infty A_i) \ = \ \sum_{i=1}^\infty P(A_i) \ = \ \sum_{i=1}^\infty f_P(x_i) \ = \ \sum_{x \in A} f_P(x).$$

Hence (1.9) holds for all $A \subset S$. By plugging in $A = S$, we find that $\sum_{x \in S} f_P(x) = P(S) = 1$, so we conclude that f_P is a probability mass function. □

Proposition 1.9. *Let f be a probability mass function on a countable set S. Then the formula*

$$P(A) \ = \ \sum_{x \in A} f(x)$$

defines a probability measure on $(S, 2^S)$ with probability mass function $f_P = f$.

Proof. Obviously $P(A) \geq 0$ for all $A \subset S$. Hence all we need to do is to verify that P is countably disjointly additive, which in this case is equivalent to

$$\sum_{x \in \cup_{i=1}^\infty A_i} f(x) \ = \ \sum_{i=1}^\infty \sum_{x \in A_i} f(x). \tag{1.10}$$

Even though (1.10) appears intuitively plausible, verifying it directly is complicated if we try to keep track of enumerations of the (possibly infinite) sets A_i for each i. An alternative, more elegant proof is based on observing that

$$P(A) \;=\; \sum_{x \in A} f(x) \;=\; \sum_{x \in S} 1_A(x) f(x),$$

in which 1_A is the indicator function defined in (1.1). Next, we observe that $1_A(x) = \delta_x(A)$ where δ_x is the Dirac measure at x. Let us now fix an enumeration $S = \{x_1, x_2, \dots\}$. Then

$$P(A) \;=\; \sum_{x \in S} 1_A(x)\, f(x) \;=\; \sum_{x \in S} f(x) \delta_x(A) \;=\; \sum_{k=1}^{\infty} f(x_k)\, \delta_{x_k}(A).$$

From this equation we recognise that $P = \sum_{k=1}^{\infty} c_k \mu_k$ is a sum of Dirac measures $\mu_k = \delta_{x_k}$ weighted by $c_k = f(x_k) \in [0,1]$. Because

$$\sum_{k=1}^{\infty} c_k \;=\; \sum_{k=1}^{\infty} f(x_k) \;=\; \sum_{x \in S} f(x) \;=\; 1,$$

we may conclude with the help of Proposition 1.7 that P is a probability measure.

Finally, we note that $f_P(y) = P(\{y\}) = \sum_{x \in \{y\}} f(x) = f(y)$ for all $y \in S$, so that the probability mass function of P equals f. □

Example 1.10 (Dirac measure)**.** We say in Example 1.3 that the Dirac measure δ_b at point $b \in S$ is a probability measure. Its probability mass function is given by

$$f_{\delta_b}(x) \;=\; \begin{cases} 1, & x = b, \\ 0, & \text{else.} \end{cases}$$

Example 1.11 (Binomial distribution)**.** Fix an integer $n \geq 1$ and a number $p \in [0,1]$. Define a function on $\mathbb{Z}_+$ by

$$f(k) \;=\; \begin{cases} \binom{n}{k}(1-p)^{n-k}p^k, & k \in \{0,\dots,n\}, \\ 0, & \text{else.} \end{cases}$$

With the help of the binomial sum formula $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$ we find that

$$\sum_{k=0}^{\infty} f(k) \;=\; \sum_{k=0}^{n} \binom{n}{k} (1-p)^{n-k} p^k \;=\; \Big((1-p)+p\Big)^n \;=\; 1.$$

Hence f is a probability mass function that defines a probability measure on $(\mathbb{Z}_+, 2^{\mathbb{Z}_+})$. This probability measure is called the *binomial distribution* with trial count n and success probability p.

Example 1.12 (Poisson distribution)**.** The *Poisson distribution* with mean $\lambda \in (0,\infty)$ is the probability measure on $(\mathbb{Z}_+, 2^{\mathbb{Z}_+})$ with probability mass function

$$f(k) \;=\; e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0,1,\dots$$

By applying the power series representation $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$ of the exponential function, we may check that $\sum_{k=0}^{\infty} f(k) = 1$.

1.7 Exercises

Exercise 1.13 (De Morgan's laws)**.** Prove formulas (1.2)–(1.3).

Exercise 1.14 (Indicator functions)**.** Are the following statements true or false in general for all subsets of a set S? Prove the statement true or find a counterexample.

(a) $1_{A\cap B}(x) = 1_A(x)1_B(x)$ for all $x \in S$.

(b) $1_{A\cup B}(x) = 1_A(x) + 1_B(x)$ for all $x \in S$.

Exercise 1.15 (Sigma-algebras or not)**.** Which of the following set families are sigma-algebras on the set $S = \{1,2,3\}$:

$$\mathcal{S}_1 = \{\emptyset, \{1,2,3\}\}, \qquad \mathcal{S}_2 = \{\{1\},\{2\},\{3\}\}, \qquad \mathcal{S}_3 = \{\emptyset, \{1\}, \{2,3\}\} \ ?$$

Exercise 1.16 (Measures and non-measures)**.** Which of the following set functions are measures on $\mathbb{Z}_+ = \{0,1,2,\dots\}$? Explain your answer rigorously.

$$\begin{aligned}
\mu_1(A) &= 0 \quad \text{for all } A,\\
\mu_2(A) &= \text{number of the elements in } A,\\
\mu_3(A) &= \text{maximum of the elements in } A,\\
\mu_4(A) &= \text{sum of the elements in } A,\\
\mu_5(A) &= \infty \quad \text{for all } A.
\end{aligned}$$

Exercise 1.17 (Monotone set limits)**.** Prove that the following statements are true in general for all subsets of a set S:

(a) $A \subset B$ if and only if $1_A(x) \le 1_B(x)$ for all $x \in S$.

(b) $A_n \uparrow A$ if and only if $1_{A_n}(x) \uparrow 1_A(x)$ for all $x \in S$.

Exercise 1.18 (Impossible and sure events)**.** Let P be a probability measure on a measurable space $(S, \mathcal{S})$.

(a) Consider an infinite list of measurable sets $A_1, A_2, \dots$ for which $P(A_n) = 0$ for all $n \geq 1$. Prove that $P(\cup_{n=1}^{\infty} A_n) = 0$.

(b) Consider an infinite list of measurable sets $A_1, A_2, \dots$ for which $P(A_n) = 1$ for all $n \geq 1$. Prove that $P(\cap_{n=1}^{\infty} A_n) = 1$.

1.8 Historical notes

First notions of mathematical probabilities date back to Gerolamo Cardano's investigations in 1560s on the sum of three dice, and Blaise Pascal's and Pierre de Fermat's studies on gambling (random walks) that inspired Christiaan Huygens to write an article in 1657 that may be considered the first scientific publication in probability theory. Preliminary versions of key theorems in probability theory and mathematical statistics were derived by several authors in the 18th and 19th centuries. However, the rigorous formulation of modern probability theory had to wait until the birth of measure and integration theory.

The foundations of measure theory based on sigma-algebras were developed by Émile Borel in the end of the 19th century, and the foundations of modern integration theory were introduced in Henri Lebesgue's PhD thesis in 1902, supervised by Borel. Maurice Fréchet in 1915 noted that Borel's and Lebesgue's measure and integration theory can be cast in an abstract measurable space. Finally, in 1933 Andrey Kolmogorov published his textbook masterpiece [Kol33] where he condensed the developments of the early 20th century into a general abstract definition of a probability measure as presented in Section 1.3. See [Kal02, SV06] for nice historical accounts.

Chapter 2

Uniform distributions

Les domaines et les ensembles sont aux fonctions ce que les tissus sont aux êtres vivants.

—*Émile Borel*

A uniform probability distribution represents complete randomness in which all outcomes are equally likely. Uniform probability distributions are easy to describe for finite sets, but not so for the continuum. The problem is that the real line contains complicated sets for which it is impossible to define any reasonable notion of size. The solution is to give up trying to assign probabilities to all subsets of the real line. Instead, we shall restrict to a sigma-algebra that is big enough to contain all sets of practical relevance, and small enough to rule out pathological cases that lead to complications. To understand this approach, we need to dig a bit deeper into measure theory and to get introduced to generators of sigma-algebras.

Key concepts: generator of a sigma-algebra, Borel set, counting measure, Lebesgue measure, cumulative distribution function

Learning outcomes:

- Get introduced to working with generators of a sigma-algebra.
- Become familiar with uniform distributions on finite sets and bounded subsets of the real line.
- Learn to recognise sets of Lebesgue measure zero.

Prerequisites: Open and closed subsets of the real line.

2.1 Discrete uniform distribution

The *counting measure* on a measurable space $(S, \mathcal{S})$ is a set function $\#\colon \mathcal{S} \to [0,\infty]$ defined by

$$\#(A) \;=\; \text{number of elements in } A \subset S. \tag{2.1}$$

Proposition 2.1. *The counting measure is a measure.*

Proof. Because $\#(\emptyset) = 0$, we only need to verify that

$$\#\Big(\bigcup_{n\ge1} A_n\Big) \;=\; \sum_{n\ge1} \#(A_n) \tag{2.2}$$

for any finite or countably infinite list of mutually disjoint sets $A_1, A_2, \dots \in \mathcal{S}$. If $\#(A_n) = \infty$ for some n, then (2.2) immediately holds with both sides being infinite.

Let us next consider the case in which the sets $A_1, A_2, \dots \in \mathcal{S}$ are finite. Then the set $S_0 = \bigcup_{n\ge1} A_n$ is countable. The number of elements in any measurable set $A \subset S_0$ can be represented as

$$\#(A) \;=\; \sum_{x\in S_0} \delta_x(A) \;=\; \nu(A),$$

where $\nu = \sum_{x\in S_0} \delta_x$. Because each Dirac measure δ_x is a measure (recall Example 1.3), Proposition 1.7 implies that ν is a measure. In particular, ν countably disjointly additive, and it follows that

$$\#\Big(\bigcup_{n\ge1} A_n\Big) \;=\; \nu\Big(\bigcup_{n\ge1} A_n\Big) \;=\; \sum_{n\ge1} \nu(A_n) \;=\; \sum_{n\ge1} \#(A_n). \qquad \square$$

Let $C \in \mathcal{S}$ be a finite nonempty set in a measurable space $(S, \mathcal{S})$. The *uniform distribution* on C is a set function defined by

$$\mu(A) \;=\; \frac{\#(A\cap C)}{\#(C)}, \qquad A \in \mathcal{S}. \tag{2.3}$$

Proposition 2.2. *The uniform distribution defined by* (2.3) *is a probability measure on* $(S, \mathcal{S})$.

Proof. Exercise 2.10. □

2.2 In search of a good sigma-algebra on $\mathbb{R}$

The real line contains pathological sets for which it is impossible to assign probabilities in a meaningful way. Therefore, the discrete sigma-algebra $2^{\mathbb{R}}$ is not suitable for probability theory. We would like to construct a smaller sigma-algebra on $\mathbb{R}$ that still contains sets of practical relevance. One option is to rule out the problematic sets of $2^{\mathbb{R}}$. This is hard because it is hard to systematically describe all problematic sets. An alternative, economical approach is to start with a minimum set family that needs to be contained in the sigma-algebra, and enlarge it.

In the minimal approach, we want all open sets to be measurable. Let

$$\mathcal{S}_1 \;=\; \{\text{open sets in } \mathbb{R}\}.$$

Is this a sigma-algebra? No, because for example the complement of the open set $(0,1)$ is not open. Any sigma-algebra containing all the open sets must hence be larger than $\mathcal{S}_1$. Let us try to enlarge this. Define

$$\mathcal{S}_2 \;=\; \{\text{open sets in } \mathbb{R}\} \cup \{\text{closed sets in } \mathbb{R}\}.$$

Is $\mathcal{S}_2$ a sigma-algebra? No, because for example the countable union of closed sets $[0,\frac{1}{2}] \cup [0,\frac{2}{3}] \cup [0,\frac{3}{4}] \cup \cdots = [0,1)$ is not open nor closed. We could try to enlarge this by defining

$$\mathcal{S}_3 \;=\; \{\text{countable unions of sets in } \mathcal{S}_2\}.$$

It is possible to check that $\mathcal{S}_3$ is still not a sigma-algebra. We should enlarge this by adding countable intersections of sets in $\mathcal{S}_3$, but unfortunately this type of recursive algorithm would not finish in a finite number of iterations. We will next develop an alternative, indirect approach.

The set family of open sets in $\mathbb{R}$ is not a sigma-algebra. There is no simple direct way to enlarge the set family of open sets into a sigma-algebra.

2.3 Generators of sigma-algebras

The sigma-algebra $\sigma(\mathcal{C})$ *generated* by a set family $\mathcal{C}$ on S is defined as the smallest sigma-algebra on S containing all members of $\mathcal{C}$. More precisely, we define

$$\sigma(\mathcal{C}) \;=\; \bigcap_i \mathcal{F}_i, \tag{2.4}$$

where the intersection on the right is takes over all sigma-algebras on S that contain $\mathcal{C}$. The following results confirms that the right side of (2.4) is a sigma-algebra, so that above definition makes sense. A set family $\mathcal{C}$ is called a *generator* of sigma-algebra $\mathcal{F}$ when $\mathcal{F} = \sigma(\mathcal{C})$.

Proposition 2.3. *The intersection $\cap \mathcal{F}_i$ of sigma-algebras $\mathcal{F}_i$ on S is a sigma-algebra on S.*

Proof. (i) Because each $\mathcal{F}_i$ is a sigma-algebra, we see that $S \in \mathcal{F}_i$ for all i. The latter property means that $S \in \cap_i \mathcal{F}_i$. A similar argument confirms that $\emptyset \in \cap_i \mathcal{F}_i$.

(ii) Assume that $A \in \cap \mathcal{F}_i$. Then $A \in \mathcal{F}_i$ for all i. Because each $\mathcal{F}_i$ is a sigma-algebra, we see that $A^c \in \mathcal{F}_i$ for all i. But this means that $A^c \in \cap \mathcal{F}_i$.

(iii) Assume that $A_1, A_2, \dots \in \cap \mathcal{F}_i$. Then $A_n \in \mathcal{F}_i$ for all i and n. Because each $\mathcal{F}_i$ is sigma-algebra, it follows that $\cup_n A_n \in \mathcal{F}_i$ and $\cap_n A_n \in \mathcal{F}_i$ for all i. We conclude that $\cup_n A_n \in \mathcal{F}$ and $\cap_n A_n \in \mathcal{F}$. □

Any set family $\mathcal{C}$ may be extended to a sigma-algebra $\sigma(\mathcal{C})$. This is how sigma-algebras are usually defined.

Note the analogy with linear algebra.

Any set of vectors $\mathcal{C} = \{v_1, \dots, v_n\}$ may be extended to a vector space $\operatorname{span}(\mathcal{C})$.

2.4 Borel sigma-algebra on the real line

The *Borel sigma-algebra*[1]

$$\mathcal{B}(\mathbb{R}) \;=\; \sigma(\text{open sets in } \mathbb{R})$$

is defined as the smallest sigma-algebra on $\mathbb{R}$ containing all open sets of $\mathbb{R}$. The members of $\mathcal{B}(\mathbb{R})$ are called *Borel sets*. By definition, all open sets are Borel sets. Because $\mathcal{B}(\mathbb{R})$ is a sigma-algebra, and sigma-algebras are by definition closed with respect to taking complements, we see that all closed sets are Borel sets as well. We also see that all members of the set families constructed in Section 2.2 are Borel sets. In addition, $\mathcal{B}(\mathbb{R})$ contains many more sets—in fact essentially all sets of practical interest.

[1] Named after Émile Borel (1871–1956). PhD 1893 @ École Normale Superieure for Jean-Gaston Darboux.

Indeed, it is not easy to write down a subset of $\mathbb{R}$ that is not a Borel set. Neither there exists a simple recipe for constructively describing all Borel sets. Luckily, it is often sufficient to work with a generator of the sigma-algebra, instead of all members. Especially, to work with $\mathcal{B}(\mathbb{R})$, it usually suffices to consider the open sets.

The following important result shows that the Borel sigma-algebra can also be generated by a much simpler set family than the family of all open sets of $\mathbb{R}$. Indeed, most sigma-algebras admit lots of different generators.

Sigma-algebras typically have several generators. For example, the Borel sigma-algebra $\mathcal{B}(\mathbb{R})$ is generated by both the family of open sets, and the family of intervals $(-\infty, x]$.

Note the analogy with linear algebra.

Vector spaces typically have several bases.

Proposition 2.4. *The set family $\mathcal{C} = \{(-\infty, x] \colon x \in \mathbb{R}\}$ is a generator of $\mathcal{B}(\mathbb{R})$.*

Proof. The prove the claim, it suffices to verify that the sigma-algebra $\sigma(\mathcal{C})$ generated by $\mathcal{C}$ satisfes (i) $\sigma(\mathcal{C}) \subset \mathcal{B}(\mathbb{R})$ and (ii) $\sigma(\mathcal{C}) \supset \mathcal{B}(\mathbb{R})$.

(i) To show that $\sigma(\mathcal{C}) \subset \mathcal{B}(\mathbb{R})$, we will first show that $\mathcal{C} \subset \mathcal{B}(\mathbb{R})$. To do this, we note that every interval $(-\infty, x]$ is the complement of an open set (x, ∞), and that $\mathcal{B}(\mathbb{R})$ by definition contains all open sets of $\mathbb{R}$. Hence $\mathcal{C} \subset \mathcal{B}(\mathbb{R})$. We conclude that $\mathcal{B}(\mathbb{R})$ is a sigma-algebra containing the set family $\mathcal{C}$. By definition, $\sigma(\mathcal{C})$ is the smallest sigma-algebra with this property. Therefore, $\sigma(\mathcal{C}) \subset \mathcal{B}(\mathbb{R})$.

(ii) To show that $\sigma(\mathcal{C}) \supset \mathcal{B}(\mathbb{R})$, we will first show that $\sigma(\mathcal{C})$ contains all open sets of $\mathbb{R}$. We proceed step-by-step to show that $\sigma(\mathcal{C})$ contains:

(a) all semi-open intervals of the form $(x, y]$,

(b) all open intervals of the form (x, z),

(c) all open sets $V \subset \mathbb{R}$.

For (a), we note that

$$(x, y] \;=\; (-\infty, y] \setminus (-\infty, x] \;=\; (-\infty, y] \cap (-\infty, x]^c.$$

We see that $(x, y]$ is from members of $\mathcal{C}$ by complements and countable (indeed, finite) intersections. Therefore, $(x, y] \in \sigma(\mathcal{C})$.

For (b), we note that

$$(x, z) = \cup_{n=1}^{\infty} (x, z - 1/n]$$

shows that (x, z) is obtained from a countable union of intervals of type (a). Because intervals of type (a) belong to $\sigma(\mathcal{C})$, so does the interval (x, z).

For (c), we note that every open set V of $\mathbb{R}$ can be written as a countable union of open intervals of type (b). Because intervals of type (b) belong to $\sigma(\mathcal{C})$, so does the set V.

We conclude that $\sigma(\mathcal{C})$ is a sigma-algebra containing the open sets of $\mathbb{R}$. By definition, $\mathcal{B}(\mathbb{R})$ is the smallest such sigma-algebra. Therefore $\sigma(\mathcal{C}) \supset \mathcal{B}(\mathbb{R})$. □

2.5 The uniform measure on the real line

Intuitively, the probability that a uniformly sampled random number is contained in a set A should be not depend on the location or orientation of the set, but only on its size. But how to define the size of a set? It is natural to require that the size of a disjoint union of sets be equal to the sum of the sizes of the constituent sets. This means that the size should be a measure. But how do we define such a measure, and does such a measure exist?

The uniformity of a measure intuitively means that the measure of a set does not change if we shift the set in space. The *shift* of a set $A \subset \mathbb{R}$ by $h \in \mathbb{R}$ is denoted by

$$A + h = \{a + h \colon a \in A\}.$$

A measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called *shift invariant* if $\mu(A + h) = \mu(A)$ for all Borel sets $A \subset \mathbb{R}$ and all $h \in \mathbb{R}$. The following result confirms that there exists a shift-variant measure on the real line which is unique up to a normalisation.

Theorem 2.5. *There exists a unique shift-invariant measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda([0, 1]) = 1$.*

Proof. This famous result was first recognised by Borel in 1895–1898. The proof has two parts: existence and uniqueness.

(i) For the existence, most modern proofs utilise Carathéodory's[2] extension theorem from 1918. Readers more interested in applications of probability theory than real analysis may skip the proof. A concise but rigorous proof is available for example in [Kal02, Theorem 2.2].

[2] Constantin Carathéodory, 1873–1950. PhD 1904 @ University of Göttingen for Hermann Minkowski.

(ii) The uniqueness can by done by applying Dynkin's identification theorem in Section 2.6. See Exercise 2.15. □

The *Lebesgue measure*[3] is the unique shift-invariant measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda([0,1]) = 1$, which is well defined due to Theorem 2.5.

Proposition 2.6. *The Lebesgue measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *satisfies:*

(i) $\lambda(\{x\}) = 0$ *for all* x*.*

(ii) $\lambda([a,b]) = b - a$ *for all* $a \le b$*.*

(iii) $\lambda(\mathbb{R}) = \infty$*.*

The Lebesgue measure of an interval equals the length of the interval. In this way, the Lebesgue measure extends the concept of length from intervals to arbitrary Borel sets.

Proof. By noting that the intervals $\big[0, \frac{1}{n}\big)$, $\big[\frac{1}{n}, \frac{2}{n}\big)$, $\big[\frac{2}{n}, \frac{3}{n}\big)$ ..., are disjoint, we see by applying shift invariance that

$$\lambda\big(\big[0, \tfrac{m}{n}\big)\big) \;=\; \lambda\Bigg(\bigcup_{k=1}^{m} \big[\tfrac{k-1}{n}, \tfrac{k}{n}\big)\Bigg) \;=\; \sum_{k=1}^{m} \lambda\big(\big[\tfrac{k-1}{n}, \tfrac{k}{n}\big)\big) \;=\; m\lambda\big(\big[0, \tfrac{1}{n}\big)\big) \qquad (2.5)$$

Equality (2.5) for $m = n$ implies that

$$\lambda\big(\big[0, \tfrac{1}{n}\big)\big) \;=\; \tfrac{1}{n}\lambda\big([0,1)\big). \qquad (2.6)$$

The assumption $\lambda([0,1]) = 1$ combined with monotonicity (Proposition 1.4) implies that $\lambda\big([0, \frac{1}{n})\big) \le \frac{1}{n}$. The monotone continuity of measures (Proposition 1.5) then implies that

$$0 \;\le\; \lambda(\{0\}) \;=\; \lambda\Bigg(\bigcap_{n=1}^{\infty} \big[0, \tfrac{1}{n}\big)\Bigg) \;=\; \lim_{n\to\infty} \lambda\big(\big[0, \tfrac{1}{n}\big)\big) \;\le\; \lim_{n\to\infty} \tfrac{1}{n} \;=\; 0.$$

Therefore $\lambda(\{0\}) = 0$, and claim (i) follows by shift invariance.

Due to (i), we see that $\lambda([0,1)) = \lambda([0,1]) = 1$. Then (2.6) implies that $\lambda\big(\big[0, \frac{1}{n}\big)\big) = \frac{1}{n}$. By substituting this back to (2.5), we conclude that $\lambda\big(\big[0, \frac{m}{n}\big)\big) = \frac{m}{n}$, in other words,

$$\lambda\big([0, q)\big) \;=\; q \qquad \text{for all rational numbers } q > 0. \qquad (2.7)$$

[3] Named after Henri Lebesgue (1875–1941). PhD 1902 @ Nancy for Émile Borel.

We may extend (2.7) to an arbitrary real number $x > 0$ as follows. Select a sequence of positive rational numbers such that $q_n \uparrow x$. Then $[0, q_n) \uparrow [0, x)$, and the monotone continuity of measures combined with (2.7) implies that

$$\lambda\big([0,x)\big) \;=\; \lim_{n\to\infty} \lambda\big([0,q_n)\big) \;=\; \lim_{n\to\infty} q_n \;=\; x.$$

Because $\lambda(\{b\}) = 0$, it follows by shift invariance that $\lambda([a,b]) = \lambda([a,b)) = \lambda([0,b-a)) = b-a$ for all $a \le b$. This confirms (ii).

Finally, by the monotone continuity of measures, $\lambda(\mathbb{R}) = \lambda(\cup_{n=1}^{\infty}[-n,n]) = \lim_{n\to\infty}\lambda([-n,n]) = \lim_{n\to\infty} 2n = \infty$. Hence (iii) is true. □

The Lebesgue measure λ on $\mathbb{R}$ and the counting measure $\#$ on $\mathbb{Z}$ are analogous. The former is the uniform measure on the continuum, and the latter is the uniform measure on the discrete integer lattice.

2.6 Identification of measures

The following result is attributed to Eugene Dynkin[4][5].

Theorem 2.7 (Dynkin's identification theorem)**.** *Let $\mathcal{C}$ be a generator of $\mathcal{S}$ that is closed under pairwise intersection. Then for any probability measures on $(S, \mathcal{S})$:*

$$\mu = \nu \qquad \textit{if and only if} \qquad \mu(A) = \nu(A) \textit{ for all } A \in \mathcal{C}. \tag{2.8}$$

Any probability measure on $(S, \mathcal{S})$ is uniquely identified by its values on any particular generator of $\mathcal{S}$ that is closed under pairwise intersection.

The proof of Theorem 2.7 is based on a monotone class argument that utilises the following fundamental set-theoretic result. A set family $\mathcal{D}$ on S is called a *Dynkin class*[6] if it contains S and is closed under subset difference[7] and increasing set limit[8].

[4] Eugene Dynkin (1924–2014). PhD 1948 @ Moscow State University for Kolmogorov.

[5] A set family that is closed under pairwise intersection is sometimes called a *π-system*.

[6] aka *Dynkin system*, *λ-system*

[7] $A_1, A_2 \in \mathcal{D},\ A_1 \subset A_2 \implies A_2 \setminus A_1 \in \mathcal{D}$.

[8] $A_1, A_2, \dots \in \mathcal{D},\ A_1 \subset A_2 \subset \cdots \implies \cup_n A_n \in \mathcal{D}$.

Theorem 2.8 (Monotone class theorem)**.** *If $\mathcal{C}$ is a set family that is closed under pairwise intersection and generates a sigma-algebra $\mathcal{S}$, then every Dynkin class containing $\mathcal{C}$ also contains $\mathcal{S}$.*

A monotone-class argument *is a proof technique for verifying that a certain property $\mathcal{P}$ is valid for all sets in a sigma-algebra $\mathcal{S}$.*

1. *Verify that property $\mathcal{P}$ holds for all members of a generator of $\mathcal{S}$ that is closed under pairwise intersection.*
2. *Verify that $\mathcal{D} = \{$sets in $\mathcal{S}$ having property $\mathcal{P}\}$ forms a Dynkin class.*
3. *Apply Theorem 2.8 to conclude that $\mathcal{D} \supset \mathcal{S}$, so that indeed $\mathcal{D} = \mathcal{S}$, and therefore all members of $\mathcal{S}$ have property $\mathcal{P}$.*

The purely set-theoretic proof of the monotone class theorem can be skipped without missing important insights in probabilistic thinking. For the curious reader, the proof is given in Appendix B.

Proof of Theorem 2.7. Let $\mathcal{C}$ be a generator of $\mathcal{S}$ that is closed under pairwise intersection. We will prove the forward implication in (2.8) (the converse is trivial). Let μ, ν be probability measures on $(S, \mathcal{S})$ such that

$$\mu(A) = \nu(A) \qquad \text{for all } A \in \mathcal{C}. \tag{2.9}$$

We will prove that $\mu = \nu$ by showing that (2.9) holds for all $A \in \mathcal{S}$.

The proof strategy, known as a monotone-class argument, is to represent the collection of sets that have the desired property as a set family

$$\mathcal{D} \;=\; \{A \in \mathcal{S} \colon \mu(A) = \nu(A)\}.$$

We verify that $\mathcal{D}$ is a Dynkin class as follows:

(i) $\mathcal{D}$ contains the ground set S because $\mu(S) = 1$ and $\nu(S) = 1$.

(ii) Assume that $A, B \in \mathcal{D}$ are such that $A \subset B$. Then

$$\mu(B \setminus A) \;=\; \mu(B) - \mu(A) \;=\; \nu(B) - \nu(A) \;=\; \nu(B \setminus A),$$

so that $B \setminus A \in \mathcal{D}$.

(iii) Assume that $A_1, A_2, \dots \in \mathcal{D}$ are such that $A_1 \subset A_2 \subset \cdots$. The monotone continuity of measures (Proposition 1.6) then implies that

$$\mu(\cup_n A_n) \;=\; \lim_{n\to\infty} \mu(A_n) \;=\; \lim_{n\to\infty} \nu(A_n) \;=\; \nu(\cup_n A_n),$$

so that $\cup_n A_n \in \mathcal{D}$.

Due to (2.9) we see that $\mathcal{D} \supset \mathcal{C}$. Because $\mathcal{C}$ is closed under pairwise intersection and generates $\mathcal{S}$, the monotone class theorem (Theorem 2.8) implies that $\mathcal{D} \supset \mathcal{S}$. This means that (2.9) holds for all $A \in \mathcal{S}$. Hence $\mu = \nu$. □

Dynkin's identification theorem has an important corollary for probability measures on the real line. The *cumulative distribution function* of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a function $F\colon \mathbb{R} \to [0,1]$ defined by

$$F(x) \;=\; \mu((-\infty, x]).$$

Theorem 2.9. *Probability measures μ_1 and μ_2 on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are equal if and only if their cumulative distributions functions F_1 and F_2 are equal.*

Proof. Assume that $F_1 = F_2$. Define a set family on the real line by $\mathcal{C} = \{(-\infty, x]\colon x \in \mathbb{R}\}$. Because $(-\infty, x] \cap (-\infty, y] = (-\infty, x \wedge y]$ for all x, y, we see that $\mathcal{C}$ is closed under pairwise intersection. Because

$$\mu_1((-\infty, x]) = F_1(x) = F_2(x) = \mu_2((-\infty, x])$$

for all real numbers x, we see that $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{C}$. We also know (Proposition 2.4) that $\mathcal{C}$ is a generator of $\mathcal{B}(\mathbb{R})$. Dynkin's identical theorem (Theorem 2.7) now implies that $\mu_1 = \mu_2$. The converse implication is immediate. □

2.7 Exercises

Exercise 2.10 (Truncated measures)**.** The *truncation* of a measure μ on $(S, \mathcal{S})$ into a set $C \in \mathcal{S}$ is a set function μ_C defined by formula $\mu_C(A) = \mu(A \cap C)$, $A \in \mathcal{S}$.

(a) Prove that μ_C is a measure on $(S, \mathcal{S})$.

(b) Prove Proposition 2.2 with the help of (a) and Proposition 2.1.

(c) Assume that S is countable. Determine the probability mass function of the probability measure in Proposition 2.2.

Exercise 2.11 (Sets of Lebesgue measure zero)**.** Which of the following sets have zero Lebesgue measure? Justify your answer carefully.

$$A_1 = \{x\}$$

$$A_2 = \mathbb{Z}$$

$$A_3 = \text{generic open set}$$

$$A_4 = \text{generic nonempty finite set}$$

$$A_5 = \text{generic countably infinite set}$$

Exercise 2.12 (Lebesgue's properties)**.** Are the following statements true or false for the Lebesgue measure on $\mathbb{R}$? Justify your answers rigorously.

(a) There exist a countably infinite set $A \subset \mathbb{R}$ such that $\lambda(A) = 0$.

(b) There exist an unbounded Borel set $A \subset \mathbb{R}$ such that $\lambda(A) = 3$.

(c) There exist a bounded Borel set $A \subset \mathbb{R}$ such that $\lambda(A) = \infty$.

Exercise 2.13 (Uniform distribution on an infinite countable set)**.** Does there exist a probability distribution P on $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ that is uniform in the sense that the probability mass function $x \mapsto P(\{x\})$ is constant? Justify your answer in detail.

Exercise 2.14 (Shift-invariant discrete measures)**.** A measure μ on $(\mathbb{Z}, 2^{\mathbb{Z}})$ is called shift-invariant if $\mu(A + h) = \mu(A)$ for all $A \subset \mathbb{Z}$ and all $h \in \mathbb{Z}$. The counting measure $\#$ on $\mathbb{Z}$ is shift-invariant and satisfies $\#(\{0\}) = 1$. Do there exist other such measures? Explain your answer rigorously.

Exercise 2.15 (Uniqueness of the Lebesgue measure)**.** Assume that λ_1, λ_2 are measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that are shift-invariant and such that $\lambda_i([0, 1]) = 1$ for $i = 1, 2$. Prove that $\lambda_1 = \lambda_2$ as follows.

(a) For an integer $n \geq 1$, define $\lambda_{i,n}(A) = \frac{1}{2n}\lambda_i(A \cap [-n, n])$ for $i = 1, 2$.

(b) Verify that $\lambda_{i,n}$ is a probability measure.

(c) Verify that the cumulative distribution functions of $\lambda_{1,n}$ and $\lambda_{2,n}$ are equal. (Hint: Proposition 2.6.).

(d) Conclude that $\lambda_{1,n} = \lambda_{2,n}$ for all n.

(e) Conclude that $\lambda_1 = \lambda_2$ (Hint: Monotone continuity of measures.)

2.8 Notes

The Lebesgue measure was first defined by Borel. Lebesgue developed the concept of integration with respect to this measure.

The monotone class theorem was proved in 1928 by a Polish mathematician Wacław Sierpiński in [Sie28], but apparently it was first applied in probability theory by a Russian–American mathematician Eugene Dynkin [Dyn61].

An excellent summary of the history of probability theory is in the appendix of [Kal02].

Chapter 3

Random variables

> *Je ne connais aucune fonction, qui ne soit sommable, je ne sais s'il en existe.*
>
> *—Henri Lebesgue*

Random variables are observable quantities associated with a random outcome. They are modelled as functions which are measurable in the sense that their preimages of measurable sets are measurable.

Key concepts: random variable, measurable function, law of a random variable

Learning outcomes:

- Get introduced to modelling random variables as functions defined on a probability space.
- Learn to operate with preimages of functions.
- Get familiar with the concept of the law of a random variable.
- Learn to construct new functions by pointwise operations on functions.
- Gain insight on how measurability is preserved in standard pointwise operations on functions.

Prerequisites: limits, suprema, and infima of number sequences

3.1 Preimages

The *preimage* of a set $B \subset S_2$ under a function $f\colon S_1 \to S_2$ is the set of points that f maps into B, denoted by

$$f^{-1}(B) \;=\; \{x \in S_1 \colon f(x) \in B\}.$$

If f is a bijection with inverse function $y \mapsto f^{-1}(y)$, then we see that $f^{-1}(B) = \{f^{-1}(y)\colon y \in B\}$. A function that is not a bijection does not admit an inverse function, but the preimages of f are well define also in this case.

The preimage operation $B \mapsto f^{-1}(B)$ maps subsets of S_2 into subsets of S_1. Preimages are always well defined for any function f.

Proposition 3.1. *For any function $f\colon S_1 \to S_2$ and for arbitrary sets $B, B_i \subset S_2$, $i \in I$,*

$$\begin{aligned} f^{-1}(B^c) &= f^{-1}(B)^c && (3.1)\\ f^{-1}(\cap_{i\in I} B_i) &= \cap_{i\in I} f^{-1}(B_i) && (3.2)\\ f^{-1}(\cup_{i\in I} B_i) &= \cup_{i\in I} f^{-1}(B_i). && (3.3) \end{aligned}$$

Proof. Equality (3.1) follows by

$$\begin{aligned} f^{-1}(B^c) &= \{x \in S_1 \colon f(x) \in B^c\}\\ &= \{x \in S_1 \colon f(x) \notin B\}\\ &= \{x \in S_1 \colon x \notin f^{-1}(B)\}\\ &= \{x \in S_1 \colon x \in f^{-1}(B)^c\}\\ &= f^{-1}(B)^c, \end{aligned}$$

and equality (3.2) by

$$\begin{aligned} f^{-1}(\cap_{i\in I} B_i) &= \{x \in S_1 \colon f(x) \in \cap_{i\in I} B_i\}\\ &= \{x \in S_1 \colon f(x) \in B_i \text{ for all } i \in I\}\\ &= \{x \in S_1 \colon x \in f^{-1}(B_i) \text{ for all } i \in I\}\\ &= \{x \in S_1 \colon x \in \cap_{i\in I} f^{-1}(B_i)\}\\ &= \cap_{i\in I} f^{-1}(B_i). \end{aligned}$$

Equality (3.3) follows by replacing '$\cap$' ↦ '$\cup$' and 'all' ↦ 'some' above. □

3.2 Measurable functions

A function $f\colon S_1 \to S_2$ associated with measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ is called *measurable* if

$$f^{-1}(B) \in \mathcal{S}_1 \quad \text{for all } B \in \mathcal{S}_2,$$

where $f^{-1}(B) = \{x\colon f(x) \in B\}$ denotes the preimage of B by f. If the associated sigma-algebras are not clear from the context, we say that a function is $\mathcal{S}_1/\mathcal{S}_2$-measurable.

A function f is measurable iff the preimages of measurable sets under f are measurable sets.

Note the analogy with topology.

A function f is continuous iff the preimages of open sets under f are open sets.

Proposition 3.2. *The indicator function 1_A of a measurable set A is measurable.*

Proof. Fix a measurable space $(S, \mathcal{S})$ and consider a set $A \in \mathcal{S}$. Observe that for any subset B of the real line (or the extended real line),

$$1_A^{-1}(B) = \begin{cases} \emptyset & \text{if } 0, 1 \notin B, \\ A & \text{if } 0 \notin B \text{ and } 1 \in B, \\ A^c & \text{if } 0 \in B \text{ and } 1 \notin B, \\ S & \text{if } 0, 1 \in B. \end{cases}$$

Therefore, all possible preimages of 1_A are the sets $\emptyset, A, A^c, S$. These are all contained in $\mathcal{S}$ because $A \in \mathcal{S}$.

□

The following result provides a convenient sufficient condition for verifying that $f\colon S_1 \to S_2$ is measurable.

Proposition 3.3. *Let $\mathcal{C}$ be a generator of $\mathcal{S}_2$. Then f is $\mathcal{S}_1/\mathcal{S}_2$-measurable if and only if*

$$f^{-1}(B) \in \mathcal{S}_1 \quad \textit{for all } B \in \mathcal{C}. \tag{3.4}$$

Proof. If f is $\mathcal{S}_1/\mathcal{S}_2$-measurable, then (3.4) follows immediately.

Verifying the less immediate direction is an example of a proof technique where we first specify a set family containing all 'good' sets, and we then prove that this set family is so large that the claim follows. In this case the good sets are the members of $\mathcal{S}_2$ for which $f^{-1}(G) \in \mathcal{S}_1$. Then we define a set family

$$\mathcal{G} \ = \ \{G \in \mathcal{S}_2 : f^{-1}(G) \in \mathcal{S}_1\}.$$

Assumption (3.4) implies that $\mathcal{C} \subset \mathcal{G}$. We will next verify that $\mathcal{G}$ is a sigma-algebra:

(i) $f^{-1}(S_2) = S_1 \in \mathcal{S}_1$ implies that $S_2 \in \mathcal{G}$. Similarly, $f^{-1}(\emptyset) = \emptyset \in \mathcal{S}_1$ implies that $\emptyset \in \mathcal{G}$. Hence $\mathcal{G}$ contains $\emptyset, S_2$.

(ii) If $B \in \mathcal{G}$, then $f^{-1}(B) \in \mathcal{S}_1$, so that in light of (3.1) it follows that $f^{-1}(B^c) = (f^{-1}(B))^c \in \mathcal{S}_1$. Therefore, $B^c \in \mathcal{G}$. Hence $\mathcal{G}$ is closed under complement.

(iii) Verifying that $\mathcal{G}$ is closed under countable intersection and countable union can be done a similar manner as in (ii) using (3.2)–(3.3).

We see that $\mathcal{G}$ is a sigma-algebra on S_2 containing the set family $\mathcal{C}$. By definition, $\sigma(\mathcal{C})$ is the smallest such sigma-algebra. Therefore, $\mathcal{S}_2 = \sigma(\mathcal{C}) \subset \mathcal{G}$. Therefore, every set $G \in \mathcal{S}_2$ belongs to $\mathcal{G}$ and hence satisfies the property $f^{-1}(G) \in \mathcal{S}_1$. Hence f is $\mathcal{S}_1/\mathcal{S}_2$-measurable. □

Proposition 3.4. *Continuous functions are measurable.*

Proof. Let $f \colon S_1 \to S_2$ where S_1 and S_2 are arbitrary topological spaces equipped with Borel sigma-algebras $\mathcal{S}_1$ and $\mathcal{S}_2$. Because f is continuous, we know that $f^{-1}(B)$ is open for every open set $B \subset S_2$. In other words,

$$f^{-1}(B) \in \mathcal{T}_1 \qquad \text{for all } B \in \mathcal{T}_2,$$

where $\mathcal{T}_1$ and $\mathcal{T}_2$ denote the families of open sets in S_1 and S_2. Because every open set is a Borel set, it follows that

$$f^{-1}(B) \in \mathcal{S}_1 \qquad \text{for all } B \in \mathcal{T}_2,$$

Because $\mathcal{T}_2$ is a generator of $\mathcal{S}_2$, the claim follows by Proposition 3.3. □

3.3 Preservation of measurability

3.3.1 Compositions

The *composition* of functions $f\colon S_1 \to S_2$ and $g\colon S_2 \to S_3$ is a function $h = g \circ f\colon S_1 \to S_3$ defined by $h(x) = g(f(x))$.

Preimages work 'backwards' under compositions of functions.

$$\begin{tikzcd} S_1 \arrow[r, "f"] & S_2 \arrow[r, "g"] & S_3 \end{tikzcd}$$

$$\begin{tikzcd} f^{-1}(g^{-1}(C)) & g^{-1}(C) \arrow[l, "f^{-1}"'] & C. \arrow[l, "g^{-1}"'] \end{tikzcd}$$

Proposition 3.5. *If $f\colon S_1 \to S_2$ is $\mathcal{S}_1/\mathcal{S}_2$-measurable and $g\colon S_2 \to S_3$ is $\mathcal{S}_2/\mathcal{S}_3$-measurable, then the composite function $g \circ f$ is $\mathcal{S}_1/\mathcal{S}_3$-measurable.*

Proof. Denote $h = g \circ f$. Fix a set $C \in \mathcal{S}_3$, and observe that

$$\begin{aligned} h^{-1}(C) &= \{x \in S_1 \colon g(f(x)) \in C\} \\ &= \{x \in S_1 \colon f(x) \in g^{-1}(C)\} \\ &= f^{-1}(g^{-1}(C)). \end{aligned}$$

Because g is measurable, we see that $B = g^{-1}(C) \in \mathcal{S}_2$. Because f is measurable, it follows that $h^{-1}(C) = f^{-1}(B) \in \mathcal{S}_1$. We conclude that h is $\mathcal{S}_1/\mathcal{S}_3$-measurable. □

Compositions of measurable functions are measurable.

3.3.2 Pointwise operations

The *pointwise multiplication* of a function $f\colon S \to \mathbb{R}$ by a scalar $c \in \mathbb{R}$ yields a function $cf\colon S \to \mathbb{R}$ defined by $(cf)(x) = cf(x)$.

Proposition 3.6. *If $f\colon S \to \mathbb{R}$ is measurable, then so is cf, for any $c \in \mathbb{R}$.*

Proof. We may represent cf as a composition $cf = \psi \circ f$ where $\psi\colon \mathbb{R} \to \mathbb{R}$ is defined by $\psi(y) = cy$. The function ψ is measurable as a continuous function (Proposition 3.4). Because compositions of measurable functions are measurable (Proposition 3.5), it follows that $cf = \psi \circ f$ is measurable. □

The *pointwise sum* of functions $f, g\colon S \to \mathbb{R}$ is a function $f+g$ defined by $(f+g)(x) = f(x) + g(x)$. The pointwise product, min, max are defined analogously.

Proposition 3.7. *If $f, g\colon S \to \mathbb{R}$ are measurable, then so are the functions $f+g$, fg, $f \wedge g$, $f \vee g$. Furthermore, f/g is measurable when $g \neq 0$ on S.*

Proof sketch. The pointwise sum of two functions can be represented as a composition

$$f+g \;=\; \psi \circ \phi \tag{3.5}$$

in which $\phi\colon S \to \mathbb{R}^2$ is defined by $\phi(x) = (f(x), g(x))$ and $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ is defined by $\psi(y_1, y_2) = y_1 + y_2$. It is intuitively clear that ϕ is measurable because both its coordinate functions are measurable. Furthermore, ψ is measurable as a continuous function (Proposition 3.4). Because compositions of measurable functions are measurable (Proposition 3.5), it follows that $f+g = \psi \circ \phi$ is measurable.

The pointwise product of two functions can be written as a similar composition $fg = \psi \circ \phi$ where ϕ is the same as above, but ψ is updated to $\psi(y_1, y_2) = y_1 y_2$. Again $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ is continuous and hence measurable. Hence the same argument as above concludes fg is measurable. Verifying the measurability of the pointwise min, max, and quotient of f and g can be completed analogously.

'It is intuitively clear that ϕ is measurable' is the part the we skipped in the proof. Verifying this is harder than what one might expect. A curious reader is recommended to consult [Kal02, Lemma 1.12] for details. □

Below $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ denotes the extended real line equipped with the Borel sigma-algebra $\mathcal{B}(\bar{\mathbb{R}})$ generated by open sets in $\bar{\mathbb{R}}$. A set is open in $\bar{\mathbb{R}}$ if and only if it can be written as a union of open intervals (a, b) and open rays of form $[-\infty, a)$ and $(b, \infty]$.

Proposition 3.8. *If the functions $f_1, f_2, \ldots : S \to \bar{\mathbb{R}}$ are measurable, then so are the pointwise infimum $\inf_n f_n$ and pointwise supremum $\sup_n f_n$.*

Proof. (i) Let $g = \sup_n f_n$. Then for any $t \in \mathbb{R}$,

$$\begin{aligned} g^{-1}([-\infty, t]) &= \{x \in S\colon g(x) \le t\} \\ &= \{x \in S\colon f_n(x) \le t \text{ for all } n\} \\ &= \cap_{n \ge 1} \{x \in S\colon f_n(x) \le t\} \\ &= \cap_{n \ge 1} f_n^{-1}([-\infty, t]). \end{aligned}$$

Because $[-\infty, t]$ is a Borel set in $\bar{\mathbb{R}}$, it follows that $f_n^{-1}([-\infty, t]) \in \mathcal{S}$ for all n. Hence $g^{-1}([-\infty, t]) \in \mathcal{S}$. In particular, $g^{-1}(B) \in \mathcal{S}$ for every B in the set family $\mathcal{C} = \{[-\infty, t] : t \in \mathbb{R}\}$. Because $\mathcal{C}$ is a generator[1] of $\mathcal{B}(\bar{\mathbb{R}})$, it follows [Kal02, Lemma 1.4] that g is measurable.

(ii) Next, by noting that $\inf_n f_n = -\sup_n(-f_n)$, it follows from (i) that $\inf_n f_n$ is measurable. □

Proposition 3.9. *If the functions $f_1, f_2, \ldots : S \to \bar{\mathbb{R}}$ are measurable, and $f_n \to f$ pointwise, then f is measurable.*

Proof. If a sequence $x_1, x_2, \ldots$ converges in $\bar{\mathbb{R}}$ according to $\lim_{n\to\infty} x_n = x$, then

$$x \;=\; \limsup_{n\to\infty} x_n \;=\; \inf_{m\geq 1} \sup_{n\geq m} x_n.$$

Under the assumptions that f is a pointwise limit, the limit function f may then be represented pointwise by

$$f \;=\; \inf_{m\geq 1} \underbrace{\sup_{n\geq m} f_n}_{g_m}.$$

Then each g_m is measurable by Proposition 3.8. Then $f = \inf_{m\geq 1} g_m$ is measurable, again by Proposition 3.8. □

3.4 Measurability of countable-range functions

Lemma 3.10. *A function $f\colon S \to \bar{\mathbb{R}}$ with a countable range is measurable if and only if $f^{-1}(\{y\}) \in \mathcal{S}$ for all y in the range of f.*

Proof. The forward implication ‘only if’ is immediate because singleton sets $\{y\}$ are Borel sets. To prove the backward implication ‘if’, denote the range of f by $f(S) = \{f(x)\colon x \in S\}$. Fix a Borel set $B \subset \mathbb{R}$. Note that $B = \cup_{y\in B}\{y\}$, so that $f^{-1}(B) = \cup_{y\in B} f^{-1}(\{y\})$ by (3.3). We also note that $f^{-1}(\{y\}) = \emptyset$ for $y \notin f(S)$. Therefore, $f^{-1}(B) = \cup_{y\in B\cap f(S)} f^{-1}(\{y\})$ equals a countable union of measurable sets $f^{-1}(\{y\})$. Because countable unions of measurable sets are measurable, it follows that $f^{-1}(B)$ is measurable. □

[1] This may be proved in the same manner as for $\mathbb{R}$ or $[0, \infty]$.

3.5 Random variables

A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$ in which Ω is a set, $\mathcal{A}$ is a sigma-algebra on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{A})$. A *random variable* on probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a measurable function $X : \Omega \to S$ where $(S, \mathcal{S})$ is a measurable space.

Commonly used informal notations $\mathbb{P}(X = x)$ and $\mathbb{P}(X \in B)$ now have precise meanings

$$\begin{aligned} \mathbb{P}(X = x) &= \mathbb{P}(X^{-1}(\{x\})) = \mathbb{P}(\{\omega \in \Omega \colon X(\omega) = x\}), \\ \mathbb{P}(X \in B) &= \mathbb{P}(X^{-1}(B)) \quad\ = \mathbb{P}(\{\omega \in \Omega \colon X(\omega) \in B\}). \end{aligned}$$

A *random vector* on probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a measurable function $X : \Omega \to \mathbb{R}^n$ where $\mathbb{R}^n$ is always seen as a measurable space equipped with the Borel sigma-algebra $\mathcal{B}(\mathbb{R}^n)$, the smallest sigma-algebra containing the open sets in $\mathbb{R}^n$. The observables may also assume values in a space $S \neq \mathbb{R}^n$. In this case we equip S with its Borel sigma-algebra $\mathcal{B}(S)$ when S is a topological space, or some other sigma-algebra otherwise. Different types of random variables are described in Table 3.1.

Table 3.1: Examples of random variables

Name	S	$\mathcal{S}$
Random integer	$\mathbb{Z}$	$2^{\mathbb{Z}}$
Random number	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$
Random vector	$\mathbb{R}^n$	$\mathcal{B}(\mathbb{R}^n)$
Random matrix	$\mathbb{R}^{m \times n}$	$\mathcal{B}(\mathbb{R}^{m \times n})$
Random sequence	$\mathbb{R}^\infty$	$\mathcal{B}(\mathbb{R}^\infty)$
Continuous stochastic process	$C[0,T]$	$\mathcal{B}(C[0,T])$
Undirected random graph	$\{0,1\}^{\binom{n}{2}}$	$2^{\{0,1\}^{\binom{n}{2}}}$

3.6 Law of a random variable

Let $X : \Omega \to S$ be a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where $(S, \mathcal{S})$ is a measurable space. The *law* or *distribution* of X is a set function $\mu \colon \mathcal{S} \to [0,1]$ defined by

$$\mu(B) = \mathbb{P}(X^{-1}(B)).$$

Proposition 3.11. *The law of X is a probability measure on $(S, \mathcal{S})$.*

Proof. (i) We note that $\mu(\emptyset) = \mathbb{P}(X^{-1}(\emptyset)) = \mathbb{P}(\emptyset) = 0$.

(ii) Assume that $B_1, B_2, \dots \in \mathcal{S}$ are disjoint. Then

$$X^{-1}(B_i) \cap X^{-1}(B_j) \overset{(3.2)}{=} X^{-1}(B_i \cap B_j) \ = \ X^{-1}(\emptyset) \ = \ \emptyset \qquad \text{for } i \neq j,$$

so that also the preimages $X^{-1}(B_1), X^{-1}(B_2), \dots \in \mathcal{S}$ are disjoint. It follows that

$$\begin{aligned} \mu\Big(\bigcup_i B_i\Big) &= \mathbb{P}\Big(X^{-1}\big(\bigcup_i B_i\big)\Big) \\ &\overset{(3.3)}{=} \mathbb{P}\Big(\bigcup_i X^{-1}(B_i)\Big) \\ &= \sum_i \mathbb{P}\Big(X^{-1}(B_i)\Big) \\ &= \sum_i \mu(B_i). \end{aligned}$$

Hence μ is countably disjointly additive, and we conclude that μ is a measure.

(iii) Finally, the observation that $\mu(S) = \mathbb{P}(X^{-1}(S)) = \mathbb{P}(\Omega) = 1$ confirms that μ is a probability measure. □

The law of X allows us to compute the probabilities of all events concerning the random variable X. When our attention is on a single random variable X, we may often ignore the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and directly operate using the law of X.

When the law of X equals μ, we say that

'X is a μ-distributed random variable'

or

'X is distributed according to μ'.

The law of X is often denoted by $\mu = \mathbb{P} \circ X^{-1}$ and called the pushforward of $\mathbb{P}$ by X.

Proposition 3.12. *For every probability measure μ on $(S, \mathcal{S})$ there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable $X\colon \Omega \to S$ such that the law of X equals μ.*

Proof. The proof is simpler than what one might expect. Define $(\Omega, \mathcal{A}, \mathbb{P}) = (S, \mathcal{S}, \mu)$ and let $X\colon \Omega \to S$ be the identity map $X(\omega) = \omega$. Then $X^{-1}(B) = B$ for any $B \in \mathcal{S}$, and we see that the law of X equals μ. □

Example 3.13. Let $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$, and $\mathbb{P} = \lambda$ be the Lebesgue measure of $\mathbb{R}$ restricted to $[0,1]$. Define a random variable $X\colon \Omega \to \mathbb{R}$ by $X(\omega) = 1_{(1/2,1]}(\omega)$. Determine the law of X.

We note that

$$X^{-1}(B) = \begin{cases} \emptyset & \text{if } 0,1 \notin B, \\ (1/2,1] & \text{if } 0 \notin B \text{ and } 1 \in B, \\ (1/2,1]^c & \text{if } 0 \in B \text{ and } 1 \notin B, \\ [0,1] & \text{if } 0,1 \in B. \end{cases}$$

Therefore,

$$\mathbb{P}(X^{-1}(B)) = \begin{cases} \lambda(\emptyset) = 0 & \text{if } 0,1 \notin B, \\ \lambda((1/2,1]) = \frac{1}{2} & \text{if } 0 \notin B \text{ and } 1 \in B, \\ \lambda((1/2,1]^c) = \frac{1}{2} & \text{if } 0 \in B \text{ and } 1 \notin B, \\ \lambda([0,1]) = 1 & \text{if } 0,1 \in B. \end{cases}$$

We find that $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

3.7 Exercises

Exercise 3.14 (Transformed random numbers)**.** Let U be a random variable uniformly distributed in $[0,1]$. Define $X = 1 - U$ and $Y = \phi(U)$ where[2] $\phi(u) = 2u - \lfloor 2u \rfloor$.

(a) Determine the cumulative distribution function of X.

(b) Determine the cumulative distribution function of Y.

(c) Are X and Y equally distributed? If yes, explain why. If no, write down a measurable set A for which $\mathbb{P}(X \in A) \neq \mathbb{P}(Y \in A)$.

[2] $\lfloor a \rfloor$ denotes the rounding down of a: the unique integer k such that $k \le a < k+1$.

Exercise 3.15 (Logs and roots random numbers)**.** Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ in which $\Omega = (0,1)$, $\mathcal{A} = \mathcal{B}((0,1))$, and $\mathbb{P}$ is the Lebesgue measure restricted to $(0,1)$. Define $X, Y\colon \Omega \to \mathbb{R}$ by

$$\begin{aligned} X(\omega) &= \log\left(\frac{1}{1-\omega}\right), & \tilde{X}(\omega) &= \log\left(\frac{1}{\omega}\right), \\ Y(\omega) &= \sqrt{\frac{1}{1-\omega}}, & \tilde{Y}(\omega) &= \sqrt{\frac{1}{\omega}}. \end{aligned}$$

(a) Convince yourself and others that $X, Y, \tilde{X}, \tilde{Y}$ are random variables.

(b) Compute the cumulative distribution function $F\colon \mathbb{R} \to [0,1]$ for X, Y.

(c) Compute the cumulative distribution function $F\colon \mathbb{R} \to [0,1]$ for $\tilde{X}, \tilde{Y}$.

(d) Some, if not all, of the laws of $X, Y, \tilde{X}, \tilde{Y}$ are familiar probability distributions. Can you recognise them?

Exercise 3.16 (Nonstandard parametric family)**.** Fix numbers $p \in [0,1]$ and $t \in (0,\infty)$ and define a set function $\mu_{p,t}\colon \mathcal{B}(\mathbb{R}) \to [0,\infty]$ by[3]

$$\mu_{p,t}(A) \;=\; \begin{cases} 1 - p + c\lambda(A \cap [0,t]) & \text{if } 0 \in A, \\ c\lambda(A \cap [0,t]) & \text{if } 0 \notin A, \end{cases}$$

where c is chosen so that $\mu_{p,t}$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(a) What is the value of c?

(b) Determine a measurable set $B \subset \mathbb{R}$ such that $\mu_{0,t}(B) = 1$ and $\mu_{1,t}(B) = 0$.

(c) Let T be a random variable in $\mathbb{R}$ distributed according to $\mu_{p,t}$. Determine values of p and t such that $\mathbb{P}(T > 0) = 0.3$ and $\mathbb{P}(4 < T < 5) = 0.01$.

Exercise 3.17 (Sampling near zero)**.** Define $P = \sum_{k=1}^{\infty} 2^{-k} \delta_{1/k^2}$.

(a) Prove that P is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(b) Compute $\mathbb{P}(X \le \frac{1}{9})$ for a P-distributed random variable X.

(c) Let Y be a random variable with a geometric distribution $\mathbb{P}(Y = k) = (1-p)^{k-1}p$, $k = 0, 1, 2, \ldots$, where $p \in (0,1)$. Determine a function ϕ such that the law of $\phi(Y)$ equals P.

[3] Unless otherwise mentioned, λ stands for the Lebesgue measure on $\mathbb{R}$.

3.8 Historical remarks

Andrey Kolmogorov made the axioms of random variables popular in his famous book from 1933. These concepts were more or less known to earlier researchers, at least Johann Radon and Maurice Fréchet. See [Kal02] for a detailed historical account.

Chapter 4

Expectations

> *Put simply, Expected Goals (xG) is a metric designed to measure the probability of a shot resulting in a goal.*
>
> *—StatsBomb Inc.*

The expectation of a real-valued random variable is the average of its possible values weighted by the corresponding probabilities. Because probabilities are represented using a measure, we need to define what is meant by a weighted average with respect to a measure. This concept is the modern definition of an integral.

Key concepts: integral against a measure

Learning outcomes:

- Learn to approximate general functions with finite-range functions
- Get introduced to the modern definition of an integral.
- Learn how to compute expected values using integrals.

Prerequisites: Previous chapters.

4.1 Approximation by finite-range functions

The integral of a nonnegative measurable function f shall be defined as limit of integrals $\int_S f_n \, d\mu$, where f_n are finite-range functions approximating the integrand. The approximating functions are defined by discretising the extended half line $\bar{\mathbb{R}}_+$ as follows.

Given an integer $n \geq 1$ we define a lattice

$$2^{-n}\mathbb{Z}_+ \;=\; \{0,\, 1 \cdot 2^{-n},\, 2 \cdot 2^{-n},\, 3 \cdot 2^{-n},\, \dots\},$$

and the *down-rounding* of $x \in \bar{\mathbb{R}}_+$ at resolution n by

$$\lfloor x \rfloor_n \;=\; \begin{cases} \max\{q \in 2^{-n}\mathbb{Z}_+ \colon q \leq x\}, & x < \infty, \\ \infty, & x = \infty. \end{cases} \tag{4.1}$$

We also define a *truncation* of $x \in \bar{\mathbb{R}}_+$ at level n by $x \wedge n = \min\{x, n\}$. Then we define an approximation of f at resolution n by

$$f_n(s) \;=\; \lfloor f(s) \rfloor_n \wedge n. \tag{4.2}$$

Proposition 4.1. *For any measurable function $f \colon S \to \bar{\mathbb{R}}_+$ and any integer $n \geq 1$, the function $f_n \colon S \to \bar{\mathbb{R}}_+$ defined by (4.2) is a measurable function with a finite range contained in $2^{-n}\mathbb{Z}_+ \cap [0, n]$. Furthermore, $f_n \uparrow f$ pointwise as $n \to \infty$.*

Proof. Observe that $f_n = \tau_n \circ \rho_n \circ f$, where $\rho_n(x) = \lfloor x \rfloor_n$ is the downrounding map, and $\tau_n(x) = x \wedge n$ is truncation map. It is not hard to verify that ρ_n is measurable (Exercise 4.16). Furthermore, $\tau_n \colon \bar{\mathbb{R}}_+ \to \bar{\mathbb{R}}_+$ is continuous, and therefore measurable (Proposition 3.4). Hence f_n is measurable, being a composition of measurable functions (Proposition 3.5).

Because the range of $\tau_n \circ \rho_n$ is contained in $2^{-n}\mathbb{Z}_+ \cap [0, n]$, so is the range of f_n. Hence the range of f_n is finite. Verifying that $f_n \to f$ is not hard. (Draw a picture.) □

All nonnegative measurable functions can be approximated by finite-range measurable functions.

4.2 Integrating nonnegative functions

A *partition* of a set S is a family of disjoint sets $A_i \subset S$ such that $\bigcup_{i \in I} A_i = S$.

A partition decomposes a set S into disjoint pieces so that each point of S is contained in exactly one piece.

The *integral* of a measurable function $f\colon S \to \bar{\mathbb{R}}_+$ against a measure μ on a measurable space $(S, \mathcal{S})$ is defined by

$$\int_S f \, d\mu \;=\; \sup_{\{A_i\}} \sum_i \Big(\inf_{s \in A_i} f(s) \Big) \, \mu(A_i), \tag{4.3}$$

where the supremum is taken with respect to all partitions $\{A_i\}$ of S into finitely many nonempty measurable sets.

The integral of a function f against a measure μ is typically denoted in many ways, including

$$\int_S f \, d\mu \;=\; \int_S f(s) \, \mu(ds) \;=\; \mu(f).$$

Proposition 4.2. *The integral of a measurable function $f\colon S \to \bar{\mathbb{R}}_+$ with a finite range $f(S) = \{f(s)\colon s \in S\}$ equals*

$$\int_S f \, d\mu \;=\; \sum_{x \in f(S)} x \, \mu\big(f^{-1}(\{x\})\big). \tag{4.4}$$

In particular, $\int_S 1_A \, d\mu = \mu(A)$ for every measurable $A \subset S$.

Proposition 4.2 will be proved using the following, slightly more general, technical result.

Lemma 4.3. *For any measurable function of form $f = \sum_{j=1}^n b_j 1_{B_j}$ in which $b_j \in \bar{\mathbb{R}}_+$ and the sets $B_1, \dots, B_n$ are disjoint, the integral equals*

$$\int_S f \, d\mu \;=\; \sum_{j=1}^n b_j \, \mu(B_j). \tag{4.5}$$

Proof. Denote $B_0 = (B_1 \cup \cdots \cup B_n)^c$ and define $b_0 = 0$. Then the sets $B_0, B_1, \dots, B_n$ form a partition of S, and $f(s) = b_j$ for all $s \in B_j$.

Let $\{A_i\colon i = 1, \dots, m\}$ be a partition of S into disjoint nonempty measurable sets, and define $a_i = \inf_{s \in A_i} f(s)$. If $A_i \cap B_j$ contains a point t, then

$a_i \le f(t) = b_j$. Therefore, $a_i \le b_j$ whenever $A_i \cap B_j$ is nonempty, and we see that

$$a_i \mu(A_i \cap B_j) \ \le \ b_j \mu(A_i \cap B_j) \qquad \text{for all } i, j.$$

By observing that $\{A_i \cap B_j \colon j = 0, \dots, n\}$ is a partition of A_i, and that $\{A_i \cap B_j \colon i = 1, \dots, m\}$ is a partition of B_j, we find that

$$\mu(A_i) \ = \ \sum_{j=0}^n \mu(A_i \cap B_j) \qquad \text{and} \qquad \mu(B_j) \ = \ \sum_{i=1}^m \mu(A_i \cap B_j).$$

It follows that

$$\sum_{i=1}^m a_i \mu(A_i) \ = \ \sum_{i=1}^m \sum_{j=0}^n a_i \mu(A_i \cap B_j) \ \le \ \sum_{i=1}^m \sum_{j=0}^n b_j \mu(A_i \cap B_j) \ = \ \sum_{j=0}^n b_j \mu(B_j).$$

By recalling the definition of a_i, we see that

$$\sum_i \Big(\inf_{s \in A_i} f(s) \Big) \mu(A_i) \ \le \ \sum_{j=0}^n b_j \mu(B_j). \tag{4.6}$$

The above computation shows that (4.6) holds for an arbitrary finite partition $\{A_i\}$. Furthermore, (4.6) holds as equality if we choose $\{A_i\}$ to be the partition of S consisting of the nonempty sets among $B_0, B_1, \dots, B_n$. We may hence conclude that

$$\int_S f \, d\mu \ = \ \sup_{\{A_i\}} \sum_i \Big(\inf_{s \in A_i} f(s) \Big) \mu(A_i) \ = \ \sum_{j=0}^n b_j \mu(B_j).$$

We may omit the first term in the sum on the right because $b_0 = 0$. □

Proof of Proposition 4.2. Enumerate the range of the integrand as $f(S) = \{b_1, \dots, b_n\}$, and denote $B_j = f^{-1}(\{b_j\})$. Then the integrand may be written as $f = \sum_{j=1}^n b_j 1_{B_j}$ with B_j disjoint. Lemma 4.3 then implies that

$$\int_S f \, d\mu \ = \ \sum_{j=1}^n b_j \, \mu(B_j) \ = \ \sum_{x \in f(S)} x \, \mu\big(f^{-1}(\{x\})\big).$$

In particular, if $f = 1_A$ for some measurable $A \subset S$, then the range of f equals $\{0, 1\}$, the preimages are given by $f^{-1}(\{0\}) = A^c$ and $f^{-1}(\{1\}) = A$, so that the above formula becomes

$$\int_S f \, d\mu \ = \ 0 \cdot \mu(A^c) + 1 \cdot \mu(A) \ = \ \mu(A).$$

□

Proposition 4.4 (Monotonicity)**.** *For any measurable functions $f, g\colon S \to \bar{\mathbb{R}}_+$, $f \le g$ pointwise $\implies \mu(f) \le \mu(g)$.*

Proof. For any finite partition of S into measurable sets A_i, we see that $\inf_{s \in A_i} f(s) \le \inf_{s \in A_i} g(s)$ for all i, so that

$$\sum_i \Big(\inf_{s \in A_i} f(s) \Big) \mu(A_i) \ \le \ \sum_i \Big(\inf_{s \in A_i} g(s) \Big) \mu(A_i).$$

The claim hence follows by taking suprema on both sides of the above inequality, and recalling the defining formula (4.3). □

Theorem 4.5 (Monotone continuity)**.** *For arbitrary measurable functions $f, f_n\colon S \to \bar{\mathbb{R}}_+$, $f_n \uparrow f \implies \mu(f_n) \uparrow \mu(f)$.*

Proof. Because $f_1 \le f_2 \le \cdots \le f$, we see by applying Proposition 4.4 that $\mu(f_1) \le \mu(f_2) \le \cdots \le \mu(f)$. Because every nondecreasing sequence in $\bar{\mathbb{R}}_+$ converges, it follows that $\mu(f_n)$ converges to a limit in $\bar{\mathbb{R}}_+$ that satisfies

$$\lim_{n \to \infty} \mu(f_n) \ \le \ \mu(f). \tag{4.7}$$

To prove a corresponding reverse inequality we will show that for every finite partition $\{A_i \colon i \in I\}$ of S into measurable sets,

$$\lim_{n \to \infty} \mu(f_n) \ \ge \ \underbrace{\sum_{i \in I} a_i \mu(A_i)}_{L}, \tag{4.8}$$

where $a_i = \inf_{s \in A_i} f(s)$.

(i) Assume that $0 < L < \infty$. Then the set

$$I_+ \ = \ \{ i \in I \colon a_i \mu(A_i) > 0 \} \tag{4.9}$$

is nonempty, and we find that $a_i \in (0, \infty)$ and $\mu(A_i) \in (0, \infty)$ for all $i \in I_+$. Fix a small enough number $\epsilon > 0$ such that $\epsilon < a_i$ for all $i \in I_+$. Define

$$A_{in} \ = \ \{ s \in A_i \colon f_n(s) > a_i - \epsilon \}, \qquad i \in I_+.$$

Because the sets $\{A_i \colon i \in I_+\}$ are disjoint, so are $\{A_{in} \colon i \in I_+\}$. The family of sets A_{in}, $i \in I_+$, augmented with the set $B_n = (\cup_{i \in I_+} A_{in})^c$, hence forms a partition of S. Therefore,

$$\begin{aligned} \mu(f_n) \ &\ge \ \sum_{i \in I_+} \Big(\inf_{s \in A_{in}} f_n(s) \Big) \mu(A_{in}) + \Big(\inf_{s \in B_n} f_n(s) \Big) \mu(B_n) \\ &\ge \ \sum_{i \in I_+} (a_i - \epsilon)\, \mu(A_{in}). \end{aligned}$$

Because $f_n \uparrow f$ pointwise, and $f(s) > a_i - \epsilon$ for all $s \in A_i$, we find that $A_{in} \uparrow A_i$ as $n \to \infty$. The monotone continuity of measures (Proposition 1.5) now implies that $\mu(A_{in}) \uparrow \mu(A_i)$, so that

$$\lim_{n\to\infty} \mu(f_n) \ \geq \ \sum_{i \in I_+} (a_i - \epsilon)\, \mu(A_i).$$

Because the above inequality is true for arbitrarily small $\epsilon > 0$, we may let $\epsilon \to 0$ above to conclude[1] that

$$\lim_{n\to\infty} \mu(f_n) \ \geq \ \sum_{i \in I_+} a_i \mu(A_i) \ \overset{(4.9)}{=} \ \sum_{i \in I} a_i \mu(A_i).$$

(ii) Assume now that $L = \infty$. Then we may fix an index $i \in I$ such that $a_i \mu(A_i) = \infty$. Then $a_i, \mu(A_i) > 0$, and at least one of a_i and $\mu(A_i)$ is infinite. Select $0 < M < a_i$ and $0 < N < \mu(A_i)$, and define $A_{in} = \{s \in A_i \colon f_n(x) > M\}$. For the partition $\{A_{in}, A_{in}^c\}$, we see that

$$\mu(f_n) \ \geq \ \Big(\inf_{s \in A_{in}} f_n(s)\Big)\mu(A_{in}) + \Big(\inf_{s \in A_{in}^c} f_n(s)\Big)\mu(A_{in}^c) \ \geq \ M\mu(A_{in}).$$

Because $f_n \uparrow f$ and $f(s) > M$ for all $s \in A_i$, it follows that $A_{in} \uparrow A_i$ as $n \to \infty$. The monotone continuity of measures (Proposition 1.5) now implies that $\lim_{n\to\infty} \mu(A_{in}) = \mu(A_i) > N$, and we conclude that

$$\lim_{n\to\infty} \mu(f_n) \ \geq \ MN.$$

In the above inequality we may let $M \uparrow a_i$ and $N \uparrow \mu(A_i)$ to conclude that

$$\lim_{n\to\infty} \mu(f_n) \ \geq \ a_i \mu(A_i) \ = \ \infty \ \overset{(L=\infty)}{=} \ \sum_{i \in I} a_i \mu(A_i).$$

(iii) We have verified that (4.8) holds whenever $L \in (0, \infty)$ or $L = \infty$. We also note that (4.8) holds trivially for $L = 0$. We conclude that

$$\lim_{n\to\infty} \mu(f_n) \ \geq \ \sum_{i \in I} \Big(\inf_{s \in A_i} f(s)\Big)\, \mu(A_i)$$

for every finite partition $\{A_i \colon i \in I\}$ of S into measurable sets. Recalling the defining formula (4.3), it follows that $\lim_{n\to\infty} \mu(f_n) \geq \mu(f)$. The claim follows by combining this observation with (4.7). □

[1] This is possible because $a_i, \mu(A_i) < \infty$ for all i.

Proposition 4.6 (Linearity)**.** *For any measurable functions $f, g\colon S \to \bar{\mathbb{R}}_+$, $\mu(af+bg) = a\mu(f) + b\mu(g)$ for all $a, b \in \bar{\mathbb{R}}_+$.*

Proof. (i) Assume first that f, g both have a finite range. Enumerate the ranges by $f(S) = \{x_1, \dots, x_m\}$ and $g(S) = \{y_1, \dots, y_n\}$. Then $f = \sum_{i=1}^m x_i 1_{A_i}$ and $g = \sum_{j=1}^n y_j 1_{B_j}$ where $A_i = f^{-1}(\{x_i\})$ and $B_j = g^{-1}(\{y_j\})$. By Proposition 4.2, we see that

$$\mu(f) \;=\; \sum_i x_i \mu(A_i), \qquad \mu(g) \;=\; \sum_j y_j \mu(B_j).$$

It follows that

$$\begin{aligned} af + bg &= \sum_{i=1}^m a x_i 1_{A_i} + \sum_{j=1}^n b y_j 1_{B_j} \\ &= \sum_{i=1}^m \sum_{j=1}^n a x_i 1_{A_i \cap B_j} + \sum_{i=1}^m \sum_{j=1}^n b y_j 1_{A_i \cap B_j} \\ &= \sum_{i=1}^m \sum_{j=1}^n (a x_i + b y_j) 1_{A_i \cap B_j}. \end{aligned}$$

Because the sets $A_i \cap B_j$ are disjoint, we see by Lemma 4.3 that

$$\begin{aligned} \mu(af + bg) &= \sum_{i=1}^m \sum_{j=1}^n (a x_i + b y_j) \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^m \sum_{j=1}^n x_i \mu(A_i \cap B_j) + b \sum_{i=1}^m \sum_{j=1}^n y_j \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^m x_i \mu(A_i) + b \sum_{j=1}^n y_j \mu(B_j) \\ &= a\mu(f) + b\mu(g). \end{aligned}$$

(ii) Assume now that $f, g\colon S \to \bar{\mathbb{R}}_+$ are general measurable functions. Then by Proposition 4.1, we may select finite-range functions $f_n, g_n\colon S \to \bar{\mathbb{R}}_+$ such that $f_n \uparrow f$ and $g_n \uparrow g$. Then $af_n + bg_n \uparrow af + bg$. Then by part (i), we see that

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

The monotone continuity of integration (Theorem 4.5) then implies that $\mu(f_n) \uparrow \mu(f)$, $\mu(g_n) \uparrow \mu(g)$, and $\mu(af_n + bg_n) \uparrow \mu(af + bg)$. The claim follows by taking limits as $n \to \infty$ in the above equality. □

Integration against a measure μ is a monotone, linear, and monotonely continuous functional on the convex cone of nonnegative measurable functions.

Recall that the sum $\mu + \nu$ of measures μ and ν is a measure (Proposition 1.7). The following result shows how we can integrate against a sum of measures.

Proposition 4.7. *For any measures μ, ν on $(S, \mathcal{S})$ and any measurable $f \colon S \to \bar{\mathbb{R}}_+$,*

$$\int_S f \, d(\mu + \nu) \;=\; \int_S f \, d\mu + \int_S f \, d\nu.$$

Proof. The proof proceeds in three steps: first for indicators functions, then for finite-range functions, and then for general nonnegative functions.

(i) For any measurable set B, the definition of the sum of measures (1.7) implies that

$$\int_S 1_B \, d(\mu+\nu) \;=\; (\mu+\nu)(B) \;=\; \mu(B) + \nu(B) \;=\; \int_S 1_B \, d\mu + \int_S 1_B \, d\nu.$$

Therefore, the claim holds for $f = 1_B$.

(ii) Assume that f has a finite range. Then $f = \sum_{i=1}^n b_i 1_{B_i}$ for some $b_i \in \bar{\mathbb{R}}_+$ and measurable $B_i \subset S$. Then by linearity (Proposition 4.6),

$$\begin{aligned}
\int_S f \, d(\mu+\nu) &\overset{\text{lin}}{=} \sum_{i=1}^n b_i \int_S 1_{B_i} \, d(\mu+\nu) \\
&\overset{\text{(i)}}{=} \sum_{i=1}^n b_i \left(\int_S 1_{B_i} \, d\mu + \int_S 1_{B_i} \, d\nu \right) \\
&\overset{\text{lin}}{=} \int_S \sum_{i=1}^n b_i 1_{B_i} \, d\mu + \int_S \sum_{i=1}^n b_i 1_{B_i} \, d\nu \;=\; \int_S f \, d\mu + \int_S g \, d\nu.
\end{aligned}$$

(iii) Assume that $f \colon S \to \bar{\mathbb{R}}_+$ is a general measurable function. Let f_n be measurable finite-range functions such that $f_n \uparrow f$. Then by monotone continuity (Theorem 4.5),

$$\begin{aligned}
\int_S f \, d(\mu+\nu) &\overset{\text{cont}}{=} \lim_{n\to\infty} \int_S f_n \, d(\mu+\nu) \\
&\overset{\text{(ii)}}{=} \lim_{n\to\infty} \left(\int_S f_n \, d\mu + \int_S f_n \, d\nu \right) \\
&\overset{\text{cont}}{=} \left(\int_S \lim_{n\to\infty} f_n \, d\mu + \int_S \lim_{n\to\infty} f_n \, d\nu \right) \;=\; \int_S f \, d\mu + \int_S f \, d\nu.
\end{aligned}$$

□

Nonnegative integration can be viewed as a functional $(f, \mu) \mapsto \int f \, d\mu$ that is linear in both of its arguments.

We finish this section by the following lemma that is needed in the next section to prove Theorem 4.13.

Lemma 4.8 (Fatou's inequality)**.** *For arbitrary measurable functions $f_n \colon S \to \bar{\mathbb{R}}_+$:*

$$\liminf_{n\to\infty} \mu(f_n) \ \geq \ \mu(\liminf_{n\to\infty} f_n).$$

Proof. Define $g_m = \inf_{n \geq m} f_n$, and note that $g_m \colon S \to \bar{\mathbb{R}}_+$ is measurable.

(i) By definition $g_1 \leq g_2 \leq \cdots$, which implies that the pointwise limit $g = \lim_{m\to\infty} g_m$ is a well-defined function $g \colon S \to \bar{\mathbb{R}}_+$. The monotone continuity of nonnegative integration then implies that

$$\mu(g) \ = \ \lim_{m\to\infty} \mu(g_m). \tag{4.10}$$

(ii) Because $g_m \leq f_n$ for all $m \leq n$, the monotonicity of nonnegative integration implies that $\mu(g_m) \leq \mu(f_n)$ for all $n \geq m$. In particular, $\mu(g_m) \leq \inf_{m \geq n} \mu(f_n)$. By taking limits as $m \to \infty$ and applying (4.10), we conclude that

$$\mu(g) \ = \ \lim_{m\to\infty} \mu(g_m) \ \leq \ \lim_{m\to\infty} \inf_{m \geq n} \mu(f_n) \ = \ \liminf_{n\to\infty} \mu(f_n).$$

The claim follows by observing that $g = \lim_{m\to\infty} g_m = \liminf_{n\to\infty} f_n$. □

4.3 Integrating general functions

The *integral* of a measurable function $f \colon S \to \bar{\mathbb{R}}$ is defined by

$$\int_S f \, d\mu \ = \ \int_S f_+ \, d\mu - \int_S f_- \, d\mu, \tag{4.11}$$

where the *positive part* and the *negative part* of f are defined by

$$f_+ = \max\{f, 0\} \qquad \text{and} \qquad f_- = \max\{-f, 0\},$$

and we use the convention[2] $\infty + (-\infty) = -\infty$. Because f_+ and f_- are nonnegative measurable functions (Exercise 4.19), the integrals on the right side of (4.11) are well defined by formula (4.3).

[2] The value of the integral $\int_S f \, d\mu$ is usually left undefined in the case where $\int_S f_+ \, d\mu$ and $\int_S f_- \, d\mu$ both are infinite. To avoid writing 'when the integral is defined' in later theorems, here we choose to define this value as $-\infty$.

Proposition 4.9 (Monotonicity). *$f \le g \implies \mu(f) \le \mu(g)$ for all measurable functions $f, g \colon S \to \bar{\mathbb{R}}$.*

Proof. Note that $f \le g$ implies that $f_+ \le g_+$ and $f_- \ge g_-$. Proposition 4.4 then implies that $\mu(f_+) \le \mu(g_+)$ and $\mu(g_-) \le \mu(f_-)$. Therefore, in the case where $\min\{\mu(f_+), \mu(f_-)\}$ and $\min\{\mu(g_+), \mu(g_-)\}$ both are finite,

$$\mu(f) \;=\; \mu(f_+) - \mu(f_-) \;\le\; \mu(g_+) - \mu(g_-) \;=\; \mu(g).$$

For the remaining cases, we note that:

(i) If $\min\{\mu(f_+), \mu(f_-)\} = \infty$, then $\mu(f_-) = \infty$.

(ii) If $\min\{\mu(g_+), \mu(g_-)\} = \infty$, then $\mu(g_-) = \infty$ implies that $\mu(f_-) = \infty$.

The convention $x + (-\infty) = -\infty$ for all $x \in \bar{\mathbb{R}}$ now implies that $\mu(f) = -\infty$ in both of the above cases. Therefore, $\mu(f) \le \mu(g)$. □

A measurable function $f \colon S \to \bar{\mathbb{R}}$ is called *intϵgrable* against μ if the value of the integral $\int_S f \, d\mu$ is an ordinary real number.

Proposition 4.10. *A measurable function $f \colon S \to \bar{\mathbb{R}}$ is integrable if and only if $\int_S |f| \, d\mu < \infty$.*

Proof. The definition together with the convention $\infty - \infty = -\infty$ implies that $\int_S f \, d\mu \in \mathbb{R}$ if and only if $\int_S f_+ \, d\mu$ and $\int_S f_- \, d\mu$ both are finite. Because $|f| = f_+ + f_-$, Proposition 4.6 implies that $\int_S |f| \, d\mu = \int_S f_+ \, d\mu + \int_S f_- \, d\mu$. Therefore, $\int_S f \, d\mu \in \mathbb{R}$ if and only if $\int_S |f| \, d\mu < \infty$. □

The integral of a measurable function is an extended real number in $\bar{\mathbb{R}}$. Integrable functions are those for which the integral is an ordinary real number in $\mathbb{R}$.

In most applications our focus will be on real-valued functions for which the integrals are real valued. The space of such functions is denoted by

$$L^1(\mu) \;=\; \left\{ f \colon S \to \mathbb{R} \text{ measurable} \colon \int_S |f| \, d\mu < \infty \right\}.$$

Proposition 4.11 (Linearity). *$L^1(\mu)$ is a real vector space, in which the equality $\mu(af+bg) = a\mu(f)+b\mu(g)$ holds for all $f, g \in L^1(\mu)$ and $a, b \in \mathbb{R}$.*

Proof. (i) Fix $f, g \in L^1(\mu)$ and $a, b \in \mathbb{R}$. Denote $h = af + bg$. Then $|h| \le |a||f| + |b||g|$. By applying Proposition 4.9 and then Proposition 4.6, we find that

$$\mu(|h|) \ \le \ \mu(|a||f|+|b||g|) \ \le \ |a|\mu(|f|)+|b|\mu(|g|) \ < \ \infty.$$

Hence $h \in L^1(\mu)$ and we conclude that $L^1(\mu)$ is a vector space.

(ii) We will verify that $\mu(h) = \mu(f) + \mu(g)$ for $h = f + g$ and for all $f, g \in L^1(\mu)$. By writing $f = f_+ - f_-$ and $g = g_+ - g_-$ and $h = h_+ - h_-$, we find that

$$h_+ - h_- \ = \ f_+ - f_- + g_+ - g_-.$$

By rearranging this equality, we see that

$$h_+ + f_- + g_- \ = \ h_- + f_+ + g_+.$$

The above equality only involves nonnegative functions. Therefore, we may apply Proposition 4.6 to conclude that

$$\mu(h_+) + \mu(f_-) + \mu(g_-) \ = \ \mu(h_-) + \mu(f_+) + \mu(g_+).$$

By (i) we know that $\mu(|h|) < \infty$. Because $f_+, f_- \le |f|$ and $g_+, g_- \le |g|$ and $h_+, h_- \le |h| \le |f| + |g|$, we conclude by Proposition 4.9 that all integrals in the above equality are finite. Hence by rearranging, we find that

$$\mu(h_+) - \mu(h_-) \ = \ \mu(f_+) - \mu(f_-) + \mu(g_+) - \mu(g_-).$$

By definition, the above equality can be rewritten as $\mu(h) = \mu(f) + \mu(g)$.

(iii) We claim that $\mu(cf) = c\mu(f)$ for all $c \in \mathbb{R}$ and $f \in L^1(\mu)$. Assume first that $c \ge 0$. Then $(cf)_+ = cf_+$ and $(cf)_- = cf_-$. Hence

$$\mu(cf) \ = \ \mu(cf_+) - \mu(cf_-) \ = \ c(\mu(f_+) - \mu(f_-)) \ = \ c\mu(f).$$

If $c < 0$, then $(cf)_+ = |c|f_-$ and $(cf)_- = |c|f_+$, and it follows that

$$\mu(cf) \ = \ \mu(|c|f_-) - \mu(|c|f_+) \ = \ |c|(\mu(f_-) - \mu(f_+)) \ = \ -|c|\mu(f) \ = \ c\mu(f).$$

□

Integration is a monotone linear functional on the vector space of integrable functions.

Proposition 4.12 (Triangle inequality)**.** *$|\mu(f)| \le \mu(|f|)$ for any measurable function $f\colon S \to \bar{\mathbb{R}}$.*

Proof. Because $|f| = f_+ + f_-$, Proposition 4.6 implies that

$$\mu(|f|) \;=\; \mu(f_+) + \mu(f_-).$$

By definition,

$$\mu(f) \;=\; \begin{cases} \mu(f_+) - \mu(f_-), & \mu(f_+) < \infty,\ \mu(f_-) < \infty, \\ \infty, & \mu(f_+) = \infty,\ \mu(f_-) < \infty, \\ -\infty, & \mu(f_+) < \infty,\ \mu(f_-) = \infty, \\ -\infty, & \mu(f_+) = \infty,\ \mu(f_-) = \infty. \end{cases}$$

When $\mu(f_+)$ and $\mu(f_-)$ both are finite, the inequality

$$|\mu(f_+) - \mu(f_-)| \;\le\; |\mu(f_+)| + |\mu(f_-)| \;=\; \mu(f_+) + \mu(f_-)$$

implies that $|\mu(f)| \le \mu(|f|)$. In the remaining cases the claim follows immediately. □

New result, added 7 Oct 2024 The following result confirms another continuity property of integration, in which the convergence of the integrands need not be monotone.

Theorem 4.13 (Dominated continuity)**.** *For arbitrary measurable functions $f_n, f, g\colon S \to \bar{\mathbb{R}}$ such that $|f_n| \le g$ for all n and $\int_S g\, d\mu < \infty$:*

$$f_n \to f \quad \Longrightarrow \quad \int_S f_n\, d\mu \to \int_S f\, d\mu.$$

Proof. (i) We will first prove the theorem under an extra assumption that

$$g(x) \in \mathbb{R} \quad \text{for all } x \in S. \tag{4.12}$$

This guarantees that f_n, f, g are real-valued functions, and we may add and subtract these pointwise without worrying about infinite values. Because $|f_n| \le g$, the monotonicity of nonnegative integration implies that f_n is integrable. The same is true for f, because $|f| = \lim_{n\to\infty} |f_n| \le |g|$ pointwise. Hence f_n, f, g are all in $L^1(\mu)$, and we may add and subtract their integrals as ordinary real numbers. Because the functions $g \pm f_n$ are nonnegative, Fatou's inequality (Lemma 4.8) and the linearity of integration imply that

$$\begin{aligned} \mu(g) + \liminf_{n\to\infty} \mu(\pm f_n) &= \liminf_{n\to\infty} \mu(g \pm f_n) \\ &\ge \mu(\liminf_{n\to\infty}(g \pm f_n)) \\ &= \mu(g \pm f) \\ &= \mu(g) \pm \mu(f). \end{aligned}$$

By subtracting $\mu(g)$ from both sides, we conclude that

$$\liminf_{n\to\infty} \mu(\pm f_n) \ \ge \ \pm\mu(f).$$

By noting that $\liminf_{n\to\infty} \mu(-f_n) = -\limsup_{n\to\infty} \mu(f_n)$, the above pair of inequalities can be rewritten as

$$\mu(f) \ \le \ \liminf_{n\to\infty} \mu(f_n) \ \le \ \limsup_{n\to\infty} \mu(f_n) \ \le \ \mu(f).$$

All quantities in the above display must be equal to each other, and we conclude that $\mu(f) = \lim_{n\to\infty} \mu(f_n)$.

(ii) Let us now get rid of the extra assumption (4.12). Because $\int g \, d\mu < \infty$, we know (Theorem 5.2) that the complement of $\tilde{S} = \{x \in S \colon g(x) \in \mathbb{R}\}$ has zero measure. Observe next that the functions $\tilde{f} = f 1_{\tilde{S}}$, $\tilde{f}_n = f_n 1_{\tilde{S}}$, and $\tilde{g} = g 1_{\tilde{S}}$ satisfy the assumptions of the theorem, and that the function $\tilde{g}$ is real-valued. Hence by (i), we conclude that

$$\mu(\tilde{f}) = \lim_{n\to\infty} \mu(\tilde{f}_n).$$

By insensitivity of integration (Proposition 5.3), we see that $\mu(\tilde{f}) = \mu(f)$ and $\mu(\tilde{f}_n) = \mu(f_n)$. Hence the claim follows from (i). □

4.4 Expectations

The *expectation* of a random variable $X\colon \Omega \to \bar{\mathbb{R}}$ on probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is defined as the integral

$$\mathbb{E}X \ = \ \int_\Omega X \, d\mathbb{P}.$$

A random variable is called *integrable* if its expectation is a well-defined real number, or equivalently (Proposition 4.10), if $\mathbb{E}|X| < \infty$. The space of real-valued integrable defined on probability space $(\Omega, \mathcal{A}, \mathbb{P})$ random variables is denoted by $L^1(\mathbb{P})$.

Theorem 4.14. *The expectation operation on the space of $\bar{\mathbb{R}}_+$-valued random variables has the following properties.*

(i) Linearity: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ *for all* $a, b \in \bar{\mathbb{R}}_+$.

(ii) Monotonicity: $X \le Y \implies \mathbb{E}X \le \mathbb{E}Y$.

(iii) Monotone continuity: $X_n \uparrow X \implies \mathbb{E}X_n \uparrow \mathbb{E}X$.

Proof. (i) follows from Proposition 4.4, (ii) from Proposition 4.6, and (iii) from Theorem 4.5. □

The expectation operator is monotone, linear, and monotonically continuous functional on the convex cone of nonnegative random variables.

Theorem 4.15. *The expectation operation on the space of $\mathbb{R}$-valued integrable random variables has the following properties.*

(i) Linearity: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ *for all* $a, b \in \mathbb{R}$.

(ii) Monotonicity: $X \le Y \implies \mathbb{E}X \le \mathbb{E}Y$.

(iii) Dominated continuity: If $|X_n| \le Y$ *for all* n *with* $\mathbb{E}Y < \infty$, *then* $X_n \to X \implies \mathbb{E}X_n \to \mathbb{E}X$.

(iv) Bounded continuity: If $|X_n| \le c$ *for all* n *for some constant* $c < \infty$, *then* $X_n \to X \implies \mathbb{E}X_n \to \mathbb{E}X$.

Proof. Property (i) and the fact that $L^1(\mathbb{P})$ is a real vector space follow from Proposition 4.11. Property (ii) follows from Proposition 4.9, and (iii) from Proposition 4.13. Property (iv) follows from (iii) by interpreting c as a constant random variable such that $Y(\omega) = c$ for all ω. □

The expectation operator is monotone linear functional on the vector space of integrable random variables.

4.5 Exercises

Exercise 4.16 (Down-rounding is measurable)**.** Fix an integer $n \ge 1$, and consider the down-rounding operation $\rho_n \colon \bar{\mathbb{R}}_+ \to \bar{\mathbb{R}}_+$ defined by (4.1).

(a) Determine the range of ρ_n, that is, the set $\rho_n(\bar{\mathbb{R}}_+) = \{\rho_n(x) \colon x \in \bar{\mathbb{R}}_+\}$.

(b) Determine the preimage $\rho_n^{-1}(\{q\}) = \{x \in \bar{\mathbb{R}}_+ \colon \rho_n(x) = q\}$ for each value q in the range ρ_n.

(c) By applying (a) and (b), prove that ρ_n is a measurable function.

Hint: The formula $f^{-1}(B) = \bigcup_{b \in B \cap f(S)} f^{-1}(\{b\})$, valid all functions $f \colon S \to T$ and all sets $B \subset T$, may be helpful.

Exercise 4.17 (Extended domain of a measure)**.** Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces such that $S \subset T$ and $A \cap S \in \mathcal{S}$ for all $A \in \mathcal{T}$. We define an *extension* of a measure μ on $(S, \mathcal{S})$ as a set function $\tilde{\mu} \colon \mathcal{T} \to \bar{\mathbb{R}}_+$ given by

$$\tilde{\mu}(A) = \mu(A \cap S), \qquad A \in \mathcal{T}.$$

(a) Prove that $\tilde{\mu}$ is a measure on $(T, \mathcal{T})$

(b) If μ is a probability measure on $(S, \mathcal{S})$, does it follow that $\tilde{\mu}$ is a probability measure on $(T, \mathcal{T})$?

(c) Let $S = \{0, 1\}$ and $\mathcal{S} = 2^S$, and define a measure μ on $(S, \mathcal{S})$ by

$$\mu(A) = \begin{cases} 0, & A = \emptyset, \\ 1 - p, & A = \{0\}, \\ p, & A = \{1\}, \\ 1, & A = \{0, 1\}. \end{cases}$$

where $p \in [0, 1]$. Let $\tilde{\mu}$ be the extension of μ to the measurable space $(T, \mathcal{T}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Write down constants c_0, c_1 such that $\tilde{\mu} = c_0 \delta_0 + c_1 \delta_1$, where δ_x is the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by $\delta_x(A) = 1_A(x)$, $A \in \mathcal{B}(\mathbb{R})$.

Exercise 4.18 (Alternative integral definition)**.** In some textbooks the integral of a measurable function $f \colon S \to \bar{\mathbb{R}}_+$ is defined by $\mu(f) = \lim_{n \to \infty} \mu(f_n)$ where $f_n \colon S \to \bar{\mathbb{R}}_+$ are arbitrary measurable finite-range functions such that $f_n \uparrow f$, and $\mu(f_n)$ for a finite-range function f_n is defined by (4.4). Prove that this definition agrees with the definition in (4.3).

Exercise 4.19 (Positive and negative parts are measurable)**.** Let $f \colon S \to \bar{\mathbb{R}}$ be a measurable function. Prove that f_+, f_- are measurable functions. You may proceed for example as follows.

(i) Write $f_+ = \rho_+ \circ f$ where $\rho_+ \colon \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ is defined by $\rho_+(x) = \max\{x, 0\}$.

(ii) Check that the preimages $\{x \in \bar{\mathbb{R}} \colon \rho_+(x) \leq t\}$ are measurable for every $t \in \bar{\mathbb{R}}_+$.

(iii) Conclude that ρ_+ is a measurable function.

(iv) Conclude that f is a measurable function.

4.6 Historical notes

Henri Lebesgue (1875–1941). PhD 1902 from Sorbonne for Émile Borel (Lebesgue was 27, Borel was 31 during that time), seminal thesis *Intégrale, longueur, aire.* In Lebesgue's thesis, original terms include *une mesure*, *un ensemble mesurable*, and so on. He did not speak of sigma-algebras, but rather measurable sets (*ensemble mesurable*) which now are called Lebesgue measurable sets.

Chapter 5

Probability densities

The probability distribution of a random variable is usually represented in terms of a probability density function with respect to a reference measure. Usually, but not alway, the reference measure is the chosen as the Lebesgue measure. In this chapter we learn what probability distributions do admit a probability density function, and how to compute expected values and probabilities using densities.

Key concepts: probability density function, almost sure event

Learning outcomes:

- Learn to compute expectations of random variables by integrating 'against the law'.
- Learn to construct probability measures as measures weighted by integrable functions.
- Learn how integrate against the Lebesgue measure in practice.
- Recognise when the modern definition of the integral corresponds to classical high-school definition of the integral.

Prerequisites: Previous chapters.

5.1 Almost sure properties

Let C be a property concerning the points of a measurable space $(S, \mathcal{S})$ that is equipped with a measure μ. We say that C holds for *μ-almost every* $s \in S$, or *μ-almost everywhere*, if the set $C = \{s \in S \colon s \text{ has property } \mathsf{C}\}$ is measurable and $\mu(C^c) = 0$. We say that a property concerning the points of a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ holds *almost surely*, if it holds for $\mathbb{P}$-almost every $\omega \in \Omega$.

Theorem 5.1. *The integral of a measurable function $f \colon S \to \bar{\mathbb{R}}_+$ against a measure μ equals $\int_S f \, d\mu = 0$ if and only if $\mu(\{f \neq 0\}) = 0$.*

For nonnegative functions, $\int_S f \, d\mu = 0$ if and only if $f = 0$ μ-almost everywhere.

For any random variable, $\mathbb{E}|X| = 0$ if and only if $X = 0$ almost surely.

Proof. Fix an integer $n \geq 1$, and define functions $f_n, g \colon S \to \bar{\mathbb{R}}_+$ by formulas $f_n = n^{-1} 1_{\{f \geq n^{-1}\}}$ and $g = \infty \cdot 1_{\{f > 0\}}$, so that

$$f_n(s) = \begin{cases} 0, & f(s) < n^{-1}, \\ n^{-1}, & f(s) \geq n^{-1}, \end{cases} \qquad \text{and} \qquad g(s) = \begin{cases} 0, & f(s) = 0, \\ \infty, & f(s) > 0. \end{cases}$$

Because f_n, g are measurable with finite ranges, Proposition 4.2 implies that

$$\mu(f_n) \; = \; 0 \cdot \mu(\{f < n^{-1}\}) + n^{-1} \mu(\{f \geq n^{-1}\}) \; = \; n^{-1} \mu(\{f \geq n^{-1}\}).$$

and

$$\mu(g) \; = \; 0 \cdot \mu(\{f = 0\}) + \infty \cdot \mu(\{f > 0\}) \; = \; \infty \cdot \mu(\{f > 0\}).$$

Because $f_n \leq f \leq g$, we see that monotonicity (Proposition 4.4) implies that $\mu(f_n) \leq \mu(f) \leq \mu(g)$. Therefore,

$$n^{-1} \mu(\{f \geq n^{-1}\}) \; \leq \; \mu(f) \; \leq \; \infty \cdot \mu(\{f > 0\}). \tag{5.1}$$

If $\mu(\{f > 0\}) = 0$, then the inequality on the right side of (5.1) implies that $\mu(f) = 0$.

Assume now that $\mu(f) = 0$. Then the inequality on the left side of (5.1) implies that $\mu(\{f \geq n^{-1}\}) = 0$ for all integers $n \geq 1$. Because $\{f \geq n^{-1}\} \uparrow \{f > 0\}$, the monotone continuity of measures (Proposition 1.5) implies that

$$\mu(\{f > 0\}) \; = \; \lim_{n \to \infty} \mu(\{f > n^{-1}\}) \; = \; 0. \qquad \square$$

New result added 4 Oct 2024

Theorem 5.2. *If the integral of a measurable function $f\colon S \to \bar{\mathbb{R}}_+$ against a measure μ satisfies $\int_S f\,d\mu < \infty$, then $\mu(\{f = \infty\}) = 0$.*

For any $\bar{\mathbb{R}}_+$-valued random variable, $\mathbb{E}|X| < \infty$ implies that $X < \infty$ almost surely.

Proof. Because $f \ge 0$, we find that

$$f \;=\; f1_{\{f<\infty\}} + \infty 1_{\{f=\infty\}} \;\ge\; \infty 1_{\{f=\infty\}}.$$

The monotonicity of nonnegative integration then implies that

$$\int_S f\,d\mu \;\ge\; \int_S \left(\infty 1_{\{f=\infty\}}\right) d\mu \;=\; \infty\mu(\{f=\infty\}).$$

Because $\int_S f\,d\mu$ is finite, we conclude that $\mu(\{f = \infty\}) = 0$. □

The integral of a function restricted to a measurable set $A \subset S$ is defined by $\int_A f\,d\mu = \int_S 1_A f\,d\mu$.

Proposition 5.3 (Insensitivity of integration)**.** *Let C be a measurable set such that $\mu(C^c) = 0$.*

(i) $\mu(A) = \mu(A \cap C)$ for all measurable sets A.

(ii) $\int_S f\,d\mu = \int_C f\,d\mu$ for all nonnegative measurable functions f.

(iii) $\int_S f\,d\mu = \int_C f\,d\mu$ for all absolutely integrable functions f.

When computing the measure of an event or the integral of a function, we may restrict to a set C with $\mu(C^c) = 0$ without affecting the result.

Proof. (i) Because $\mu(A \cap C^c) \le \mu(C^c) = 0$, we conclude that $\mu(A \cap C^c) = 0$. Because the sets $A \cap C$ and $A \cap C^c$ are disjoint, it follows that

$$\mu(A) \;=\; \mu((A\cap C) \cup (A \cap C^c)) \;=\; \mu(A\cap C) + \mu(A \cap C^c) \;=\; \mu(A \cap C).$$

(ii) Let $f\colon S \to \bar{\mathbb{R}}_+$ be measurable. Because the sets C and C^c are disjoint, we find that $f = f1_C + f1_{C^c}$ is a sum of measurable nonnegative functions. By linearity of integration (Proposition 4.6), we find that

$$\int_S f\,d\mu \;=\; \int_S f1_C\,d\mu + \int_S f1_{C^c}\,d\mu. \tag{5.2}$$

Observe next that $f(s)1_{C^c}(s) > 0$ can occur only if $s \in C^c$, so that

$$\{f1_{C^c} > 0\} \ \subset \ C^c.$$

Therefore, $\mu(\{f1_{C^c} > 0\}) = 0$, and Theorem 5.1 implies that $\int_S f1_{C^c}\, d\mu = 0$. The second claim now follows from (5.2).

(iii) Consider now an absolutely integrable measurable function $f\colon S \to \bar{\mathbb{R}}$. Then $\mu(|f1_C|) \le \mu(|f|) < \infty$ shows that also $f1_C$ is absolutely integrable. Then $\mu(f) = \mu(f_+) - \mu(f_-)$, in which both integrals on the right side are finite. By (ii), we know that $\mu(f_+) = \mu(f_+1_C)$ and $\mu(f_-) = \mu(f_-1_C)$. Hence

$$\begin{aligned} \mu(f1_C) &= \mu((f1_C)_+) - \mu((f1_C)_-) \\ &= \mu(f_+1_C) - \mu(f_-1_C) \\ &= \mu(f_+) - \mu(f_-) \\ &= \mu(f). \end{aligned}$$

□

5.2 Weighted measures and densities

The *weighting* of a measure ν on $(S, \mathcal{S})$ by a measurable function $f\colon S \to \bar{\mathbb{R}}_+$ produces a set function $\mu\colon \mathcal{S} \to \bar{\mathbb{R}}_+$ defined[1] by

$$\mu(A) \ = \ \int_A f\, d\nu. \tag{5.3}$$

When (5.3) holds for all $A \in \mathcal{S}$, we say that f is a *density function*[2] of μ with respect to ν, and abbreviate this by writing $d\mu = f d\nu$.

Proposition 5.4. *For any measure ν on $(S, \mathcal{S})$ and any measurable function $f\colon S \to \bar{\mathbb{R}}_+$, the set function μ defined by (5.3) is a measure on $(S, \mathcal{S})$. The measure μ is a probability measure if and only if $\int_S f\, d\nu = 1$.*

Weighting by a nonnegative function f provides a mechanism for constructing new measures from an existing measure ν. The weighted measure is often abbreviated as $f d\nu$.

[1] Recall that $\int_B f d\nu = \int_S 1_B f d\nu$.

[2] In real analysis, density functions are called Radon–Nikodym derivatives.

Proof. (i) Let us first verify that μ is a well-defined set function. For any measurable set $A \subset S$, the function 1_A is measurable, and the same is true for f by our assumption. Hence also the function $1_A f$ is measurable, and the integral $\int_A f \, d\nu = \int_S 1_A f \, d\nu$ on the right side of (5.3) is well defined. Hence μ is a well-defined set function from $\mathcal{S}$ into $\bar{\mathbb{R}}_+$.

(ii) Because $1_\emptyset(s) f(s) = 0$ for all $s \in S$, we see that the function $1_\emptyset f$ is identically zero, and it follows that $\mu(\emptyset) = \int_S 1_\emptyset f \, d\nu = 0$.

(iii) To verify disjoint countable additivity, let $A_1, A_2, \dots$ be disjoint measurable sets. Denote $B_n = \cup_{k=1}^n A_k$. Then $1_{B_n} = \sum_{k=1}^n 1_{A_k}$ implies that $1_{B_n} f = \sum_{k=1}^n 1_{A_k} f$. The linearity of integration (Proposition 4.6) implies that

$$\mu(B_n) \;=\; \int_S 1_{B_n} f \, d\nu \;=\; \sum_{k=1}^n \int_S 1_{A_k} f \, d\nu \;=\; \sum_{k=1}^n \mu(A_k).$$

Because $B_n \uparrow \cup_{k=1}^\infty A_k$, the monotone continuity of measures (Proposition 1.5) implies that

$$\mu\Big(\bigcup_{k=1}^\infty A_k \Big) \;=\; \lim_{n\to\infty} \mu(B_n) \;=\; \lim_{n\to\infty} \sum_{k=1}^n \mu(A_k) \;=\; \sum_{k=1}^\infty \mu(A_k).$$

Hence μ is countably disjointly additive, and we conclude that μ is a measure.

(iv) Finally, we note that $\mu(S) = \int_S f \, d\nu$, so that the measure μ is a probability measure if and only if $\int_S f \, d\nu = 1$. □

Todo: Lebesgue integral is the ordinary integral.

Example 5.5 (Normal distribution)**.** The probability measure $\mu = f \, d\lambda$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is called the *standard normal distribution*.

Example 5.6 (Exponential distribution)**.** The probability measure $\mu = f \, d\lambda$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with $f(x) = 1_{(0,\infty)} a e^{-ax}$ is called the *exponential distribution* with parameter $a \in (0, \infty)$.

Example 5.7 (Binomial distribution)**.** The probability measure $\mu = f \, d\#$ on $(\mathbb{Z}, 2^{\mathbb{Z}})$ with $f(x) = 1_{[0,n]}(x) \binom{n}{x} (1-p)^{n-x} p^x$ is called the *binomial distribution* with trial count $n \geq 1$ and success rate $p \in [0, 1]$.

Todo: Examples with no density

5.3 Integrating using densities

The following result demonstrates how to integrate against a weighted measure.

Proposition 5.8. *The integral of a measurable function $g\colon S \to \bar{\mathbb{R}}_+$ against the weighted measure $d\mu = f d\nu$ defined by (5.3) can be computed according to*

$$\int_S g(x)\,\mu(dx) \;=\; \int_S g(x) f(x)\,\nu(dx). \tag{5.4}$$

Rewriting (5.4) as $\int g\,d\mu = \int g f\,d\nu$ shows that integration against μ is converted to integration against ν by symbolically replacing '$d\mu \mapsto f d\nu$'. This motivates abbreviating the f-weighted modification of ν by $d\mu = f d\nu$.

Proof. We employ a common proof technique in probability theory, where the claim is proved in three stages: first for indicator functions, then for finite-range functions, and finally for general functions.

(i) Assume that $g = 1_A$ for some measurable set A. Then

$$\int_S g\,d\mu \;=\; \int_S 1_A\,d\mu \;=\; \mu(A) \;=\; \int_A f\,d\nu \;=\; \int_S 1_A f\,d\nu.$$

Hence (5.4) holds for indicator functions g.

(ii) Assume next that g has a finite range enumerated as $\{a_1, \dots, a_n\}$. The function g may then be represented as $g = \sum_{k=1}^n a_k 1_{A_k}$ with $A_k = g^{-1}(\{a_k\})$. Then by (i) and the linearity of integration, we see that

$$\int_S \underbrace{\sum_{k=1}^n a_k 1_{A_k}}_{g}\,d\mu = \sum_{k=1}^n a_k \int_S 1_{A_k}\,d\mu \overset{\text{(i)}}{=} \sum_{k=1}^n a_k \int_S 1_{A_k} f\,d\nu = \int_S \underbrace{\sum_{k=1}^n a_k 1_{A_k}}_{g} f\,d\nu.$$

so that (5.4) holds for all measurable finite-range functions $g\colon S \to \bar{\mathbb{R}}_+$.

(iii) Let $g\colon S \to \bar{\mathbb{R}}_+$ be measurable. Let g_n be the down-rounded and truncated modification of g as defined in (4.2). Proposition 4.1 then implies that $g_n \uparrow g$. Because f is nonnegative, we also see that $f g_n \uparrow f g$. Monotone continuity of integration (Theorem 4.5) then implies that $\int_S g_n\,d\mu \uparrow \int_S g\,d\mu$ and $\int_S f g_n\,d\nu \uparrow \int_S f g\,d\nu$. By applying (ii), we may now conclude that

$$\int_S g\,d\mu \;=\; \lim_{n\to\infty} \int_S g_n\,d\mu \overset{\text{(ii)}}{=} \lim_{n\to\infty} \int_S f g_n\,d\nu \;=\; \int_S f g\,d\nu.$$

□

5.4 Almost uniqueness

Proposition 5.9. *If f is a density of a measure μ with respect to a measure ν, then so is every measurable function $\tilde{f}\colon S \to \bar{\mathbb{R}}_+$ such that $\tilde{f}(s) = f(s)$ for ν-almost every $s \in S$.*

Proof. Let f be a density function of μ with respect to ν. Let $C = \{s \in S\colon f(s) = \tilde{f}(s)\}$. Then $\nu(C^c) = 0$. Proposition 5.3 then implies that for any measurable set A,

$$\int_A \tilde{f}\, d\nu \;=\; \int_S 1_A \tilde{f}\, d\nu \;=\; \int_C 1_A \tilde{f}\, d\nu \;=\; \int_C 1_A f\, d\nu \;=\; \int_S 1_A f\, d\nu \;=\; \mu(A).$$

Hence $\tilde{f}$ is also a density of μ with respect to ν. □

Density functions and are usually not unique: A density function remains a density function when its values are modified in a set of reference measure zero.

The following result shows that a density function, when it exists, is almost unique. The proof requires a mild regularity condition. A measure ν on $(S, \mathcal{S})$ is called *sigma-finite* if there exist measurable sets $S_n \uparrow S$ such that $\nu(S_n) < \infty$ for all integers $n \geq 1$. All finite measures, and in particular, all probability measures, are sigma-finite. So is the Lebesgue measure, because $\lambda(S_n) = 2n < \infty$ for $S_n = [-n, n]$.

Proposition 5.10. *Assume that f and g are densities of a measure μ with respect to a sigma-finite measure ν. Then $f(s) = g(s)$ for ν-almost every s.*

This could be generalised as: Assume that nonnegative measurable functions f, g satisfy $\int_B f\, d\nu = \int_B g\, d\nu$ for all measurable B. Then $f(s) = g(s)$ for ν-almost every s.

Proof. Fix some measurable sets $S_n \uparrow S$ such that $\nu(S_n) < \infty$, and define $h_n = 1_{B_n}(g - f)$, where

$$B_n \;=\; \{s \in S_n\colon g(s) > f(s),\; f(s) \leq n\}.$$

The assumption that f and g are densities of the same measure implies that

$$\int_S 1_{B_n} f\, d\nu \;=\; \int_S 1_{B_n} g\, d\nu \tag{5.5}$$

with both sides being equal to $\mu(B_n)$. Because $f(s) \le n$ for all $s \in B_n$, we see that

$$\int_S 1_{B_n} f \, d\nu \;\le\; \int_S 1_{B_n} n \, d\nu \;=\; n\nu(B_n) \;\le\; n\nu(S_n) \;<\; \infty,$$

so that $1_{B_n} f \in L^1(\nu)$. Equality (5.5) shows that also $1_{B_n} g \in L^1(\nu)$. Because $L^1(\nu)$ is a vector space (Proposition 4.11), we see that $h_n = 1_{B_n} g - 1_{B_n} f \in L^1(\nu)$, and

$$\int_S h_n \, d\nu \;=\; \int_S 1_{B_n} g \, d\nu - \int_S 1_{B_n} f \, d\nu \;\overset{(5.5)}{=}\; 0.$$

The definition of B_n implies that h_n is nonnegative. Therefore, Proposition 5.1 implies that $\nu(\{h_n > 0\}) = 0$. Because $h_n(s) > 0$ for all $s \in B_n$, we see that $B_n \subset \{h_n > 0\}$, and therefore $\nu(B_n) = 0$ by monotonicity of measures (Proposition 1.4).

Finally, we note that $B_n \uparrow \{g > f\}$, so the monotone continuity of measures (Proposition 1.5) implies that

$$\nu(\{g > f\}) \;=\; \lim_{n\to\infty} \nu(B_n) \;=\; 0.$$

A symmetric argument where we swap the roles of f and g shows that $\nu(\{f > g\}) = 0$. As a conclusion, $\nu(\{f \ne g\}) = \nu(\{g > f\}) + \nu(\{f > g\}) = 0$. □

Example 5.11 (Discrete probability densities)**.** Recall from Section 1.6 that a probability mass function on a countable set S is a function $f \colon S \to [0,1]$ such that $\sum_{x\in S} f(x) = 1$. We say in Proposition 1.9 that the formula

$$P(A) \;=\; \sum_{x\in A} f(x)$$

defines a probability measure. By rewriting the above equation as $P(A) = \int_A f \, d\#$, we see that P is a f-weighted modification of the counting measure $\#$ on $(S, 2^S)$. Hence the probability mass function f is a probability density function of P with respect to the counting measure.

Probability mass functions are probability density functions with respect to the counting measure.

5.5 Probability density functions

Let μ and ν be measures on a measurable space $(S, \mathcal{S})$. A measurable function $f \colon S \to \bar{\mathbb{R}}_+$ is called a *density function* of μ with respect to ν if

$$\mu(A) \;=\; \int_A f \, d\nu \qquad \text{for all } A \in \mathcal{S}. \tag{5.6}$$

A density function of a probability measure μ is called a *probability density function*. Observe that (5.6) coincides with (5.3) used to define weighted measures. Therefore, f being a density function means that μ equals the f-weighted modification of ν.

The measure ν is often called a reference measure.

Not all measures admit a density function with respect to a given reference measure ν. Even when they do, a density function might not be unique.

Weighted measures defined using a weight function integrating to one yield probability measures. Let ν be a measure on $(S, \mathcal{S})$ and let $f \colon S \to [0, \infty]$ be a measurable function such that $\int_S f \, d\nu = 1$. Proposition 5.4 implies that $\mu(B) = \int_B f \, d\nu$ is a probability measure on $(S, \mathcal{S})$. We say that f is a *density* of μ with respect to reference measure ν. Note that ν does not need to be a finite measure.

5.6 Integrating against the Lebesgue measure

A bounded function $f \colon [a, b] \to \mathbb{R}$ is called *Riemann-integrable* when its *Riemann integral*

$$\int_a^b f(x) \, dx,$$

exists as a well-defined real number. The following key result tells us how the classical Riemann integral coincides with the modern definition of the integral against the Lebesgue measure λ on the real line.

Theorem 5.12. *A bounded measurable function $f \colon [a, b] \to \mathbb{R}$ is Riemann integrable if and only if f is continuous at λ-almost every point $[a, b]$, and in this case*

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

Todo: Put proof to appendix, can be omitted from main text.

Proof. Let $f \colon [a, b] \to \mathbb{R}$ be a bounded measurable function. Then there exists [Rud76, Theorem 11.33] a nested sequence $\mathcal{A}_n$ of finite partitions of $[a, b]$ into intervals such that the lower and upper Riemann integrals of f equal

$$R_-(f) = \lim_{n \to \infty} \lambda(L_n), \qquad R_+(f) = \lim_{n \to \infty} \lambda(U_n), \tag{5.7}$$

where $L_n, U_n \colon [a,b] \to \mathbb{R}$ are finite-range measurable functions given by

$$L_n = \sum_{A \in \mathcal{A}_n} \Big(\inf_{s \in A} f(s) \Big) 1_A, \qquad U_n = \sum_{A \in \mathcal{A}_n} \Big(\sup_{s \in A} f(s) \Big) 1_A.$$

Then functions L_n, U_n are bounded by

$$L_1 \le L_2 \le \cdots \le f \le \cdots \le U_2 \le U_1. \tag{5.8}$$

As a consequence, the pointwise limits $L = \lim_n L_n$ and $U = \lim_n U_n$ are well defined bounded real functions on $[a,b]$.

Because pointwise limits of measurable functions are measurable (Proposition 3.9), we see that $L, U \colon [a,b] \to \mathbb{R}$ are measurable. We also see from (5.8) that $L \le f \le U$, so that the monotonicity (Proposition 4.9) implies that $\lambda(L) \le \lambda(f) \le \lambda(U)$. Because $L_n \uparrow L$ and $U_n \downarrow U$, monotone continuity (Proposition 4.5) implies that $\lambda(L_n) \uparrow \lambda(L)$ and $\lambda(U_n) \downarrow \lambda(U)$. In light (5.7), we conclude that

$$R_-(f) = \lambda(L), \qquad R_+(f) = \lambda(U). \tag{5.9}$$

Let C be the set of points in $[a,b]$ at which f is continuous, and let B the set of points that appear as a boundary point of an interval in $\mathcal{A}_n$ for some n. By inspecting the proof [Rud76, Theorem 11.33], one may check that for any $x \in B^c$, $x \in C$ if and only if $L(x) = U(x)$. Therefore, $C \cap B^c = \{L = U\} \cap B^c$. With a little work one may check that C is a measurable set (Exercise 5.21). Furthermore, $\lambda(B) = 0$ implies that

$$\lambda(C) \ = \ \lambda(\{L = U\} \cap B^c) \ = \ \lambda(\{L = U\}).$$

By taking complements, it follows that

$$\lambda(C^c) \ = \ \lambda(\{L \ne U\}).$$

By definition, f is Riemann integrable if and only if $R_-(f) = R_+(f)$. By (5.9), this is equivalent to $\lambda(L) = \lambda(U)$, which is further equivalent to $\lambda(U - L) = 0$. Because $U - L \ge 0$, Proposition 5.1 implies that $\lambda(\{L \ne U\}) = 0$. Hence f is Riemann integrable if and only if $\lambda(C^c) = 0$. The first claim follows.

Assume now that f is bounded, measurable, and Riemann integrable. Then the Riemann integral of f equals $\int_a^b f(x)dx = R_-(f) = \lambda(L)$. We saw above that then $\lambda(L) = \lambda(U)$, so that $\lambda(\{L \ne U\}) = 0$. Because $L \le f \le U$, it follows that $\{f \ne L\} \subset \{L \ne U\}$, so that $\lambda(\{f \ne L\}) = 0$. Hence, by Proposition 5.3,

$$\int_a^b f(x)dx \ = \ \lambda(L) \ = \ \lambda(f).$$

□

The integral of a Riemann integrable measurable function against the Lebesgue measure corresponds to the classical Riemann integral.

Example 5.13. Compute the integral $\int_{(0,1]} x^{-\alpha} \lambda(dx)$ for $\alpha \neq 1$. By monotone continuity of integration (Theorem 4.5), and noting that $x^{-\alpha} 1_{[1/n,1]} \uparrow x^{-\alpha} 1_{(0,1]}$, we see that

$$\int_{(0,1]} x^{-\alpha} \lambda(dx) \;=\; \lim_{n\to\infty} \int_{[1/n,1]} x^{-\alpha} \lambda(dx).$$

Because $x^{-\alpha}$ is continuous on the closed interval $[1/n, 1]$, we see (Theorem 5.12) that

$$\int_{[1/n,1]} x^{-\alpha} \lambda(dx) \;=\; \int_{1/n}^{1} x^{-\alpha} dx \;=\; \Big|_{1/n}^{1} \frac{x^{1-\alpha}}{1-\alpha} \;=\; \frac{1-(1/n)^{1-\alpha}}{1-\alpha}.$$

Hence

$$\int_{(0,1]} x^{-\alpha} \lambda(dx) \;=\; \lim_{n\to\infty} \frac{1-(1/n)^{1-\alpha}}{1-\alpha} \;=\; \begin{cases} \frac{1}{1-\alpha}, & \alpha < 1, \\ \infty, & \alpha > 1. \end{cases}$$

5.7 Integrating against the law

Let X be a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{S})$. Given a measurable function $g\colon S \to \bar{\mathbb{R}}$, we may interpret $g(X)$ as a random variable $\omega \mapsto g(X(\omega))$. We will prove that, under mild regularity,

$$\mathbb{E} g(X) \;=\; \int_S g(x)\, \mu(dx)$$

where $\mu = \mathbb{P} \circ X^{-1}$ is the law of X.

Theorem 5.14. *For any random variable $X\colon \Omega \to S$ with law P_X and any measurable function $g\colon S \to \bar{\mathbb{R}}$,*

$$\mathbb{E} g(X) \;=\; \int_S g(x)\, P_X(dx). \tag{5.10}$$

Proof. (i) Consider first the case in which $g = 1_A$ is the indicator function of a measurable set $A \in S$. Then $g(X(\omega)) = 1_A(X(\omega)) = 1_{X^{-1}(A)}(\omega)$. Hence the left side of (5.10) equals

$$\mathbb{E} g(X) \;=\; \mathbb{E} 1_{X^{-1}(A)} \;=\; \mathbb{P}(X^{-1}(A)),$$

and the right side of (5.10) equals

$$\int_S 1_A \, dP_X \;=\; P_X(A).$$

We see that (5.10) holds because $P_X(A) = \mathbb{P}(X^{-1}(A))$ by the definition of P_X.

(ii) Assume that g is nonnegative with a finite range. Then $g = \sum_{i=1}^m a_i 1_{A_i}$. By linearity and (i), we find that

$$\mathbb{E}g(X) \;=\; \sum_{i=1}^m a_i \mathbb{E}1_{X^{-1}(A)} \;=\; \sum_{i=1}^m a_i \mathbb{P}(X^{-1}(A)) \;=\; \int_S g(x)\, P_X(dx).$$

(iii) Assume now that $g\colon S \to \bar{\mathbb{R}}_+$. Let g_n be nonnegative finite-range measurable functions such that $g_n \uparrow g$. Then by (ii) and the monotone continuity of integration,

$$\mathbb{E}g(X) \;=\; \lim_{n\to\infty} \mathbb{E}g_n(X) \;=\; \lim_{n\to\infty} \int_S g_n(x)\, P_X(dx) \;=\; \int_S g(x)\, P_X(dx).$$

(iv) Assume now that $g\colon S \to \bar{\mathbb{R}}$ is general. By (iii), we see that

$$\begin{aligned} \mathbb{E}g_+(X) \;&=\; \int_S g_+(x)\, P_X(dx), \\ \mathbb{E}g_-(X) \;&=\; \int_S g_-(x)\, P_X(dx). \end{aligned}$$

We find that $\mathbb{E}|g(X)| < \infty$ iff both $\mathbb{E}g_+(X), \mathbb{E}g_-(X)$ are finite. This is equivalent to both $\int_S g_+(x)\, P_X(dx)$ and $\int_S g_-(x)\, P_X(dx)$ being finite. This is equivalent to $\int_S |g(x)|\, P_X(dx)$ being finite. Todo: Finish this part .

□

Todo: Add examples.

5.8 Exercises

Exercise 5.15 (Density of a nonstandard parametric family)**.** The probability measure $\mu_{p,t}$ in Exercise 3.16 admits a probability density function f with respect to the measure $\nu = \delta_0 + \lambda$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where δ_0 is the Dirac measure at 0, and λ is the Lebesgue measure on the real line.

(a) Determine an expression for f.

(b) Is the probability density function unique?

Exercise 5.16 (Almost surely finite random variable)**.** Let $X\colon \Omega \to \bar{\mathbb{R}}_+$ be a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

(a) Prove that $\mathbb{E}(X) < \infty$ implies that $X < \infty$ almost surely.

(b) Is the converse implication true?

Exercise 5.17 (Chained densities)**.** Let μ_0, μ_1, μ_2 be probability measures on a measurable space $(\Omega, \mathcal{A})$. Assume that f_1 is a probability density function of μ_1 with respect to μ_0, and f_2 is a probability density function of μ_2 with respect to μ_1.

(a) Does μ_2 admit a probability density function with respect to μ_0? If yes, explain how the density function may be determined using f_0 and f_1. If no, explain why.

(b) Do the conclusions of (a) extend to general measures that are not necessarily probability measures?

Exercise 5.18 (Power integrals)**.** Compute the values of the integrals $\int_{[1,\infty)} x^{-\alpha} \, \lambda(dx$ and $\int_{(0,\infty)} x^{-\alpha} \, \lambda(dx)$ for all $\alpha \in (0, \infty)$. Proceed rigorously, as in Example 5.13.

Hint. $[1, \infty) = \cup_{n=1}^{\infty} [1, n]$ and $(0, \infty) = \cup_{n=1}^{\infty} [1/n, n]$.

Exercise 5.19. Prove that the set C in the proof of Theorem 5.12 is a Borel set. (Hint B is a countable set.) Also prove that $\lambda(C^c) = \lambda(\{L \neq U\})$.

Exercise 5.20 (Almost sure events)**.** Let $A_1, A_2 \subset \Omega$ be events that occur almost surely.

(a) Prove that also the event $A_1 \cup A_2$ occurs almost surely.

(b) Does the same conclusion also hold for the union of a countable list of events $A_1, A_2, \ldots$? Justify your answer.

(c) Does the same conclusion also hold for the union of an uncountably infinite set of events A_i, $i \in I$? Justify your answer.

Exercise 5.21. Prove that the set C in the proof of Theorem 5.12 is measurable.

Chapter 6

Multivariate random variables

Multivariate random variables, or random vectors, are random variables with multiple coordinates or components. The components of a random vector can be stochastically dependent or independent. Tensor products of measures form the basic building block for constructing probabilistic models.

Key concepts: product measure, product sigma-algebra, random vector, convolution

Learning outcomes:

- Learn to work with product measures and their densities.
- Learn to apply Fubini theorem to change the order of integration.

Prerequisites: Previous chapters.

6.1 Product of sigma-algebras

In multivariate probability and statistics we often work with products of two or more measurable spaces. Given measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, the Cartesian *product* of sets S_1 and S_2 is denoted by

$$S_1 \times S_2 \;=\; \{(s_1, s_2) \colon s_1 \in S_1,\, s_2 \in S_2\}.$$

We also need to define a sigma-algebra on the product space $S_1 \times S_2$. It is natural to require that the coordinate projection maps $\pi_i \colon S_1 \times S_2 \to S_i$ defined by

$$\begin{aligned} \pi_1((s_1, s_2)) &= s_1, \\ \pi_2((s_1, s_2)) &= s_2, \end{aligned}$$

should be measurable functions. Then all preimages $\pi_1^{-1}(B_1)$ with $B_1 \in \mathcal{S}_1$ and $\pi_2^{-1}(B_2)$ with $B_2 \in \mathcal{S}_2$ should be measurable sets in $S_1 \times S_2$. In other words, the set families $\pi_i^{-1}(\mathcal{S}_i) = \{\pi_i^{-1}(B) \colon B \in \mathcal{S}_i\}$ should be contained in the sigma-algebra on $S_1 \times S_2$. The *product sigma-algebra*

$$\mathcal{S}_1 \otimes \mathcal{S}_2 \;=\; \sigma(\pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2)) \tag{6.1}$$

is by definition the smallest sigma-algebra on $S_1 \times S_2$ such that the projection maps π_1, π_2 are measurable.

$\mathcal{S}_1 \otimes \mathcal{S}_2$ is the sigma-algebra on $S_1 \times S_2$ that contains the preimages of π_1 and π_2, complements and countable intersections and unions thereof, complements and countable intersections and unions thereof, and so on.

Another natural candidate for a sigma-algebra on the product space would be the family of sets $B_1 \times B_2$ with $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. However, this set family is not in general closed under complement nor countable union. The following result confirms that if we extended this set family into a sigma-algebra, we obtain the same sigma-algebra as above.

Proposition 6.1. *The set family $\mathcal{C} = \{B_1 \times B_2 \colon B_1 \in \mathcal{S}_1,\, B_2 \in \mathcal{S}_2\}$ is a generator of $\mathcal{S}_1 \otimes \mathcal{S}_2$.*

Proof. (i) Any set in $\mathcal{C}$ can be written as

$$B_1 \times B_2 \;=\; (B_1 \times S_2) \cap (S_1 \times B_2) \;=\; \pi_1^{-1}(B_1) \cap \pi_2^{-1}(B_2)$$

with $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. Because $\pi_1^{-1}(B_1)$ and $\pi_2^{-1}(B_2)$ belong to $\mathcal{S}_1 \otimes \mathcal{S}_2$, the above equality implies that so does $B_1 \times B_2$. Therefore, $\mathcal{C} \subset \mathcal{S}_1 \otimes \mathcal{S}_2$, and in particular, $\sigma(\mathcal{C}) \subset \mathcal{S}_1 \otimes \mathcal{S}_2$.

(ii) Any set in $\pi_1^{-1}(\mathcal{S}_1)$ can be written as

$$\pi_1^{-1}(B_1) \;=\; B_1 \times S_2$$

for some $B_1 \in \mathcal{S}_1$. Because $S_2 \in \mathcal{S}_2$, we conclude from the above equality that $\pi_1^{-1}(B_1) \in \mathcal{C}$. Therefore, $\pi_1^{-1}(\mathcal{S}_1) \subset \mathcal{C}$. A similar computation implies that also $\pi_2^{-1}(\mathcal{S}_2) \subset \mathcal{C}$, so that $\pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2) \subset \mathcal{C}$. As a consequence,

$$\mathcal{S}_1 \otimes \mathcal{S}_2 \;=\; \sigma(\pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2)) \subset \sigma(\mathcal{C}). \qquad \square$$

Products of measurable sets $B_1 \times B_2$ are sometimes called measurable rectangles, although geometrically they might appear more like barcodes.

6.2 Vector-valued functions

Let $(S_0, \mathcal{S}_0)$, $(S_1, \mathcal{S}_1)$, $(S_2, \mathcal{S}_2)$ be measurable spaces. Unless otherwise mentioned, a product space $S_i \times S_j$ will always be equipped with the product sigma-algebra $\mathcal{S}_i \otimes \mathcal{S}_j$. A vector-valued function $f \colon S_0 \to S_1 \times S_2$ is characterised by the *ccordinate functions* $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$, so that the output values of f can be represented as

$$f(\omega) \;=\; (f_1(\omega), f_2(\omega)).$$

The following result provides a simple way to check whether or not the function f is measurable.

Proposition 6.2. *A function $f \colon S_0 \to S_1 \times S_2$ is measurable if and only if its coordinate functions $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$ are measurable.*

Proof. By definition (6.1), the set family $\mathcal{C} = \pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2)$ is a generator of $\mathcal{S}_1 \otimes \mathcal{S}_2$. Therefore, by Proposition 3.3, f is $\mathcal{S}_0/(\mathcal{S}_1 \otimes \mathcal{S}_2)$-measurable if and only if $f^{-1}(C) \in \mathcal{S}_0$ for all $C \in \mathcal{C}$. Because all sets in $\mathcal{C}$ are either of the form $C = \pi_1^{-1}(B)$ for some $B \in \mathcal{S}_1$ or $C = \pi_2^{-1}(B)$ for some $B \in \mathcal{S}_2$, we see that

$$\begin{aligned}
&f \text{ is } \mathcal{S}_0/(\mathcal{S}_1 \otimes \mathcal{S}_2)\text{-measurable} \\
\iff\; & f^{-1}(C) \in \mathcal{S}_0 \text{ for all } C \in \mathcal{C} \\
\iff\; & f^{-1}(\pi_i^{-1}(B)) \in \mathcal{S}_0 \text{ for all } B \in \mathcal{S}_i \text{ and for all } i \in \{1,2\} \\
\iff\; & (\pi_i \circ f)^{-1}(B) \in \mathcal{S}_0 \text{ for all } B \in \mathcal{S}_i \text{ and for all } i \in \{1,2\} \\
\iff\; & \pi_i \circ f \text{ is } \mathcal{S}_0/\mathcal{S}_i\text{-measurable for all } i \in \{1,2\}.
\end{aligned}$$

□

From now on, unless otherwise mentioned, $S_1 \times S_2$ is always considered a measurable space equipped with the product sigma-algebra $\mathcal{S}_1 \otimes \mathcal{S}_2$. The proof of the following simple statement is harder that what one might expect, and can be skipped without fear of losing probabilistic intuition.

A measurable function on $S_1 \times S_2$ remains measurable if we freeze one of the input variables.

Proposition 6.3. *The following functions are measurable:*

(i) $s_1 \mapsto (s_1, s_2)$, *for any* s_2.

(ii) $s_2 \mapsto (s_1, s_2)$, *for any* s_1.

Proof. (i) Fix $s_2 \in S_2$ and consider a function $g\colon S_1 \to S_1 \times S_2$ defined by $g(s_1) = (s_1, s_2)$. Its coordinate functions are $g_1(s_1) = s_1$ and $g_2(s_1) = s_2$. The first coordinate function $g_1\colon S_1 \to S_1$ is obviously measurable, being the identity function. The second coordinate function $g_2\colon S_1 \to S_2$ is measurable, being a constant function. Proposition 6.2 now implies that g is measurable.

(ii) Analogous to the proof of (i). □

Proposition 6.4. *For any measurable function* $f\colon S_1 \times S_2 \to S_3$, *the following functions are measurable:*

(i) $s_1 \mapsto f(s_1, s_2)$, *for any* s_2.

(ii) $s_2 \mapsto f(s_1, s_2)$, *for any* s_1.

Proof. (i) Fix $s_2 \in S_2$ and note that the function $g\colon S_1 \to S_1 \times S_2$ defined by $g(s_1) = (s_1, s_2)$ is measurable by Proposition 6.3. Because compositions of measurable functions are measurable (Proposition 3.5), we find that the function $f \circ g\colon S_1 \to S_3$ is measurable. This function, given by $(f \circ g)(s_1) = f(s_1, s_2)$, is the one that we wanted to prove measurable.

(ii) Analogous to the proof of (i). □

Proposition 6.5. *For any measurable function* $f\colon S_1 \times S_2 \to \bar{\mathbb{R}}_+$ *and any* *finite measures* μ_1 *and* μ_2*,*

$$s_1 \mapsto \int_{S_2} f(s_1,s_2)\,\mu_2(ds_2) \text{ is measurable,} \tag{6.2}$$

$$s_2 \mapsto \int_{S_1} f(s_1,s_2)\,\mu_1(ds_1) \text{ is measurable.} \tag{6.3}$$

This is a difficult proof that could go to the Appendix (can be skipped).

Proof. By Proposition 6.4, the function $s_2 \mapsto f(s_1,s_2)$ is measurable for every s_1. Therefore, the function

$$g(s_1) \;=\; \int_{S_2} f(s_1,s_2)\,\mu_2(ds_2)$$

is well defined. We will next verify that g is measurable.

(i) Assume that $f = 1_{B_1\times B_2}$ for some $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. Because $1_{B_1\times B_2}(s_1,s_2) = 1_{B_1}(s_1)1_{B_2}(s_2)$, we see that

$$g(s_1) \;=\; \int_{S_2} 1_{B_1}(s_1)\,1_{B_2}(s_2)\,\mu_2(ds_2) \;=\; 1_{B_1}(s_1)\,\mu_2(B_2).$$

This indicates that g is a constant multiple of the indicator function 1_{B_1}. Therefore, g is measurable.

(ii) Let us verify that g is measurable whenever $f = 1_B$ for some $B \in \mathcal{S}_1 \otimes \mathcal{S}_2$. To this end, freeze a value $s_2 \in S_2$, and define

$$\mathcal{G} \;=\; \left\{ B \in \mathcal{S}_1 \otimes \mathcal{S}_2 \colon s_1 \mapsto \int_{S_2} 1_B(s_1,s_2)\,\mu_2(ds_2) \text{ is measurable} \right\}.$$

Part (i) tells us that the set family $\mathcal{C} = \{B_1 \times B_2 \colon B_1 \in \mathcal{S}_1, B_2 \in \mathcal{S}_2\}$ is contained in $\mathcal{G}$. We will next verify the following:

- $\mathcal{G}$ is closed under subset difference. If $B, C \in \mathcal{G}$ are such that $B \subset C$, then by writing $1_{C\setminus B} = 1_C - 1_B$, and noting that the functions $s_1 \mapsto 1_B(s_1,s_2)$ and $s_2 \mapsto 1_C(s_1,s_2)$ are in $L^1(\mu_2)$, we see that

$$\int_{S_2} 1_{C\setminus B}(s_1,s_2)\,\mu_2(ds_2) \;=\; \int_{S_2} 1_C(s_1,s_2)\,\mu_2(ds_2) - \int_{S_2} 1_B(s_1,s_2)\,\mu_2(ds_2).$$

  Because terms on the right, considered as functions of s_1, are measurable, and because linear combinations of measurable functions are measurable, it follows that also the term on the left is measurable as a function of s_1. Therefore, $C \setminus B \in \mathcal{G}$.

- $\mathcal{G}$ is closed under increasing set limit. If $B_n \in \mathcal{G}$ are such that $B_n \uparrow B$. Then $1_{B_n} \uparrow 1_B$ pointwise on $S_1 \times S_2$, so that the functions $f_n(s_2) = 1_{B_n}(s_1, s_2)$ and $f(s_2) = 1_B(s_1, s_2)$ and satisfy $f_n \uparrow f$ pointwise. By monotone continuity, it follows that $\mu_2(f_n) \uparrow \mu_2(f)$. That is,

$$\int_{S_2} 1_B(s_1, s_2)\, \mu_2(ds_2) \;=\; \lim_{n\to\infty} \int_{S_2} 1_{B_n}(s_1, s_2)\, \mu_2(ds_2).$$

  The above equality means that the left side, considered as a function of s_1, is a pointwise limit of measurable functions, and therefore measurable. Therefore, $B \in \mathcal{G}$.

The above facts indicate that $\mathcal{G}$ is a Dynkin class. The monotone class theorem (Theorem 2.8) now implies that $\mathcal{G} \supset \mathcal{S}_1 \otimes \mathcal{S}_2$. Therefore, (6.2) holds for all indicator functions 1_B with $B \in \mathcal{S}_1 \otimes \mathcal{S}_2$.

(iii) Assume now that f is measurable and has a finite range. Then $f = \sum_{i=1}^n b_i 1_{B_i}$ for some constants $b_i \in \bar{\mathbb{R}}_+$ and sets $B_i \in \mathcal{S}_1 \otimes \mathcal{S}_2$. Then with the help of (ii), we see that $g(s_1) = \sum_{i=1}^n b_i \int_{S_2} 1_{B_i}(s_1, s_2)\, \mu(ds_2)$ is a linear combination of measurable functions. Because linear combinations of measurable functions are measurable, we conclude that (6.2) holds.

(iv) For a general measurable function $f\colon S_1 \times S_2 \to \bar{\mathbb{R}}_+$ there exist $f_n \uparrow f$ which are finite-range and measurable. Then with the help of (iii) and monotone continuity, we see that $g(s_1) = \lim_{n\to\infty} \int_{S_2} f_n(s_1, s_2)\, \mu(ds_2)$ is a pointwise limit of measurable functions. Because limits of measurable functions are measurable, we conclude that (6.2) holds.

We have proved (6.2). The other statement (6.3) follows analogously. □

6.3 Product of measures

Çınlar [Çın11] does kernels first, then Fubini, saves effort.

A measure is *sigma-finite* if there exists a sequence of measurable sets $A_1, A_2, \dots$ such that $\cup_n A_n = S$ and $\mu(A_n) < \infty$ for all n. All finite measures, and in particular all probability measures, are sigma-finite. Examples of infinite but sigma-finite measures include the Lebesgue measure, and the counting measure on a countably infinite space (homework).

The *product measure* of sigma-finite measures μ_1 and μ_2 is a set function on $\mathcal{S}_1 \otimes \mathcal{S}_2$ defined by

$$(\mu_1 \otimes \mu_2)(B) \;=\; \int_{S_1} \left(\int_{S_2} 1_B(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1). \tag{6.4}$$

Proposition 6.6. *The product measure $\mu_1 \otimes \mu_2$ is the unique measure on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ such that*

$$(\mu_1 \otimes \mu_2)(B_1 \times B_2) \;=\; \mu_1(B_1)\mu_2(B_2) \tag{6.5}$$

for all $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$.

Proof. Let us first verify that $\mu_1 \otimes \mu_2$ is indeed a measure. For any $B \in \mathcal{S}_1 \otimes \mathcal{S}_2$, the right side in (6.4) is a well-defined element in $\bar{\mathbb{R}}_+$ because:

(i) $s_2 \mapsto 1_B(s_1, s_2)$ is a measurable function from S_2 into $\bar{\mathbb{R}}_+$ for any fixed $s_1 \in S_1$ (see Proposition 6.3), so that the inner integral in (6.4) assigns each $s_1 \in S_1$ a well-defined value in $\bar{\mathbb{R}}_+$.

(ii) $s_1 \mapsto \int_{S_2} 1_B(s_1, s_2)\, \mu_2(ds_2)$ is a measurable function from S_1 into $\bar{\mathbb{R}}_+$ (see Proposition 6.5), so that the outer integral in (6.4) yields a well-defined value in $\bar{\mathbb{R}}_+$.

Observe next that $\mu_1 \otimes \mu_2(\emptyset) = 0$ because $1_\emptyset$ equals the zero function. To verify countable disjoint additivity, let $B_1, B_2, \dots$ be disjoint measurable sets in $\mathcal{S}_1 \otimes \mathcal{S}_2$. Observe that

$$1_{\bigcup_{i=1}^\infty B_i}(s_1, s_2) \;=\; \sum_{i=1}^\infty 1_{B_i}(s_1, s_2) \;=\; \lim_{n\to\infty} \sum_{i=1}^n 1_{B_i}(s_1, s_2).$$

Therefore, by monotone continuity (Theorem 4.5) and linearity (Proposition 4.6), we see that

$$\begin{aligned}
(\mu_1 \otimes \mu_2)\Big(\bigcup_{i=1}^\infty B_i\Big) \;&=\; \int_{S_1} \left(\int_{S_2} \lim_{n\to\infty} \sum_{i=1}^n 1_{B_i}(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1) \\
&\overset{\text{cont}}{=} \lim_{n\to\infty} \int_{S_1} \left(\int_{S_2} \sum_{i=1}^n 1_{B_i}(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1) \\
&\overset{\text{lin}}{=} \lim_{n\to\infty} \sum_{i=1}^n \int_{S_1} \left(\int_{S_2} 1_{B_i}(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1).
\end{aligned}$$

Because the last expression above equals $\sum_{i=1}^\infty (\mu_1 \otimes \mu_2)(B_i)$, we conclude that $\mu_1 \otimes \mu_2$ is a measure.

(iii) Let us verify uniqueness. Assume that μ and ν are measures on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ both satisfying (6.5). Because μ_1 and μ_2 are sigma-finite, we may fix measurable sets $S_{1,n} \uparrow S_1$ and $S_{2,n} \uparrow S_2$ such that $\mu_1(S_{1,n}) < \infty$ and $\mu_2(S_{2,n}) < \infty$ for all n. Define truncated measures $\mu^{(n)}(B) = \mu(B \cap (S_{1,n} \times$

$S_{2,n}))$ and $\nu^{(n)}(B) = \nu(B \cap (S_{1,n} \times S_{2,n}))$. Then we find that $\mu^{(n)}$ and $\nu^{(n)}$ agree on the set family $\mathcal{C} = \{B_1 \times B_2 \colon B_1 \in \mathcal{S}_1, B_2 \in \mathcal{S}_2\}$, and both measures have total mass equal to $\mu_1(S_{1,n})\mu_2(S_{2,n}) < \infty$. If the the total mass is nonzero, then we may normalise these measures to probability measures, and Dynkin's identification theorem (Theorem 2.7) then implies that $\mu^{(n)} = \nu^{(n)}$. If the the total mass is zero, then trivially $\mu^{(n)} = \nu^{(n)}$. For any measurable set B, the fact $B \cap (S_{1,n} \times S_{2,n}) \uparrow B$ combined with the monotone continuity of measures (Proposition 1.5) then implies that

$$\mu(B) \;=\; \lim_{n\to\infty} \mu^{(n)}(B) \;=\; \lim_{n\to\infty} \nu^{(n)}(B) \;=\; \nu(B).$$

Hence $\mu = \nu$, and indeed both must be equal to $\mu_1 \otimes \mu_2$. □

6.4 Integrating against the product measure

The following important result tells us two things: (i) how to integrate against a product measure, and (ii) when can the order of integration be changed in iterated integrals. The version of the result for integrable functions is called *Fubini's theorem*[1]. and the version of the result of nonnegative functions is called *Tonelli's theorem*[2].

Theorem 6.7 (Fubini–Tonelli theorem). *For any sigma-finite measures μ_1, μ_2 on measurable spaces $(S_1, \mathcal{S}_1)$, $(S_2, \mathcal{S}_2)$, and for any measurable function $f \colon S_1 \times S_2 \to \bar{\mathbb{R}}$ such that either $f \geq 0$ or $f \in L^1(\mu_1 \otimes \mu_2)$,*

$$\int_{S_1 \times S_2} f \, d(\mu_1 \otimes \mu_2) \;=\; \int_{S_1} \left(\int_{S_2} f(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1). \tag{6.6}$$

and

$$\int_{S_1} \left(\int_{S_2} f(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1) \;=\; \int_{S_2} \left(\int_{S_1} f(s_1, s_2)\, \mu_1(ds_1) \right) \mu_2(ds_2). \tag{6.7}$$

Proof. (i) Assume first that $f = 1_B$ for some $B \in \mathcal{S}_1 \otimes \mathcal{S}_2$. Then (6.6) is valid by the defining formula (6.4).

(ii) Assume that $f \geq 0$ is measurable with a finite range. Then we may write $f = \sum_{i=1}^n b_i 1_{B_i}$ where $\{b_1, \dots, b_n\}$ is an enumeration of the range of f,

[1] Guido Fubini (1879 – 1943) PhD 1900 from Pisa.

[2] Leonida Tonelli (1885–1946). PhD 1907 from Bologna.

and $B_i = f^{-1}(\{b_i\})$. Then (6.6) follows by applying (i) and the linearity of nonnegative integration (Proposition 4.6).

(iii) Assume that $f \ge 0$ is measurable. Let $f_n \ge 0$ be finite-range measurable functions such that $f_n \uparrow f$ (these exist by Proposition 4.1). Then by applying (ii) and the monotone continuity of nonnegative integration (Theorem 4.5), we conclude that (6.6) holds for an arbitrary measurable $f \ge 0$.

(iv) Assume that $f\colon S \to \mathbb{R}$ satisfies $f \in L^1(\mu_1 \otimes \mu_2)$. Denote the inner integral in (6.6) by

$$J^f(s_1) \;=\; \int_{S_2} f(s_1, s_2)\, \mu_2(ds_2).$$

By definition, $J^f(s_1) = J^{f_+}(s_1) - J^{f_-}(s_1)$, where

$$J^{f_\pm}(s_1) \;=\; \int_{S_2} f_\pm(s_1, s_2)\, \mu_2(ds_2).$$

It is possible that $J^{f_\pm}(s_1)$ are both infinite for some values of s_1, in which case $J^f(s_1)$ is ill defined, and we assign $J^f(s_1) = -\infty$ by our convention. Nevertheless, the set of such input values s_1 is negligible, as we will next verify. Note that the functions $J^{f_\pm}\colon S_1 \to \bar{\mathbb{R}}_+$ are measurable due to Proposition 6.5, and we see by (iii) that

$$\int_{S_1} J^{f_\pm}(s_1)\, \mu_1(ds_1) \;=\; \int_{S_1 \times S_2} f_\pm \, d(\mu_1 \otimes \mu_2) \;<\; \infty.$$

As a consequence, the set $A = A_+ \cap A_-$ with $A_\pm = \{s_1\colon J^{f_\pm}(s_1) < \infty\}$ satisfies $\mu_1(A^c) = 0$. It follows that the functions $1_A J^{f_\pm}$ are real valued and belong to $L^1(\mu_1)$. Because $L^1(\mu_1)$ is a real vector space (Proposition 4.11), it follows that also the function $1_A J^f = 1_A J^{f_+} - 1_A J^{f_-}$ is real valued and in $L^1(\mu_1)$, and furthermore,

$$\int_{S_1} 1_A J^f \, d\mu_1 \;=\; \int_{S_1} 1_A J^{f_+} \, d\mu_1 - \int_{S_1} 1_A J^{f_-} \, d\mu_1.$$

Because $\mu(A^c) = 0$, we may omit the indicator 1_A from each integral in the above equality (Proposition 5.3), and it follows that

$$\begin{aligned}
\int_{S_1} J^f \, d\mu_1 \;&=\; \int_{S_1} J^{f_+} \, d\mu_1 - \int_{S_1} J^{f_-} \, d\mu_1 \\
&=\; \int_{S_1} f_+ \, d(\mu_1 \otimes \mu_2) - \int_{S_1} f_- \, d(\mu_1 \otimes \mu_2) \\
&=\; \int_{S_1} f \, d(\mu_1 \otimes \mu_2).
\end{aligned}$$

Hence (6.6) is valid for all $f \in L^1(\mu_1 \otimes \mu_2)$.

(v) Let us now prove (6.7). Define two set functions on $\mathcal{S}_1 \otimes \mathcal{S}_2$ by

$$\mu(B) = \int_{S_1} \left(\int_{S_2} 1_B(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1), \tag{6.8}$$

$$\nu(B) = \int_{S_2} \left(\int_{S_1} 1_B(s_1, s_2)\, \mu_1(ds_1) \right) \mu_2(ds_2). \tag{6.9}$$

Then $\mu = \mu_1 \otimes \mu_2$ equals the product measure in (6.4). Proposition 6.6 tells us that μ is a measure on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$, and (6.6) implies that

$$\int_{S_1} \left(\int_{S_2} f(s_1, s_2)\, \mu_2(ds_2) \right) \mu_1(ds_1) = \int_{S_1 \times S_2} f\, d\mu. \tag{6.10}$$

By applying the same results with the roles of μ_1 and μ_2 interchanged, we see that also ν is a measure on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$, and that

$$\int_{S_2} \left(\int_{S_1} f(s_1, s_2)\, \mu_1(ds_1) \right) \mu_2(ds_2) = \int_{S_1 \times S_2} f\, d\nu. \tag{6.11}$$

By substituting $B = B_1 \times B_2$ into (6.8)–(6.9), we see that

$$\mu(B_1 \times B_2) = \mu_1(B_1)\mu_2(B_2) = \nu(B_1 \times B_2)$$

for all $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. Proposition 6.6 states that $\mu_1 \otimes \mu_2$ is the unique measure with this property. Therefore, μ and ν are both equal to $\mu_1 \otimes \mu_2$, and it follows that the integrals in (6.10)–(6.11) are all equal to each other.

□

The first integral above is a double integral, *and the latter two are* iterated integrals.

For a nonnegative measurable function $f \colon S \to \bar{\mathbb{R}}_+$, we may always swap the order of the iterated integrals $\int_{S_1} \int_{S_2}$ and $\int_{S_2} \int_{S_1}$.

For a signed measurable function $f \colon S \to \bar{\mathbb{R}}$, we apply Fubini–Tonelli theorem twice: First to the nonnegative function $|f|$ to verify that f is integrable, and then to the function f itself to swap the order of the iterated integrals

6.5 Independent random variables

Random variables $X_1 \colon \Omega \to S_1$ and $X_2 \colon \Omega \to S_2$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ are called stochastically *independent* if

$$\mathbb{P}(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}) \;=\; \mathbb{P}(\{X_1 \in B_1\})\, \mathbb{P}(\{X_2 \in B_2\}) \tag{6.12}$$

for all $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. This is denoted by $X_1 \perp\!\!\!\perp X_2$.

New result: Added Thu 26 Sep 2024

Proposition 6.8. *$X_1 \perp\!\!\!\perp X_2$ if and only if $f_1(X_1) \perp\!\!\!\perp f_2(X_2)$ for all measurable $f_1 \colon S_1 \to \bar{\mathbb{R}}_+$ and $f_2 \colon S_2 \to \bar{\mathbb{R}}_+$.*

Independence is preserved under deterministic transformations.

Proof. (i) Assume that $f_1(X_1) \perp\!\!\!\perp f_2(X_2)$ for all measurable f_1, f_2. Fix $A_i \in \mathcal{S}_i$ and define $f_i = 1_{A_i}$. Then

$$\begin{aligned}\mathbb{P}(X_1 \in A_1,\, X_2 \in A_2) &= \mathbb{P}(f_1(X_1) = 1,\, f_2(X_2) = 1)\\ &= \mathbb{P}(f_1(X_1) = 1)\, \mathbb{P}(f_2(X_2) = 1)\\ &= \mathbb{P}(X_1 \in A_1)\, \mathbb{P}(X_2 \in A_2)\end{aligned}$$

confirms that $X_1 \perp\!\!\!\perp X_2$.

(ii) Assume that $X_1 \perp\!\!\!\perp X_2$. Fix measurable functions $f_i \colon S_i \to \bar{\mathbb{R}}_+$. Then for any measurable sets $B_i \subset \bar{\mathbb{R}}_+$,

$$\begin{aligned}\mathbb{P}(f_1(X_1) \in B_1,\, f_2(X_2) \in B_2) &= \mathbb{P}(X_1 \in f_1^{-1}(B_1),\, X_2 \in f_2^{-1}(B_2))\\ &= \mathbb{P}(X_1 \in f_1^{-1}(B_1))\, \mathbb{P}(X_2 \in f_2^{-1}(B_2))\\ &= \mathbb{P}(f_1(X_1) \in B_1)\, \mathbb{P}(f_2(X_2) \in B_2).\end{aligned}$$

Hence $f_1(X_1) \perp\!\!\!\perp f_2(X_2)$. □

Proposition 6.9. *X_1 and X_2 are independent if and only if the law $P_{(X_1,X_2)}$ of the random vector $X = (X_1, X_2)$ factorises according to*

$$P_X \;=\; P_{X_1} \otimes P_{X_2},$$

where P_{X_1} and P_{X_2} are the laws of X_1 and X_2.

Random variables are independent if and only if their joint law factorises into a product measure of the laws of the individual random variables.

Proof. Observe first that the set $X_1^{-1}(B_1) \cap X_2^{-1}(B_2)$ can be written as

$$\begin{aligned}
\{X_1 \in B_1\} \cap \{X_2 \in B_2\} &= \{\omega \colon X_1(\omega) \in B_1 \text{ and } X_2(\omega) \in B_2\} \\
&= \{\omega \colon (X_1(\omega), X_2(\omega)) \in B_1 \times B_2\} \\
&= \{\omega \colon X(\omega) \in B_1 \times B_2\} \\
&= X^{-1}(B_1 \times B_2).
\end{aligned}$$

(i) Assume that $P_X = P_{X_1} \otimes P_{X_2}$. Then for any $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$,

$$\begin{aligned}
\mathbb{P}(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}) &= \mathbb{P}(X^{-1}(B_1 \times B_2)) \\
&= P_X(B_1 \times B_2) \\
&= (P_{X_1} \otimes P_{X_2})(B_1 \times B_2) \\
&\overset{(6.5)}{=} P_{X_1}(B_1) P_{X_2}(B_2) \\
&= \mathbb{P}(\{X_1 \in B_1\})\, \mathbb{P}(\{X_2 \in B_2\}).
\end{aligned}$$

(ii) Assume that X_1 and X_2 are independent. Therefore, the left side of (6.12) equals

$$\mathbb{P}(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}) \;=\; P_X(B_1 \times B_2).$$

On the other hand, the left side of (6.12) equals

$$\mathbb{P}(X_1^{-1}(B_1))\, \mathbb{P}((X_2^{-1}(B_2)) \;=\; P_{X_1}(B_1)\, P_{X_2}(B_2).$$

By plugging these formulas into (6.12), we see that the probability measure P_X satisfies

$$P_X(B_1 \times B_2) \;=\; P_{X_1}(B_1)\, P_{X_2}(B_2)$$

for all $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$. Proposition 6.6 tells us that the product measure $P_{X_1} \otimes P_{X_2}$ is the unique measure on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ with the above property. Therefore, we conclude that $P_X = P_{X_1} \otimes P_{X_2}$. □

Proposition 6.10. *Let X_1 and X_2 be independent random variables such that the law of X_i admits a density function f_i with respect to a reference measure ν_i. Then the law of (X_1, X_2) admits a density function $f(s_1, s_2) = f(s_1) f(s_2)$ with respect to $\nu_1 \otimes \nu_2$.*

Proof. The independence of X_1 and X_2 implies (Proposition 6.9) that their joint law factorises according to $P_{X_1,X_2} = P_{X_1} \otimes P_{X_2}$. By Fubini–Tonelli (Theorem 6.7) and Proposition 5.8,

$$\begin{aligned}
P_{X_1,X_2}(B) &= \int_{S_1\times S_2} 1_B(s_1,s_2)\, P_{X_1,X_2}(ds_1,ds_2) \\
&= \int_{S_1\times S_2} 1_B(s_1,s_2)\, (P_{X_1}\otimes P_{X_2})(ds_1,ds_2) \\
&\overset{\text{Fub}}{=} \int_{S_1} \left(\int_{S_2} 1_B(s_1,s_2) P_{X_2}(ds_2) \right) P_{X_1}(ds_1) \\
&= \int_{S_1} \left(\int_{S_2} 1_B(s_1,s_2) f_2(s_2)\nu(ds_2) \right) f_1(s_1)\nu_1(ds_1) \\
&\overset{\text{Fub}}{=} \int_{S_1\times S_2} 1_B(s_1,s_2) f_1(s_1) f_2(s_2)\, (\nu_1\otimes\nu_2)(ds_1,ds_2).
\end{aligned}$$

Therefore,

$$P_{X_1,X_2}(B) = \int_B f(s_1,s_2)\,(\nu_1\otimes\nu_2)(ds_1,ds_2).$$

□

Example 6.11. Let X_1, X_2 independent real-valued random variables with laws admitting densities f_1, f_2 with respect to the Lebesgue measure on $\mathbb{R}$. Then $f(x_1,x_2) = f_1(x_1)f_2(x_2)$ is a density function of the law of the random vector (X_1,X_2) with respect to the 2-dimensional Lebesgue measure $\lambda_2 = \lambda\otimes\lambda$ on $\mathbb{R}^2$. In particular, for any Borel set $B \subset \mathbb{R}^2$,

$$\mathbb{P}((X_1,X_2)\in B) = \int_{\mathbb{R}}\int_{\mathbb{R}} 1_B(x_1,x_2)\, f_1(x_1)f_2(x_2)\, dx_1 dx_2,$$

where simply write dx_1 instead of $\lambda(dx_1)$.

6.6 Sum of independent random variables

The *convolution* of probability measures μ and ν on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is a set function defined by

$$(\mu * \nu)(B) = \int_{\mathbb{R}}\int_{\mathbb{R}} 1_B(x+y)\,\mu(dx)\,\nu(dy).$$

Proposition 6.12. *The convolution of probability measures on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *is a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof. One may prove the claim directly using the linearity and monotone continuity of nonnegative integration. Here is an alternative, perhaps more elegant, proof. Denote $h(x, y) = x + y$. We note that for any measurable $B \subset \mathbb{R}$,

$$\begin{aligned}(\mu * \nu)(B) &= \int_{\mathbb{R}^2} 1_B(x+y)\,(\mu \otimes \nu)(dx, dy) \\ &= \int_{\mathbb{R}^2} (1_B \circ h)(x, y)\,(\mu \otimes \nu)(dx, dy) \\ &= \int_{\mathbb{R}} 1_B(z)\,((\mu \otimes \nu) \circ h^{-1})(dz) \\ &= ((\mu \otimes \nu) \circ h^{-1})(B).\end{aligned}$$

Therefore, $\mu * \nu = (\mu \otimes \nu) \circ h^{-1}$ is just the pushforward of the probability measure $\mu \otimes \nu$ on $\mathbb{R}^2$ by the measurable function $h \colon \mathbb{R}^2 \to \mathbb{R}$. Because pushforwards of probability measures are probability measures, the claim follows. □

The following result gives more probabilistic insight.

Proposition 6.13. *The law of the sum* $X + Y$ *of independent real-valued random variables is given by* $P_{X+Y} = P_X * P_Y$.

Proof. Let X and Y be independent. Then their joint law factorises as $P_{X,Y} = P_X \otimes P_Y$. It follows that

$$\begin{aligned}\mathbb{P}(X + Y \in B) &= \int_{S_1 \times S_2} 1_B(x+y)\,P_{X,Y}(dx, dy) \\ &= \int_{S_1 \times S_2} 1_B(x+y)\,(P_X \otimes P_Y)(dx, dy) \\ &\overset{\text{Fub}}{=} \int_{S_1} \int_{S_2} 1_B(x+y)\,P_X(dx)P_Y(dy)\end{aligned}$$

We conclude that $\mathrm{Law}(X + Y)$ equals the convolution $P_X * P_Y$. □

When X and Y are independent, the law of $X+Y$ is fully determined by the laws P_X and P_Y by convolution. In general, this is not the case, and the law of the sum is also affected by the stochastic dependence structure between X and Y.

Example 6.14. Let X, Y independent real-valued random variables with laws P_X and P_Y admitting densities f, g with respect to the Lebesgue measure on $\mathbb{R}$. Then the law of $X+Y$ equals the convolution $P_{X+Y} = P_X * P_Y$. By Fubini–Tonelli theorem, the convolution may be computed as

$$\begin{aligned}(P_X * P_Y)(B) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_B(x+y) f(x) \lambda(dx) \right) g(y) \lambda(dx) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_B(x+y) g(y) \lambda(dy) \right) f(x) \lambda(dx).\end{aligned}$$

For any x, the inner integral may be written as

$$\int_{\mathbb{R}} \phi_x(y+x) \lambda(dy)$$

where $\phi_x(y) = 1_B(y) g(y-x)$. Because λ is shift-invariant, we see (Exercise 6.19) that

$$\int_{\mathbb{R}} \phi_x(y+x) \lambda(dy) = \int_{\mathbb{R}} \phi_x(y) \lambda(dy) \qquad \text{for all } x.$$

It follows that (again, the integration order does not matter due to Fubini–Tonelli)

$$\begin{aligned}(P_X * P_Y)(B) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_B(x+y) g(y) \lambda(dy) \right) f(x) \lambda(dx) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_B(y) g(y-x) \lambda(dy) \right) f(x) \lambda(dx) \\ &= \int_B (f * g)(y) \lambda(dy),\end{aligned}$$

where the function $f * g \colon \mathbb{R} \to \mathbb{R}_+$ is defined by $(f * g)(y) = \int_{\mathbb{R}} f(x) g(y-x) \lambda(dx)$. We see that the law of $X+Y$ admits a probability density function $f * g$ with respect to the Lebesgue measure. The density is called the *convolution* of functions f and g.

6.7 Higher-order product sigma-algebras

The *product* of sets $S_1, S_2, \dots$ is defined by

$$S_1 \times S_2 \times \cdots \;=\; \{(s_1, s_2, \dots) \colon s_1 \in S_1, s_2 \in S_2, \dots\}.$$

The *product sigma-algebra* of sigma-algebras $\mathcal{S}_1, \mathcal{S}_2, \dots$ is defined by

$$\mathcal{S}_1 \otimes \mathcal{S}_2 \times \cdots \;=\; \sigma\Big(\pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2) \cup \cdots\Big).$$

The *product measure* $\mu_1 \otimes \cdots \otimes \mu_n$ of sigma-finite measures $\mu_1, \dots, \mu_n$ is defined by iteration the construction of $n = 2$.

Add details, or move to later place.

Example 6.15. The *multivariate standard normal distribution* is the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ defined by $\mu_n = f_n \cdot \lambda_n$ where λ_n is the Lebesgue measure on $\mathbb{R}^n$ and

$$f_n(x) \;=\; (2\pi)^{-n/2} e^{-\|x\|^2/2}, \qquad x \in \mathbb{R}^n.$$

By writing $\|x\|^2 = \sum_{i=1}^n x_i^2$, we see that the probability density function of μ_n factorises according to

$$f_n(x) \;=\; (2\pi)^{-n/2} e^{-\sum_{i=1}^n x_i^2/2} \;=\; \prod_{i=1}^n (2\pi)^{-1/2} e^{-x_i^2/2},$$

so that $f_n(x) = f_1(x_1) \cdots f_1(x_n)$. Hence we find by ref that the n-variate standard normal distribution is the n-fold tensor product of the univariate standard normal distribution $\mu_n = \underbrace{\mu_1 \otimes \cdots \otimes \mu_1}_{n}$.

6.8 Exercises

Exercise 6.16 (Min and max of random integers)**.** Let $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = \max\{X_1, X_2\}$, where X_1 and X_2 are independent integer-valued random variables, both distributed according to a probability mass function $f \colon \mathbb{Z} \to [0, 1]$. Determine:

(a) The probability mass function of Y_1.

(b) The probability mass function of the vector (Y_1, Y_2).

Exercise 6.17 (Min and max of random numbers). Let $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = \max\{X_1, X_2\}$, where X_1 and X_2 are independent real-valued random variables distributed according to a probability density function f with respect to the Lebesgue measure on $\mathbb{R}$, so that $\mathbb{P}(X_i \in B) = \int_B f(x)\,\lambda(dx)$. Determine:

(a) The cumulative distribution function of Y_1.

(b) The probability density function of Y_1 with respect to λ.

(c) A probability density function $g \colon \mathbb{R}^2 \to \mathbb{R}_+$ of the random vector (Y_1, Y_2) such that

$$\mathbb{P}((Y_1, Y_2) \in B) \;=\; \int_B g(y_1, y_2)\, dy_1 dy_2.$$

Hint: Consider the integral $\int_{\mathbb{R}^2} 1_B(x_1 \wedge x_2, x_1 \vee x_2)\, f(x_1) f(x_2)\, dx_1) dx_2$. You may split the integral according into two regions based on whether or not $x_1 < x_2$, and apply symmetry.

Exercise 6.18 (Three coins). Let $p \in [0,1]$ and define a probability measure on $(\{0,1\}, 2^{\{0,1\}})$ by $\mu = (1-p)\delta_0 + p\delta_1$. Let $\Omega = \{0,1\}^3$, and define a probability measure on $(\Omega, 2^\Omega)$ by $\mu^{\otimes 3} = \mu \otimes \mu \otimes \mu$.

(a) Compute the probability $\mu^{\otimes 3}(\{0\})$.

(b) Determine a probability density function for the random variable $X(\omega) = \sum_{i=1}^{3} \omega_i$ with respect to the counting measure on $(\mathbb{Z}, 2^{\mathbb{Z}})$.

(c) Compute the expected value $\mathbb{E}X$.

(d) Can you generalise your results from $n = 3$ to a general integer n?

Exercise 6.19 (Lebesgue integral of a shifted function). Prove that for every measurable function $f \colon \mathbb{R} \to \bar{\mathbb{R}}$

$$\int_{\mathbb{R}} f(x+t)\,\lambda(dx) \;=\; \int_{\mathbb{R}} f(x)\,\lambda(dx) \qquad \text{for all } t \in \mathbb{R},$$

where λ is the Lebesgue measure on $\mathbb{R}$. You may proceed using the 'standard machine' as follows:

(a) Prove the claim in the special case where $f = 1_B$ for some measurable set $B \subset \mathbb{R}$.

(b) Generalise your result from (a) to the case where f has a finite range.

(c) Generalise your result from (b) to the case where f is a general nonnegative measurable function.

Hint: You may imitate selected parts from the proof of Theorem 6.7. Recall that the Lebesgue measure is shift invariant.

Exercise 6.20. Prove that $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$.

Hint: Every open set in $\mathbb{R}^2$ can be written as a union of sets of form $(a_1, b_1) \times (a_2, b_2)$ in which a_1, b_1, a_2, b_2 are rational numbers.

Exercise 6.21. Let $\#_{\mathbb{Z}}$ be the counting measure on $(\mathbb{Z}, 2^{\mathbb{Z}})$. Prove that $\#_{\mathbb{Z}} \otimes \#_{\mathbb{Z}} = \#_{\mathbb{Z}^2}$.

Chapter 7

Second moment analysis

The average $\frac{1}{n}(X_1 + \cdots + X_n)$ of independent random numbers is a random variable that is expected to take on values close to its mean when n is large. In other words, it is expected that the law of $\frac{1}{n}(X_1 + \cdots + X_n)$ concentrates most of its mass near its mean. This phenomenon is called *concentration of measure*. Most principles of statistical learning, spreading of financial risk, and thermodynamical equilibria are based on this phenomenon. There are several methods in probability theory for analysing the concentration of measure. We get introduced to some of these.

Add: Concentration of mass

7.1 Second-moment method

The second moment method studies square-integrable random variables. A real-valued random variable is called *square integrable* if $\mathbb{E}X^2 < \infty$.

Preliminaries:

- Square-integrable random variables are absolutely integrable: $\mathbb{E}X^2 < \infty \implies \mathbb{E}|X| < \infty$
- When X, Y are square integrable, then XY is absolutely integrable. (Cauchy–Schwarz, or simpler for independent)
- $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$ for independent. (easy to do first for nonnegative, then for signed)

7.2 Jensen's inequality

A function $f\colon \mathbb{R} \to \mathbb{R}$ is *convex* if $f((1-\alpha)x + \alpha y) \le (1-\alpha)f(x) + \alpha f(y)$ for all $\alpha \in [0,1]$ and $x, y \in \mathbb{R}$. The following important result, known as *Jensen's inequality*, is valid for all integrable random variables X. The random variable $f(X)$ does not need to be integrable. Jensen's inequality actually shows that $-\infty < \mathbb{E}f(X) \le \infty$.

Theorem 7.1 (Jensen's inequality). *For any integrable random variable $X\colon \Omega \to I$ with values in an interval $I \subset \mathbb{R}$,*

$$f(\mathbb{E}X) \ \le \ \mathbb{E}f(X)$$

for all convex functions $f\colon I \to \mathbb{R}$.

Proof. Because f is convex, there exists a tangent line $\{(x,y) \in \mathbb{R}^2 \colon y = ax+b\}$ such that

$$ax + b \ \le \ f(x) \qquad \text{for all } x \in I \tag{7.1}$$

and (7.1) holds as equality for $x = \mathbb{E}X$. (Exercise 7.18 confirms that $\mathbb{E}X$ is a point in I). For example, we may select $a = f'_+(x_0)$ and $b = f(x_0) - ax_0$, where $f'_+(x_0) = \lim_{h\downarrow 0} \frac{f(x_0+h)-f(x_0)}{h}$ is the right-derivative at $x_0 = \mathbb{E}X$, which is well defined whenever $\mathbb{E}X$ it not the rightmost point of I (otherwise we may replace a by the left derivative of f at x_0).

By plugging $x = X(\omega)$ into (7.1) and integrating against $\mathbb{P}$, the monotonicity of integration (Proposition 4.9) implies that

$$\mathbb{E}(aX+b) \ = \ \int_\Omega (aX+b)\,d\mathbb{P} \ \le \ \int_\Omega f(X)\,d\mathbb{P} \ = \ \mathbb{E}f(X).$$

By recalling that (7.1) holds as equality for $x = \mathbb{E}X$, and applying the linearity of integration (Proposition 4.11) to the random variables $X \in L^1(\mathbb{P})$, we find that

$$f(\mathbb{E}X) \ = \ a\mathbb{E}X + b \ = \ \mathbb{E}(aX+b).$$

The claim follows by combining the above two displays. □

7.3 Power-integrable random variables

A real-valued random variable is called p-th *power integrable* if

$$\|X\|_{L^p(\mathbb{P})} \ = \ (\mathbb{E}|X|^p)^{1/p}$$

is finite. The collection of such random variables is denoted $L^p(\mathbb{P})$. Power-integrable random variables with $p = 1$ are called *integrable*, and those with $p = 2$ *square integrable*. The following result implies in particular that $L^2(\mathbb{P}) \subset L^1(\mathbb{P})$.

Proposition 7.2. *$(\mathbb{E}|X|^p)^{1/p} \le (\mathbb{E}|X|^q)^{1/q}$ for all $0 < p \le q < \infty$ and all $\bar{\mathbb{R}}$-valued random variables, and in particular, $\mathbb{E}|X| \le (\mathbb{E}X^2)^{1/2}$.*

L^p seminorms associated with random variables are ordered in p.

Proof. Jensen's inequality (Proposition 7.1) applied to the nonnegative random variable $Y = |X|^p$ with the convex function $f(x) = x^{q/p}$ on $\mathbb{R}_+$ implies that

$$(\mathbb{E}|X|^p)^{q/p} \;=\; f(\mathbb{E}Y) \;\le\; \mathbb{E}f(Y) \;=\; \mathbb{E}|X|^q.$$

The claim follows by raising both sides above to the power $1/q$. □

Proposition 7.3 (Cauchy–Schwarz inequality)**.** *For all random variables $X, Y : \Omega \to \bar{\mathbb{R}}$,*

$$\mathbb{E}|XY| \;\le\; (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2}.$$

Products of square integrable functions are integrable.

Products of square integrable random variables are integrable random variables with a finite mean $\mathbb{E}X$.

Proof. We prove Cauchy–Schwarz for all measurable functions, and all measures μ. To avoid trivialities, let us assume that the right side is finite. We may assume that the right side is nonzero, because otherwise $(\int_S f^2\, d\mu)^{1/2} = 0$ or $(\int_S g^2\, d\mu)^{1/2} = 0$ implies (Theorem 5.1) that either f or g is zero almost everywhere, and this would imply that $\int_S |fg|\, d\mu = 0$.

Let us now assume that $\|f\|_2 = (\int_S f^2\, d\mu)^{1/2}$ and $\|g\|_2 = (\int_S g^2\, d\mu)^{1/2}$ are finite and strictly positive. By applying the inequality $\frac{1}{2}(x^2 + y^2) = \frac{1}{2}(x-y)^2 + xy \ge xy$, we find that

$$\frac{|f|}{\|f\|_2}\,\frac{|g|}{\|g\|_2} \;\le\; \frac{1}{2}\left(\frac{f^2}{\|f\|_2^2} + \frac{g^2}{\|g\|_2^2}\right) \qquad \text{pointwise.}$$

Integrating both sides against μ reveals that

$$\frac{\int |fg|\,d\mu}{\|f\|_2\|g\|_2} \;\le\; \frac{1}{2}\left(\frac{\int_S f^2\,d\mu}{\|f\|_2^2} + \frac{\int_S g^2\,d\mu}{\|g\|_2^2}\right) \;=\; 1,$$

and the claim follows by multiplying both sides by $\|f\|_2\|g\|_2$. □

Proposition 7.4. *$L^2(\mu)$ is a real vector space, and for all $f, g \in L^2(\mu)$,*

(i) $\|cf\|_{L^2(\mu)} = |c|\,\|f\|_{L^2(\mu)}$ *for all* $c \in \mathbb{R}$*,*

(ii) $\|f+g\|_{L^2(\mu)} \le \|f\|_{L^2(\mu)} + \|g\|_{L^2(\mu)}$.

Linear combinations of square-integrable random variables are square-integrable random variables.

Proof. (i) The first claim follows by noting that $\|cf\|_{L^2(\mu)}^2 = \int_S c^2 f^2\,d\mu = c^2\|f\|_{L^2(\mu)}^2$, and taking square roots.

(ii) Observe that $(f+g)^2 \le f^2 + 2|fg| + g^2$ pointwise. Monotonicity (Proposition 4.4) and linearity (Proposition 4.6) combined with the Cauchy–Schwarz inequality (Proposition 7.3) show that

$$\begin{aligned}
\|f+g\|_{L^2(\mu)}^2 &= \int (f+g)^2\,d\mu \\
&\overset{\text{mon}}{\le} \int \left(f^2 + 2|fg| + g^2\right) d\mu \\
&\overset{\text{lin}}{=} \int f^2\,d\mu + 2\int |fg|\,d\mu + \int g^2\,d\mu \\
&\overset{\text{CS}}{\le} \int f^2\,d\mu + 2\Big(\int f^2\,d\mu\Big)^{1/2}\Big(\int g^2\,d\mu\Big)^{1/2} + \int g^2\,d\mu \\
&= \Big(\|f\|_{L^2(\mu)} + \|g\|_{L^2(\mu)}\Big)^2,
\end{aligned}$$

so the claim follows by taking square roots. □

7.4 Independent products

The following result is needed later for proving that independent square-integrable random variables have zero covariance (Proposition 7.8).

Proposition 7.5. *For all independent random numbers such that either $X, Y \geq 0$ or $X, Y \in L^1(\mathbb{P})$,*

$$\mathbb{E}XY \;=\; \mathbb{E}X\,\mathbb{E}Y. \tag{7.2}$$

Independence allows to factorise, also in the case of expected values.

Proof. (i) Assume that $X, Y\colon \Omega \to \bar{\mathbb{R}}_+$ have finite range. Then $X = \sum_{i=1}^m a_i 1_{A_i}$ and $Y = \sum_{j=1}^n b_j 1_{B_j}$ for some $a_i, b_j \in \bar{\mathbb{R}}_+$ with $A_i = X^{-1}(\{a_i\})$ and $B_j = Y^{-1}(\{b_j\})$. Proposition 4.2 implies that

$$\mathbb{E}X\,\mathbb{E}Y \;=\; \left(\sum_{i=1}^m a_i \mathbb{P}(A_i)\right)\left(\sum_{j=1}^n b_j \mathbb{P}(B_j)\right) \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathbb{P}(A_i)\mathbb{P}(B_j). \tag{7.3}$$

We also note that

$$XY \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j 1_{A_i} 1_{B_j} \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j 1_{A_i \cap B_j},$$

so that by linearity (Proposition 4.6),

$$\mathbb{E}XY \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathbb{E}1_{A_i \cap B_j} \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathbb{P}(A_i \cap B_j). \tag{7.4}$$

The independence of X and Y implies that $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j)$. By combining this observation with (7.3) and (7.4), we conclude that (7.2) holds.

(ii) Now consider general random variables $X, Y\colon \Omega \to \bar{\mathbb{R}}_+$. Fix finite-range random variables $X_n \uparrow X$ and $Y_n \uparrow Y$. Then $X_n Y_n \uparrow XY$, and monotone continuity (Theorem 4.5) implies that

$$\mathbb{E}XY \;=\; \lim_{n\to\infty} \mathbb{E}X_n Y_n \;\overset{\text{(i)}}{=}\; \lim_{n\to\infty} \mathbb{E}X_n\,\mathbb{E}Y_n \;=\; \mathbb{E}X\,\mathbb{E}Y.$$

Hence (7.2) holds.

(iii) Now consider independent random variables $X, Y \in L^1(\mathbb{P})$. Then $\phi(X) \perp\!\!\!\perp \psi(Y)$ for all measurable functions $\phi, \psi\colon \mathbb{R} \to \mathbb{R}_+$. Then by applying (ii) we see that $\mathbb{E}\phi(X)\psi(Y) = \mathbb{E}\phi(X)\,\mathbb{E}\phi(Y)$. In particular, $\phi(X)\psi(Y) \in L^1(\mathbb{P})$ for all measurable $\phi, \psi\colon \mathbb{R} \to \mathbb{R}_+$ bounded pointwise by $\phi(x) \leq |x|$ and $\psi(x) \leq |x|$. It follows that $XY \in L^1(\mathbb{P})$, and

$$XY \;=\; (X_+ - X_-)(Y_+ - Y_-) \;=\; X_+Y_+ + X_-Y_- - X_+Y_- - X_-Y_+,$$

where all random variables appearing on the right are in $L^1(\mathbb{P})$. Hence by linearity (Proposition 4.11),

$$\begin{aligned}
\mathbb{E}XY &= \mathbb{E}X_+Y_+ + \mathbb{E}X_-Y_- - \mathbb{E}X_+Y_- - \mathbb{E}X_-Y_+ \\
&= \mathbb{E}X_+\,\mathbb{E}Y_+ + \mathbb{E}X_-\,\mathbb{E}Y_- - \mathbb{E}X_+\,\mathbb{E}Y_- - \mathbb{E}X_-\,\mathbb{E}Y_+ \\
&= (\mathbb{E}X_+ - \mathbb{E}X_-)(\mathbb{E}Y_+ - \mathbb{E}Y_-) \\
&= \mathbb{E}X\,\mathbb{E}Y.
\end{aligned}$$

□

7.5 Variances and covariances

The *variance* of a random variable $X \in L^2(\mathbb{P})$ is defined by

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

Proposition 7.2 implies that the mean $\mathbb{E}X$ is a well-defined real number for any $X \in L^2(\mathbb{P})$. Because $L^2(\mathbb{P})$ is a vector space (Proposition 7.4), and constant random variables are square integrable, we see that $X - \mathbb{E}X \in L^2(\mathbb{P})$. Therefore, the variance is a well-defined number in $\mathbb{R}_+$.

Proposition 7.6. *For any square-integrable random variable,* $\mathrm{Var}(X + t) = \mathrm{Var}(X)$ *and* $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$ *for all* $t \in \mathbb{R}$ *and* $c \in \mathbb{R}_+$*.*

The variance operator is shift invariant, but not scale invariant.

Proof. Exercise. todo □

Proposition 7.7. *For any square-integrable random variable,* $\mathrm{Var}(X) = 0$ *if and only if* $X = \mathbb{E}X$ *almost surely.*

Zero variance means zero randomness.

Proof. Proposition 5.1 implies that $\mathbb{E}(X - \mathbb{E}X)^2 = 0$ if and only if $X - \mathbb{E}X = 0$ almost surely. □

The *covariance* of random variables $X, Y \in L^2(\mathbb{P})$ is defined by

$$\operatorname{Cov}(X,Y) \;=\; \mathbb{E}\Big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\Big).$$

The Cauchy–Schwarz inequality (Proposition 7.3) implies that $\operatorname{Cov}(X,Y)$ is a well-defined real number and bounded by

$$|\operatorname{Cov}(X,Y)| \;\le\; \operatorname{Var}(X)^{1/2}\operatorname{Var}(Y)^{1/2}.$$

Proposition 7.8. $\operatorname{Cov}(X,Y) = 0$ *for independent square-integrable random variables.*

Independence implies zero covariance, but not vice versa.

Proof. When X, Y are square integrable, so are the random variables $\tilde{X} = X - \mathbb{E}X$ and $\tilde{Y} = Y - \mathbb{E}Y$. In particular, $\tilde{X}$ and $\tilde{Y}$ are independent and in $L^1(\mathbb{P})$. By linearity (Proposition 4.11), $\mathbb{E}\tilde{X} = 0$ and $\mathbb{E}\tilde{Y} = 0$. With the help of Proposition 7.5, we find that

$$\operatorname{Cov}(X,Y) \;=\; \mathbb{E}\tilde{X}\tilde{Y} \;=\; \mathbb{E}\tilde{X}\,\mathbb{E}\tilde{Y} \;=\; 0.$$ □

Proposition 7.9. *For all square-integrable random variables X_i, Y_j, and all real numbers a_i, b_j,*

$$\operatorname{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) \;=\; \sum_{i=1}^m \sum_{j=1}^n a_i b_j \operatorname{Cov}(X_i, Y_j).$$

Covariance is a bilinear functional on $L^2(\mathbb{P})$.

Proof. Because $L^2(\mathbb{P})$ is a vector space (Proposition 7.4), we see that the centered random variables $\tilde{X}_i = X_i - \mathbb{E}X_i$ and $\tilde{Y}_j = Y_j - \mathbb{E}Y_j$ are square integrable, and hence integrable (Proposition 7.2). The same conclusions are true for the random variables $X = \sum_{i=1}^m a_i X_i$ and $Y = \sum_{j=1}^n b_j Y_j$. By linearity (Proposition 4.11), we see that

$$X - \mathbb{E}X \;=\; \sum_{i=1}^m a_i \tilde{X}_i, \qquad Y - \mathbb{E}Y \;=\; \sum_{j=1}^n b_j \tilde{Y}_j.$$

By multiplying the above equalities, we find that

$$(X - \mathbb{E}X)(Y - \mathbb{E}Y) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \tilde{X}_i \tilde{Y}_j.$$

Because $\tilde{X}_i, \tilde{Y}_j \in L^2(\mathbb{P})$, Cauchy–Schwarz inequality (Proposition 7.3) tells us that $\tilde{X}_i\tilde{Y}_j \in L^1(\mathbb{P})$. Hence by taking expectations on both sides of the above equality, and applying linearity, it follows that

$$\operatorname{Cov}(X,Y) \;=\; \mathbb{E}\Big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\Big) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathbb{E}\tilde{X}_i \tilde{Y}_j.$$

The claim follows by noting that $\mathbb{E}\tilde{X}_i\tilde{Y}_j = \operatorname{Cov}(X_i, Y_j)$. □

Proposition 7.10. *The variances of the sum* $S_n = \sum_{i=1}^n X_i$ *and the arithmetic average* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *of independent square-integrable random variables* $X_1, \dots, X_n$ *are given by*

$$\operatorname{Var}(S_n) = \sum_{i=1}^n \operatorname{Var}(X_i) \quad \textit{and} \quad \operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n},$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \operatorname{Var}(X_i)$.

The variance of the arithmetic average of independent and identically distributed square-integrable random variables tends to zero as $n \to \infty$.

Proof. The independence assumption combined with Proposition 7.8 tells us that $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$. Observe also that $\operatorname{Var}(S_n) = \operatorname{Cov}(S_n, S_n)$ by definition. The bilinearity of the covariance operation (Proposition 7.9) then implies that

$$\begin{aligned}
\operatorname{Var}(S_n) \;&=\; \operatorname{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \\
&=\; \sum_{i=1}^n\sum_{j=1}^n \operatorname{Cov}(X_i, X_j) \\
&=\; \sum_{i=1}^n \operatorname{Cov}(X_i, X_i) \;=\; \sum_{i=1}^n \operatorname{Var}(X_i).
\end{aligned}$$

Observe next that $\bar{X}_n = n^{-1}S_n$. Therefore, Proposition 7.6 implies that $\operatorname{Var}(\bar{X}_n) = n^{-2}\operatorname{Var}(S_n) = n^{-1}\sigma^2$. □

This formula is remarkable as it shows that the variance of the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$, that is itself a random variable, decreases to zero as $n \to \infty$.

7.6 Markov's and Chebyshev's inequalities

Lemma 7.11 (Markov's inequality)**.** *For any $t > 0$ and any nonnegative random variable X, $\mathbb{P}(X \ge t) \le t^{-1}\mathbb{E}X$.*

Proof. Denote $A = \{X \ge t\}$. Because $t1_A \le X1_A \le X$ pointwise, the monotonicity of integration tells us that

$$\mathbb{P}(X \ge t) \;=\; \mathbb{E}1_A \;=\; t^{-1}\mathbb{E}t1_A \;\le\; t^{-1}\mathbb{E}X.$$

□

Lemma 7.12 (Chebyshev's inequality)**.** *For any $\epsilon > 0$ and any square-integrable random variable X, $\mathbb{P}(|X - \mathbb{E}X| \ge \epsilon) \le \epsilon^{-2}\operatorname{Var}(X)$.*

Proof. Note that $\mathbb{P}(|X - \mathbb{E}X| \ge \epsilon) = \mathbb{P}((X - \mathbb{E}X)^2 \ge \epsilon^2)$. Then apply Markov's inequality (Lemma 7.11) to the nonnegative random variable $(X - \mathbb{E}X)^2$. □

7.7 Weak law of large numbers

We say that a sequence of random variables X_n *converges in probability* to random variable X if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$.

Theorem 7.13. *Let $X_1, \dots, X_n$ be mutually independent square-integrable random numbers with a common mean m and common variance σ^2. Then*

$$\frac{1}{n}\sum_{i=1}^n X_i \;\to\; m \quad \textit{in probability.}$$

Proof. Denote $M_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\mathbb{E}(M_n) = m$ and $\operatorname{Var}(M_n) = \frac{\sigma^2}{n}$. Chebyshev's inequality (Lemma 7.12) tells that for any $\epsilon > 0$,

$$\mathbb{P}(|M_n - m| > \epsilon) \;\le\; \epsilon^{-2}\operatorname{Var}(M_n) \;=\; \frac{\sigma^2}{\epsilon^2 n} \;\to\; 0$$

as $n \to \infty$. □

7.8 Exercises

Exercise 7.14 (Laws of independence)**.** Which of the following statements are true in general for real-valued random variables? Explain why a statement is generally true, or give a counterexample.

(a) $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp Z \implies X \perp\!\!\!\perp Z$.

(b) $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z \implies X \perp\!\!\!\perp (Y, Z)$.

(c) $X \perp\!\!\!\perp Y$ and $X, Y \geq 0 \implies \mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y$.

Exercise 7.15 (Eleventh hour)**.** Let $T = H + \frac{1}{60}M + \frac{1}{3600}S$, in which H, M, S are mutually independent random variables such that $\mathrm{Law}(H) = \frac{1}{24}\sum_{h=0}^{23}\delta_h$, $\mathrm{Law}(M) = \frac{1}{60}\sum_{m=0}^{59}\delta_m$, and $\mathrm{Law}(S) = \frac{1}{60}\lambda_{(0,60)}$, and $\lambda_{(0,60)}$ denotes the restriction of the Lebesgue measure on the interval $(0, 60)$.

(a) Determine $\mathbb{E}T$.

(b) Determine $\mathrm{Var}\,T$.

(c) Determine $\mathrm{Cov}(H, T)$.

(d) Determine $\mathbb{P}(23\frac{3599}{3600} < T < 24)$.

Exercise 7.16 (Triangles in a random graph)**.** Denote by V_k the family of unordered k-element subsets of $V = \{1, \dots, n\}$. Let $(I_A \colon A \in V_2)$ be mutually independent $\mathrm{Ber}(p)$-distributed random variables. These random variables can be used to generate a random graph $G = (V, E)$ with node set V and link set $E = \{A \in V_2 \colon I_A = 1\}$, in which the number of triangles equals

$$T = \sum_{B \in V_3} \theta_B, \qquad \text{where} \quad \theta_B = \prod_{A \in V_2 : A \subset B} I_A.$$

(a) Explain why the random variables $(\theta_B \colon B \in V_3)$ are not independent.

(b) Determine the law of the random variable θ_B and verify that $\mathbb{E}\theta_B = p^3$ and $\mathrm{Var}(\theta_B) = p^3(1 - p^3)$.

(c) Prove that $\mathrm{Cov}(\theta_B, \theta_{B'}) = 0$ when $|B \cap B'| \leq 1$ and compute the value of $\mathrm{Cov}(\theta_B, \theta_{B'})$ when $|B \cap B'| = 2$.

(d) With the help of (b)–(c), prove that $\mathbb{E}T = \binom{n}{3}p^3$ and $\mathrm{Var}(T) = 12\binom{n}{4}(p^5 - p^6) + \binom{n}{3}(p^3 - p^6)$.
(Hint: Recall the proof of Proposition 7.10.)

Exercise 7.17 (Triangles in a sparse random graph)**.** Let T_n be the number of triangles in a random graph G_n with node count n and link probability p_n, in which all nodes pairs are linked with probability p_n, independently of other pairs. We saw in Exercise 7.16 that $\mathbb{E}T_n = \binom{n}{3}p_n^3$ and $\mathrm{Var}(T_n) = 12\binom{n}{4}(p_n^5 - p_n^6) + \binom{n}{3}(p_n^3 - p_n^6)$. Prove that

(a) $\lim_{n\to\infty} \mathbb{P}(T_n = 0) = 1$ when $np_n \to 0$.
(Hint: Markov's inequality; $T_n > 0 \iff T_n \geq 1$.)

(b) $\lim_{n\to\infty} \mathbb{P}(T_n > 0) = 1$ when $np_n \to \infty$.
(Hint: Chebyshev's inequality; $T_n = 0 \implies |T_n - \mathbb{E}T_n| \geq \mathbb{E}T_n$.)

Exercise 7.18 (Range of expectations)**.** Let X be a random variable such that $\mathbb{P}(X \in (0,1)) = 1$.

(a) Prove that $\mathbb{E}X > 0$.

(b) Prove that $\mathbb{E}X \in (0,1)$.

Hint: Recall Theorem 5.1 and Proposition 5.3.

Chapter 8

Random sequences and limits

Key concepts: IID sequence, convergence in probability, convergence almost surely

Learning outcomes:

- Learn to generate independent coin flips from a single random number
- Learn to generate independent random variables from a single random number

Prerequisites: Previous chapters.

8.1 Cumulative distribution functions

The *cumulative distribution function* of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a function $F\colon \mathbb{R} \to [0,1]$ defined by

$$F(t) \ = \ \mu((-\infty, t]).$$

The cumulative distribution function F_X of a real-valued random variable X is defined as the cumulative distribution function of the law of X, so that

$$F_X(t) \ = \ \mathbb{P}(X \le t).$$

Proposition 8.1. *Any cumulative distribution is nondecreasing, right-continuous, and has limits* $\lim_{t\to-\infty} F(t) = 0$ *and* $\lim_{t\to\infty} F(t) = 1$.

Proof. (i) For any $s \le t$, we see that $(-\infty, s] \subset (-\infty, t]$, so that by monotonicity,

$$F(s) \ = \ \mu((-\infty, s]) \ \le \ \mu((-\infty, t]) \ = \ F(t).$$

(ii) For any t, monotone continuity combined with the observation that $(-\infty, t + \frac{1}{n}] \downarrow (-\infty, t]$ implies that

$$\lim_{n\to\infty} F(t + 1/n) \ = \ \lim_{n\to\infty} \mu((-\infty, t + 1/n]) \ = \ \mu((-\infty, t]) \ = \ F(t).$$

(iii) Because $(-\infty, -n] \downarrow \emptyset$ and $(-\infty, n] \uparrow \mathbb{R}$ as $n \to \infty$, we find by monotone continuity that

$$\lim_{n\to-\infty} F(n) \ = \ \lim_{n\to\infty} \mu(-\infty, -n] \ = \ \mu(\emptyset) \ = \ 0$$

and

$$\lim_{n\to\infty} F(n) \ = \ \lim_{n\to\infty} \mu(-\infty, n] \ = \ \mu(\mathbb{R}) \ = \ 1.$$

□

Proposition 8.2. *If the cumulative distribution function of μ of is differentiable everywhere with derivative $F'(t) = f(t)$ being a measurable function, then f is a density function of μ with respect to the Lebesgue measure, so that*

$$\mu(B) \ = \ \int_B f(t)\, dt.$$

Proof. This follows by the fundamental theorem of calculus. □

8.2 Quantile functions

A *quantile function* of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a function $Q\colon (0,1) \to \mathbb{R}$ such that

$$\mu\Big((-\infty, Q(p))\Big) \ \le \ p \ \le \ \mu\Big((-\infty, Q(p)]\Big) \qquad \text{for all } p \in (0,1). \tag{8.1}$$

A quantile function of a real-valued random variable X is a quantile function of the law of X.

A quantile function of X assigns to every probability value $p \in (0,1)$ a number $t = Q(p)$ such that

$$\mathbb{P}(X < t) \ \le \ p \qquad \text{and} \qquad \mathbb{P}(X > t) \ \ge \ 1 - p.$$

If the cumulative distribution function F of μ is a strictly increasing function, then $F\colon \mathbb{R} \to (0,1)$ is invertible, and the inverse function F^{-1} is the unique quantile function of μ. (See Proposition 8.8.) In general, a quantile function always exists, but is usually nonunique.

In what follows, $f(x+) = \lim_{y \downarrow x} f(y)$ and $f(x-) = \lim_{y \uparrow x} f(y)$. Then (8.1) can be written in terms of the cumulative distribution function F as

$$F(Q(p)-) \ \le \ p \ \le \ F(Q(p)). \tag{8.2}$$

Lemma 8.3. *Any quantile function of a probability measure μ with cumulative distribution function F is nondecreasing and satisfies*

(i) $Q(s) > x \implies s \ge F(x)$,

(ii) $Q(s) \le x \implies s \le F(x)$,

(iii) $s < F(x) \implies Q(s) \le x$,

(iv) $s > F(x) \implies Q(s) > x$,

(v) $Q(F(x)-) \le x \le Q(F(x)+)$,

(vi) $(F(a), F(b)) \subset Q^{-1}(a,b] \subset [F(a), F(b)]$,

(vii) $F^{-1}(s,t) \subset [Q(s), Q(t)) \subset F^{-1}[a,b]$.

Proof. Let us first verify that Q is increasing. Fix some numbers in $(0,1)$ such that $s_1 \le s_2$. We will show that in this case $Q(s_1) > Q(s_2)$ leads into a contradiction. So, assume now that $Q(s_1) > Q(s_2)$ were true. Then because F is increasing and right-continuous, it follows that $F(Q(s_1)-) \geq F(Q(s_2))$, and hence (8.2) implies that

$$s_1 \;\geq\; F(Q(s_1)-) \;\geq\; F(Q(s_2)) \;\geq\; s_2.$$

We also have $s_1 \neq s_2$ because $Q(s_1) \neq Q(s_2)$. Therefore we conclude that $s_1 > s_2$ which is the desired contradiction.

To prove (i), note first that $F(x) \leq F(y-)$ for all $x < y$. Implication (i) hence follows from the first inequality in (8.2). On the other hand, implication (ii) follows immediately from the second inequality in (8.2). Furthermore, (iii) and (iv) follow by negating the implications (i) and (ii).

By taking a limit $s \uparrow F(x)$, we see with the help of (iii) that $Q(F(x)-) \leq x$. Similarly by taking a limit $s \downarrow F(x)$, we see with the help of (iv) that $Q(F(x)+) \geq x$. Hence we have shown (v).

By combining (i) and (ii), we obtain the implications

$$\begin{aligned} a < Q(s) \leq b \;&\Longrightarrow\; F(a) \leq s \leq F(b), \\ Q(s) \leq x < Q(t) \;&\Longrightarrow\; s \leq F(x) \leq t \end{aligned}$$

and by combining (iii) and (iv), the implications

$$\begin{aligned} F(a) < s < F(b) \;&\Longrightarrow\; a < Q(s) \leq b, \\ s < F(x) < t \;&\Longrightarrow\; Q(s) \leq x < Q(t), \end{aligned}$$

from which we may confirm the validity of (vi) and (vii).

□

Proposition 8.4. *Let Q be a quantile function of a probability measure μ, and let U be a uniformly distributed random number in $(0,1)$. Then the law of $Q(U)$ equals μ.*

For any probability measure μ on the real line, we may sample a μ-distributed random variable X using a quantile function by setting $X = Q(U)$ where U a uniformly distributed random number in $(0,1)$.

Proof. Note first that Q is measurable, being a nondecreasing function. Denote $Q^{-1}(-\infty, x] = \{s \in (0,1) : Q(s) \leq x\}$. By Lemma 8.3 we see that $Q(s) \leq x$ for $s < F(x)$. Hence $(0, F(x)) \subset Q^{-1}(-\infty, x]$. Lemma 8.3 also shows that $Q(s) \leq x$ implies $s \leq F(x)$, so we may conclude that

$$(0, F(x)) \subset Q^{-1}(-\infty, x] \subset (0, F(x)].$$

Hence

$$\mathbb{P}(U < F(x)) \ \le \ \mathbb{P}(Q(U) \le x) \ \le \ \mathbb{P}(U \le F(x)),$$

from which we may conclude that $\mathbb{P}(Q(U) \le x) = F(x)$ for all $x \in \mathbb{R}$. Therefore, the cumulative distribution function of $Q(U)$ equals F. □

Proposition 8.5. *If Q is a quantile function of a random variable X, then*

$$\mathbb{E}\,\phi(X)1(a < X \le b) \ = \ \int_{F(a)}^{F(b)} \phi(Q(u))du$$

for any ϕ such that the expectation on the left exists. Also,

$$\begin{aligned} \mathbb{E}\,\phi(X)1(X > a) \ &= \ \int_{F(a)}^{1} \phi(Q(u))du, \\ \mathbb{E}\,\phi(X)1(X \le b) \ &= \ \int_{0}^{F(b)} \phi(Q(u))du, \end{aligned}$$

whenever the expectations on the left exist.

Proof. Denote $I = \mathbb{E}\,\phi(X)1_{(a,b]}(X)$ and assume first that $\phi \ge 0$. By Proposition 8.4,

$$I = \int_0^1 \phi(Q(u))1_{(a,b]}(Q(u))\,du.$$

By Lemma 8.3,

$$1_{(F(a),F(b))}(u) \ \le \ 1_{Q^{-1}(a,b]}(u) \ \le \ 1_{[F(a),F(b)]}(u)$$

Hence

$$\int_0^1 \phi(Q(u))1_{(F(a),F(b))}(u)\,du \ \le \ I \ \le \ \int_0^1 \phi(Q(u))1_{[F(a),F(b)]}(u)\,du.$$

The first claim follows for $\phi \ge 0$ because both sides above are equal to $\int_{F(a)}^{F(b)} \phi(Q(u))du$. The same claim follows for signed ϕ by treating the positive and negative separately. The latter claims follow by taking limits $a \to -\infty$ and $b \to \infty$. □

The *left-continuous quantile function* of a probability measure with cumulative distribution function F is defined by

$$Q(s) = \sup\{x \in \mathbb{R} : F(x) < s\}. \tag{8.3}$$

By noting that the set on the right side of (8.3) is nonempty and bounded from above for every $s \in (0,1)$, it follows that the above formula defines a function $Q\colon (0,1) \to \mathbb{R}$.

Lemma 8.6. *For any $s \in (0,1)$,*

$$\sup\{x \in \mathbb{R} : F(x) < s\} \;=\; \inf\{x \in \mathbb{R} : F(x) \geq s\}. \tag{8.4}$$

Proof. Fix $s \in (0,1)$. Let $A = \{x : F(x) < s\}$ and $B = \{x : F(x) \geq s\}$. The A and B form a partition of the real line. Also, the monotonicity of F implies that $x < y$ for all $x \in A$ and $y \in B$. Hence we see that $\sup A \leq \inf B$. This inequality cannot be strict because otherwise there would exists a number z such that $\sup A < z < \inf B$, and such a number could not belong to A or B, which contradicts the fact that A and B form a partition. Hence $\sup A = \inf B$ and the first claim is proved. □

Proposition 8.7. *The left-continuous quantile function is a quantile function.*

Proof. We need to verify that the function Q defined by (8.3) satisfies (8.2). Fix $s \in (0,1)$. Formula (8.3) implies that there exists a sequence $x_n \uparrow Q(s)$ such that $x_n \in \{x \colon F(x) < s\}$ for all n. In particular, $F(x_n) < s$ for all n, and it follows that

$$F(Q(s)-) \;=\; \lim_{n\to\infty} F(x_n) \;\leq\; s.$$

Formula (8.4) implies that there exists a sequence $y_n \downarrow Q(s)$ such that $y_n \in \{x \colon F(x) \geq s\}$ for all n. In particular, $F(y_n) \geq s$ for all n, and it follows that

$$F(Q(s)+) \;=\; \lim_{n\to\infty} F(y_n) \;\geq\; s.$$

The right-continuity of F implies that $F(Q(s)) = F(Q(s)+) \geq s$, and the claim follows. □

Proposition 8.8. *Any probability measure with a strictly increasing cumulative distribution function $F \colon \mathbb{R} \to (0,1)$ has a unique quantile function given by $Q = F^{-1}$, the inverse of F.*

Proof. Assume that F is invertible with inverse function[1] F^{-1}. Then $F(F^{-1}(p)) = p$. Because F is increasing, $F(x-) \leq F(x)$ for all x. In particular this is true for $x = F^{-1}(p)$. Hence $F(F^{-1}(p)-) \leq p \leq F(F^{-1}(p))$. Hence (8.2) is valid, and F^{-1} is a quantile function.

Assume now that Q is a quantile function in the sense of (8.2). Fix $p \in (0,1)$ and denote $t = Q(p)$. ... FINISHME □

[1] Usually $F^{-1}(B)$ refers to a preimage, but now $F^{-1}(p)$ is the inverse function of F.

8.3 Random sequences

An *random sequence* is a list $(X_1, X_2, \dots)$ in which $X_i \colon \Omega \to S_i$ is a random variable for each i, and each S_i is a measurable space equipped with a sigma-algebra $\mathcal{S}_i$.

A random sequence can also be viewed as a random variable $X \colon \Omega \to S_1 \times S_2 \times \cdots$ where we equip the infinite product space with the *product sigma-algebra*

$$\mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots \ = \ \sigma\Big(\pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2) \cup \cdots\Big).$$

Usually, all the sets $S_1 = S_2 = \cdots = S$. Then we write $S^\infty = S_1 \times S_2 \times \cdots$

Proposition 6.2 extends to infinite products, and justifies the above statements.

Proposition 8.9. *A function $f \colon S_0 \to S_1 \times S_2 \times \cdots$ is measurable if and only if its coordinate functions $f_i = \pi_i \circ f$ are measurable.*

Proof. By definition, the set family $\mathcal{C} = \pi_1^{-1}(\mathcal{S}_1) \cup \pi_2^{-1}(\mathcal{S}_2) \cup \cdots$ is a generator of $\mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots$. Therefore, by Proposition 3.3, f is $\mathcal{S}_0/(\mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots)$-measurable if and only if $f^{-1}(C) \in \mathcal{S}_0$ for all $C \in \mathcal{C}$. Because all sets in $\mathcal{C}$ are either of the form $C = \pi_i^{-1}(B)$ for some $B \in \mathcal{S}_i$ and some i, we see that

$$\begin{aligned}
& f \text{ is } \mathcal{S}_0/(\mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots)\text{-measurable} \\
\iff & f^{-1}(C) \in \mathcal{S}_0 \text{ for all } C \in \mathcal{C} \\
\iff & f^{-1}(\pi_i^{-1}(B)) \in \mathcal{S}_0 \text{ for all } B \in \mathcal{S}_i \text{ and for all } i \ge 1 \\
\iff & (\pi_i \circ f)^{-1}(B) \in \mathcal{S}_0 \text{ for all } B \in \mathcal{S}_i \text{ and for all } i \ge 1 \\
\iff & \pi_i \circ f \text{ is } \mathcal{S}_0/\mathcal{S}_i\text{-measurable for all } i \ge 1.
\end{aligned}$$

□

8.4 Generating independent random sequences

A sequence of random variables $X_1, X_2, \dots$ with values in a measurable space $(S, \mathcal{S})$ is *stochastically independent*, if

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) \ = \ \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n)$$

for all integers $n \ge 1$ and all $B_1, \dots, B_n \in \mathcal{S}$.

How do we generate a sequence of independent random variables $X_1, X_2, \dots$, all distributed according to a probability measure μ, on some probability space? Is this always possible?

Assume that we have a random variable U that is uniformly distributed in $[0,1]$. Let us compute a binary expansion for a number $u = 0.b_1b_2\cdots$ in $[0,1]$. To fix a unique representation, we require that 0 has binary representation $0.000\cdots$ and all nonzero numbers in $[0,1]$ have infinitely many 1's on their binary representations. For example, $0.5 = 0.011111\cdots$.

This is a 'decision tree' (check boundaries, or don't care, as they have probability zero)

1. Set $b_1 = 1_{(\frac{1}{2},\frac{2}{2}]}(u)$
2. Set $b_2 = 1_{(\frac{1}{4},\frac{1}{2}]}(u) + 1_{(\frac{3}{4},\frac{4}{4}]}(u)$.
3. Set $b_3 = \dots$
4. In general, set $b_n = c_n - 2c_{n-1}$ for $c_n = \lceil 2^n u - 1 \rceil$.
5. Then $u = \sum_{n=1}^{\infty} b_n 2^{-n}$.

Proposition 8.10. *For* $U = \sum_{n=1}^{\infty} B_n 2^{-n}$ *with* $B_n \in \{0,1\}$*, the following are equivalent:*

(i) U *is uniformly distributed in* $(0,1)$*.*

(ii) $B_1, B_2, \dots$ *are independent and distributed according to* $\mathrm{Ber}(\frac{1}{2})$*.*

We may sample an infinite sequence of fair coin flips by a deterministic maps determined from a single uniformly distributed random number U*.*

Proof. (i) $\implies$ (ii). Assume that U is a uniformly distributed random variable defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $U = 0.B_1B_2\cdots$ be its binary representation. Then each $B_i \colon \Omega \to \{0,1\}$ is a random variable. Then $\mathbb{P}(B_1 = k_1) = \frac{1}{2}$ for all $k_1 \in \{0,1\}$. Also, $\mathbb{P}(B_1 = k_1,\, B_2 = k_2) = \frac{1}{4}$ for all $k_1, k_2 \in \{0,1\}$. This generalises to

$$\mathbb{P}(B_1 = k_1, \dots, B_n = k_n) \;=\; 2^{-n}$$

for all $k_1, \dots, k_n \in \{0,1\}$. If sum this over $k_1, \dots, k_{n-1}$, we see that $\mathbb{P}(B_n = k_n) = \frac{1}{2}$ for all $k_n \in \{0,1\}$. Hence B_n is $\mathrm{Ber}(\frac{1}{2})$-distributed, for all n. Then it follows that

$$\mathbb{P}(B_1 = k_1, \dots, B_n = k_n) \;=\; \mathbb{P}(B_1 = k_1) \cdots \mathbb{P}(B_n = k_n).$$

Because this is true for all integers n, we conclude that $B_1, B_2, \ldots$ are mutually independent and $\mathrm{Ber}(\frac{1}{2})$-distributed random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$.

Alternatively, $B = (B_1, B_2, \ldots)$ is a random variable $\Omega \to \{0,1\}^\infty$. Its law is a probability measure on $(\{0,1\}^\infty, (2^{\{0,1\}})^{\otimes\infty})$.

(ii) $\implies$ (i). Assume that $B_1, B_2, \ldots$ are mutually independent and $\mathrm{Ber}(\frac{1}{2})$-distributed. Let U' be a uniformly distributed in $(0,1)$ and let $B_1', B_2', \ldots$ be the binary expansion of U'. Then the sequences $(B_1, B_2, \ldots)$ and $(B_1', B_2', \ldots)$ have the same law in $\{0,1\}^\infty$. Explain this in detail. As a consequence, also the random variables $U = \sum_{n=1}^\infty B_n 2^{-n}$ and $U' = \sum_{n=1}^\infty B_n' 2^{-n}$ have the same law. Hence the law of U is uniform in $(0,1)$. □

For any probability distribution μ on the real line, there exists a deterministic algorithm that takes one perfect sample U from the uniform distribution as its input, and outputs an infinite sequence of independent μ-distributed random variables $X_1, X_2, \ldots$

Proposition 8.11. *Given any probability measures $\mu_1, \mu_2, \ldots$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, there exist deterministic functions $f_i \colon [0,1] \to \mathbb{R}$ such that when U is a perfect uniform sample, then the random variables $X_1 = f_1(U), X_2 = f_2(U), \ldots$ are mutually independent and distributed according to $\mathrm{Law}(X_i) = \mu_i$ for all i.*

Proof. Fix a random variable $U \in [0,1]$. Compute random variables $U_1, U_2, \ldots$ as the output of the following infinite matrix product, in which $B_1, B_2, \dots \in \{0,1\}$ are the digits in the binary expansion of U:

$$
\begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ \vdots \end{bmatrix}
=
\begin{bmatrix}
B_1 & B_2 & B_4 & B_5 & B_9 & \cdots \\
B_3 & B_5 & B_6 & \iddots & & \\
B_6 & B_7 & \iddots & & & \\
B_8 & \iddots & & & & \\
\iddots & & & & &
\end{bmatrix}
\begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{8} \\ \frac{1}{16} \\ \vdots \end{bmatrix}
$$

Proposition 8.10 implies that $B_1, B_2, \ldots$ are independent and $\mathrm{Ber}(\frac{1}{2})$-distributed. The same proposition also implies that each U_i is uniformly distributed in $(0,1)$. Because the rows of the above infinite random matrix are independent, it follows that the random variables $U_1, U_2, \ldots$ are independent. We have thus generated an infinite sequence of independent random variables $U_1, U_2, \ldots$ from a single seed random variable U.

Finally, define $X_i = Q_i(U_i)$ where Q_i is a quantile function of μ_i, for example the left-continuous quantile function defined in (8.3). Then define $f_i = Q_i \circ h \circ g_i$ where g_i maps a number u into the i-th row of infinite matrix, and $h \colon \{0,1\}^\infty \to [0,1]$ maps a binary sequence (the i-th row of the matrix) into $g(b) = \sum_{n=1}^\infty b_n 2^{-n}$. This gives us what we wanted. □

8.5 Exercises

Exercise 8.12 (Close in expectation)**.** Let X_n, X be a real-valued random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|X_n - X| \le \frac{1}{n}$ for all $n \ge 1$. For each of the following, prove that the statement is true, or give a counterexample confirming that the statement is false.

(a) $X_n \xrightarrow{\mathbb{P}} X$.

(b) $\mathbb{P}(X_n = 0) \to \mathbb{P}(X = 0)$.

(c) $W_1(\mathrm{Law}(X_n), \mathrm{Law}(X)) \to 0$, where the Wasserstein distance of order 1 between probability measures μ and ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is defined by

$$W_1(\mu, \nu) \;=\; \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^2} |x-y| \, \gamma(dx, dy),$$

with $\Gamma(\mu, \nu)$ denoting the collection of probability measures on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ having first marginal μ and second marginal ν.

Hint: Markov's inequality $\mathbb{P}(Z \ge \epsilon) \le \epsilon^{-1}\mathbb{E}Z$, valid for $Z \ge 0$ and $\epsilon > 0$, may be helpful for answering one of (a)–(c).

Chapter 9

Probability metrics

A concept of distance between the laws of two random variables corresponds to a metric on a space of probability measures. In this chapter we get introduced to the total variation distance.

Key concepts: total variation distance, coupling

Learning outcomes:

- Learn to think of probability distributions as points in a geometric space.
- Learn to apply couplings to get upper bounds on the total variation distance.
- Recognise the connection between L^1-distances of densities and the total variation distance between probability measures.

Prerequisites: Previous chapters.

9.1 Total variation distance

The *total variation distance* between probability measures μ_1 and μ_2 on a measurable space $(S, \mathcal{S})$ is defined by

$$d_{\mathrm{tv}}(\mu_1, \mu_2) \;=\; \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_2(A)|. \tag{9.1}$$

Proposition 9.1. *The total variation distance d_{tv} is a metric on the space of probability measures on $(S, \mathcal{S})$.*

Probability measures on $(S, \mathcal{S})$ can viewed as points in a geometric space with distances given by the total variation distance.

Proof. (i) Let us verify that $d_{\mathrm{tv}}(\mu_1, \mu_2) = 0$ if and only if $\mu_1 = \mu_2$. For the forward implication, it suffices to observe that $d_{\mathrm{tv}}(\mu_1, \mu_2) = 0$ implies that $|\mu_1(A) - \mu_2(A)| = 0$ for all $A \in \mathcal{S}$, and therefore $\mu_1 = \mu_2$. The backward implication is immediate.

(ii) The symmetry property $d_{\mathrm{tv}}(\mu_1, \mu_2) = d_{\mathrm{tv}}(\mu_2, \mu_1)$ is obvious.

(iii) Let us verify the triangle inequality. Let μ_1, μ_2, μ_3 be probability measures on $(S, \mathcal{S})$. The triangle inequality for the absolute value on the real line implies that

$$\begin{aligned} \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_3(A)| \;&\le\; \sup_{A \in \mathcal{S}} \Big(|\mu_1(A) - \mu_2(A)| + |\mu_2(A) - \mu_3(A)| \Big) \\ &\le\; \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_2(A)| + \sup_{A \in \mathcal{S}} |\mu_2(A) - \mu_3(A)|. \end{aligned}$$

Therefore, $d_{\mathrm{tv}}(\mu_1, \mu_3) \le d_{\mathrm{tv}}(\mu_1, \mu_2) + d_{\mathrm{tv}}(\mu_2, \mu_3)$. □

Example 9.2. The total variation distance between Dirac measures δ_x and δ_y on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ equals one if $x \neq y$, and zero otherwise (Exercise 9.11).

The following result provides a helpful symmetry property for densities of probability measures.

Lemma 9.3. *For any measurable functions $f_1, f_2 \colon S \to \mathbb{R}_+$ such that $\int_S f_1 \, d\nu = 1$ and $\int_S f_2 \, d\nu = 1$,*

$$\int_S (f_1 - f_2)_+ \, d\nu \;=\; \int_S (f_2 - f_1)_+ \, d\nu \;=\; \frac{1}{2} \int_S |f_1 - f_2| \, d\nu \tag{9.2}$$

and

$$\int_S (f_1 \wedge f_2)\, d\nu \;=\; 1 - \frac{1}{2}\int_S |f_1 - f_2|\, d\nu. \tag{9.3}$$

Proof. (i) Observe that $(f_1 - f_2)_+ = (f_1 - f_2)1_A$ and $(f_2 - f_1)_+ = (f_2 - f_1)1_{A^c}$, where $A = \{x\colon f_1(x) > f_2(x)\}$. By integrating these equalities, we find that

$$\int_S (f_1 - f_2)_+\, d\nu \;=\; \int_A (f_1 - f_2)\, d\nu, \tag{9.4}$$

$$\int_S (f_2 - f_1)_+\, d\nu \;=\; \int_{A^c} (f_2 - f_1)\, d\nu. \tag{9.5}$$

Because $\int_S f_1\, d\mu = \int_S f_2\, d\mu$, it follows that

$$0 \;=\; \int_S (f_2 - f_1)\, d\mu \;=\; \int_A (f_2 - f_1)\, d\mu + \int_{A^c} (f_2 - f_1)\, d\mu,$$

from which we see that the integrals on the right sides of (9.4)–(9.5) are equal to each other. This confirms the first equality in (9.2).

(ii) Observe that $|x - y| = (x - y)_+ + (y - x)_+$ where we recall that $t_+ = \max\{t, 0\}$ denotes the positive part of t. Then

$$\int_S |f_1 - f_2|\, d\nu \;=\; \int_S (f_1 - f_2)_+\, d\nu + \int_S (f_2 - f_1)_+\, d\nu. \tag{9.6}$$

In part (i) we saw that both integrals on right are equal to each other. This confirms the second equality in (9.2).

(iii) Finally, we note that

$$\begin{aligned}
\int_S (f_1 \wedge f_2)\, d\nu \;&=\; \int_A f_2\, d\nu + \int_{A^c} f_1\, d\nu \\
&=\; 1 - \int_A (f_1 - f_2)\, d\nu \\
&\overset{(9.4)}{=}\; 1 - \int_S (f_1 - f_2)_+\, d\nu \;\overset{(9.2)}{=}\; 1 - \frac{1}{2}\int_S |f_1 - f_2|\, d\nu.
\end{aligned}$$

□

Proposition 9.4. *Let μ_1, μ_2 be probability measures admitting densities $f_1, f_2\colon S \to \mathbb{R}_+$ with respect to a reference measure ν on $(S, \mathcal{S})$. Then*

$$d_{\mathrm{tv}}(\mu_1, \mu_2) \;=\; \frac{1}{2}\int_S |f_1(x) - f_2(x)|\, \nu(dx). \tag{9.7}$$

In particular, for any probability measures on a countable space S with *probability mass functions* $f_i(x) = \mu_i(\{x\})$*,*

$$d_{\mathrm{tv}}(\mu_1,\mu_2) \;=\; \frac{1}{2}\sum_{x\in S}|f_1(x)-f_2(x)|. \tag{9.8}$$

Proof. (i) By writing $(f_1-f_2)_+ = (f_1-f_2)1_A$ for $A = \{x\colon f_1(x) > f_2(x)\}$, we see that

$$\int_S (f_1-f_2)_+\,d\nu \;=\; \int_A f_1\,d\nu - \int_A f_2\,d\nu \;=\; \mu_1(A)-\mu_2(A) \;\le\; |\mu_1(A)-\mu_2(A)|.$$

Hence $\int_S (f_1-f_2)_+\,d\nu \le d_{\mathrm{tv}}(\mu_1,\mu_2)$. With the help of Lemma 9.3 we may then conclude that $\frac{1}{2}\int_S |f_1-f_2|\,d\nu \le d_{\mathrm{tv}}(\mu_1,\mu_2)$.

(ii) Observe that $(f_1-f_2)1_B \le (f_1-f_2)_+1_B \le (f_1-f_2)_+$ pointwise for an arbitrary measurable set B.

$$\mu_1(B)-\mu_2(B) \;=\; \int_B f_1\,d\nu - \int_B f_2\,d\nu \;=\; \int_S (f_1-f_2)1_B\,d\nu \;\le\; \int_S (f_1-f_2)_+\,d\nu.$$

Similarly, we find that

$$\mu_2(B)-\mu_1(B) \;\le\; \int_S (f_2-f_1)_+\,d\nu.$$

In light of Lemma 9.3, the rightmost integrals appearing in the above inequalities are both equal to $\frac{1}{2}\int_S |f_1-f_2|\,d\nu$. As a consequence,

$$|\mu_1(B)-\mu_2(B)| \;\le\; \frac{1}{2}\int_S |f_1-f_2|\,d\nu.$$

Because this holds for all $B \in \mathcal{S}$, we see that $d_{\mathrm{tv}}(\mu_1,\mu_2) \le \frac{1}{2}\int_S |f_1-f_2|\,d\nu$. □

Example 9.5. Determine the total variation distance between Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ with parameters p and q.

Recall that $\mathrm{Ber}(p)$ is a probability measure with density

$$f_p(x) \;=\; \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & \text{else}, \end{cases}$$

with respect to the counting measure # on $(\mathbb{Z}, 2^{\mathbb{Z}})$. By Proposition 9.4,

$$\begin{aligned} d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) &= \frac{1}{2} \int_{\mathbb{Z}} |f_p(x) - f_q(x)| \, \#(dx) \\ &= \frac{1}{2} \sum_{x \in \mathbb{Z}} |f_p(x) - f_q(x)| \\ &= \frac{1}{2} \Big(|(1-p) - (1-q)| + |p-q| \Big) \\ &= |p-q|. \end{aligned}$$

9.2 Couplings

A *coupling* of probability measures μ_1 on $(S_1, \mathcal{S}_1)$ and μ_2 on $(S_2, \mathcal{S}_2)$ is a pair (X_1, X_2) of random variables defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathrm{Law}(X_1) = \mu_1$ and $\mathrm{Law}(X_2) = \mu_2$.

The set of couplings corresponds to the set of possible dependence structures between a pair of probability measures. The trivial coupling corresponds to the degenerate dependence structure: independence.

Example 9.6 (Trivial coupling)**.** Let X_1 and X_2 be independent random variables distributed according to probability measures μ_1 and μ_2. Then (X_1, X_2) constitutes a coupling, called the *trivial coupling* of μ_1 and μ_2.

Proposition 9.7. *The total variation distance between probability measures μ_1 and μ_2 on a measurable space $(S, \mathcal{S})$ is bounded by*

$$d_{\mathrm{tv}}(\mu_1, \mu_2) \;\le\; \mathbb{P}(X_1 \ne X_2) \tag{9.9}$$

for all couplings (X_1, X_2) of μ_1 and μ_2. Furthermore, there exists a coupling for which the above bound holds as equality.

Proof. (i) Assume that (X_1, X_2) is a coupling of μ_1 and μ_2. Then for any measurable set $A \subset S$,

$$\begin{aligned} \mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A) &= \mathbb{E} 1_A(X_1) - \mathbb{E} 1_A(X_2) \\ &= \mathbb{E}\Big(1_A(X_1) - 1_A(X_2) \Big). \end{aligned}$$

We note that $1_A(X_1) - 1_A(X_2) = 0$ whenever $X_1 = X_2$. Therefore,

$$|1_A(X_1) - 1_A(X_2)| \;\le\; 1_D$$

where $D = \{\omega \in \Omega \colon X_1(\omega) \neq X_2(\omega)\}$. Because $\mathrm{Law}(X_1) = \mu_1$ and $\mathrm{Law}(X_2) = \mu_2$, it follows that

$$\begin{aligned} |\mu_1(A) - \mu_2(A)| &= \big|\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)\big| \\ &\le \mathbb{E}\big|1_A(X_1) - 1_A(X_2)\big| \\ &\le \mathbb{E}1_D \\ &= \mathbb{P}(X_1 \neq X_2). \end{aligned}$$

Because the above inequality holds for all $A \in \mathcal{S}$, we conclude that

$$d_{\mathrm{tv}}(\mu_1, \mu_2) = \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_2(A)| \le \mathbb{P}(X_1 \neq X_2).$$

Proving the existence of an optimal coupling is beyond the scope of this course. This is usually done by verifying that in the weak topology of probability measures on $\mathbb{R}^2$: (i) the function $\mathrm{Law}(X,Y) \mapsto \mathbb{P}(X \neq Y)$ is a lower semicontinuous, and (ii) the collection of probability measures $\{\mathrm{Law}(X,Y) \colon (X,Y) \in \Gamma(\mu,\nu)\}$ is compact. See [Vil09]. □

Example 9.8 (Coupling two coins)**.** A trivial coupling of Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ with parameters $p < q$ corresponds to pair independent random variables such that $\mathrm{Law}(X_1) = \mathrm{Ber}(p)$ and $\mathrm{Law}(X_2) = \mathrm{Ber}(q)$. The trivial coupling combined with (9.9) provides an upper bound

$$\begin{aligned} d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) &\le \mathbb{P}(X_1 \neq X_2) \\ &= \mathbb{P}(X_1 = 0, X_2 = 1) + \mathbb{P}(X_1 = 1, X_2 = 0) \\ &= (1-p)q + p(1-q). \end{aligned}$$

For $p = 0.5$ and $q = 0.6$ this yields $d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \le 0.5$.

Let us try to construct a better coupling. Define $X_1 = \theta_1$ and $X_2 = \theta_1 \vee \theta_2$ where θ_1, θ_2 are independent Bernoulli random variables with parameters r_1, r_2. We want (X_1, X_2) to be a coupling of $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$. For this property to be true, it is necessary that

$$\begin{aligned} p &= \mathbb{P}(X_1 = 1) = \mathbb{P}(\theta_1 = 1), \\ q &= \mathbb{P}(X_2 = 1) = \mathbb{P}(\theta_1 = 1) + \mathbb{P}(\theta_1 = 0)\mathbb{P}(\theta_2 = 1). \end{aligned}$$

This is equivalent to $r_1 = p$ and $r_1 + (1 - r_1) r_2 = q$. Then we solve $r_2 = \frac{q-p}{1-p}$. With these choices, it follows that (X_1, X_2) is a coupling of $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$. For this coupling,

$$\mathbb{P}(X_1 \neq X_2) = \mathbb{P}(\theta_1 = 0, \theta_2 = 1) = (1 - r_1) r_2 = (1-p)\frac{q-p}{1-p} = q - p.$$

In light of Example 9.5, we see that this is indeed an optimal coupling. For $p = 0.5$ and $q = 0.6$ this yields $d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \le 0.1$.

9.3 Convergence in total variation

Convergence in total variation for discrete probability spaces corresponds to pointwise convergence of probability mass functions. Somewhat surprisingly, pointwise convergence and L_1-convergence are equivalent in this setting.

Proposition 9.9. *Let S be countable. Then the following are equivalent for probability measures μ_n, μ on $(S, 2^S)$ with probability mass functions f_n, f:*

(i) $d_{\mathrm{tv}}(\mu_n, \mu) \to 0$.

(ii) $f_n(x) \to f(x)$ *for every* $x \in S$.

(iii) $\sum_{x \in S} |f_n(x) - f(x)| \to 0$.

Proof. (i) $\iff$ (iii) follows by Proposition 9.4.

(iii) $\implies$ (ii) is obvious.

(ii) $\implies$ (iii). Assume that $f_n(x) \to f(x)$ for every $x \in S$. Enumerate $S = \{x_1, x_2, \dots\}$. Fix $\epsilon > 0$. Because $\sum_{k=1}^\infty f(x_k) = 1$, we may fix an integer $K \geq 1$ such that $\sum_{k>K}^\infty f(x_k) \leq \epsilon$. Then

$$\begin{aligned}\sum_{k>K} f_n(x_k) &= \sum_{k>K} f(x_k) + \sum_{k>K} (f_n(x_k) - f(x_k)) \\ &= \sum_{k>K} f(x_k) + \sum_{k\leq K} (f(x_k) - f_n(x_k)) \\ &\leq \sum_{k>K} f(x_k) + \sum_{k\leq K} |f_n(x_k) - f(x_k)|.\end{aligned}$$

Hence

$$\begin{aligned}\sum_{x\in S} |f_n(x) - f(x)| &= \sum_{k\leq K} |f_n(x_k) - f(x_k)| + \sum_{k>K} |f_n(x_k) - f(x_k)| \\ &\leq \sum_{k\leq K} |f_n(x_k) - f(x_k)| + \sum_{k>K} (f_n(x_k) + f(x_k)) \\ &\leq 2\sum_{k\leq K} |f_n(x_k) - f(x_k)| + 2\sum_{k>K} f(x_k) \\ &\leq 2K \max_{k\leq K} |f_n(x_k) - f(x_k)| + 2\epsilon.\end{aligned}$$

By taking limits as $n \to \infty$, we find that

$$\limsup_{n\to\infty} \sum_{x\in S} |f_n(x) - f(x)| \leq 2\epsilon.$$

Because the above inequality is true for all $\epsilon > 0$, we conclude that (iii) holds. □

9.4 Poisson approximation

The binomial distribution $\mathrm{Bin}(n,p)$ represents the law of a sum

$$S_n \;=\; X_1 + \cdots + X_n$$

of independent $\mathrm{Ber}(p)$-distributed random variables $X_1, \dots, X_n$. A classical result, known as Poisson's[1] *law of small numbers*, tells us that the sum S_n is approximately Poisson distributed when n is large and the mean $\mathbb{E}S_n = np$ is bounded.

Proposition 9.10. *When $p_n = \alpha/n$ for some constant $0 < \alpha < \infty$, then $\mathrm{Bin}(n, p_n) \to \mathrm{Poi}(\lambda)$ in total variation as $n \to \infty$.*

Proof. Fix an integer $n \geq 1$. We construct a coupling of $\mathrm{Bin}(n, p_n)$ and $\mathrm{Poi}(\lambda)$ as follows. Let $(X, \tilde{X})$ be an optimal coupling of $\mathrm{Ber}(p_n)$ and $\mathrm{Poi}(p_n)$, so that $\mathbb{P}(X \neq \tilde{X}) = d_{\mathrm{tv}}(\mathrm{Ber}(p_n), \mathrm{Poi}(p_n))$. Define

$$\begin{aligned} S_n &= X_1 + \cdots + X_n, \\ \tilde{S}_n &= \tilde{X}_1 + \cdots + \tilde{X}_n, \end{aligned}$$

where $(X_1, \tilde{X}_1), \dots, (X_n, \tilde{X}_n)$ are independent copies of $(X, \tilde{X})$. Then we see that $\mathrm{Law}(S_n) = \mathrm{Bin}(n, p_n)$ and[2] $\mathrm{Law}(\tilde{S}_n) = \mathrm{Poi}(np_n)$. Hence $(S_n, \tilde{S}_n)$ constitutes a coupling of $\mathrm{Bin}(n, p_n)$ and $\mathrm{Poi}(np_n)$. The construction of the coupling shows that $S_n \neq \tilde{S}_n$ is possible only when $X_i \neq \tilde{X}_i$ for some $i = 1, \dots, n$. Hence the union bound implies that

$$\mathbb{P}(S_n \neq \tilde{S}_n) \;\leq\; \sum_{i=1}^n \mathbb{P}(X_i \neq \tilde{X}_i) \;=\; n\mathbb{P}(X \neq \tilde{X}).$$

We conclude by Proposition 9.7 that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n, p_n), \mathrm{Poi}(np_n)) \;\leq\; n\, d_{\mathrm{tv}}(\mathrm{Ber}(p_n), \mathrm{Poi}(p_n)). \tag{9.10}$$

Next, with the help of Proposition 9.4 we note that (Exercise 9.13)

$$d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Poi}(p)) \;=\; p(1 - e^{-p}) \qquad \text{for all } 0 \leq p \leq 1. \tag{9.11}$$

[1] Siméon Poisson, 1781 – 1840, PhD École Polytechnique 1800 for Lagrange and Laplace.

[2] We know that the sum of independent Poisson random variables is Poisson.

By plugging this into (9.10) and applying the bound $1-t \le e^{-t}$, we conclude that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n,p_n),\mathrm{Poi}(np_n)) \;\le\; np_n^2.$$

Recalling that $p_n = \alpha/n$, we see that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n,p_n),\mathrm{Poi}(\alpha)) \;\le\; \alpha^2/n \;\to\; 0 \qquad \text{as } n \to \infty.$$

□

9.5 Exercises

Example 9.11. Prove the statement of Example 9.2 directly from the definition.

Exercise 9.12. Compute the total variation distance for the following instances of probability measures on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$:

(a) $d_{\mathrm{tv}}(\delta_a,\delta_b)$.

(b) $d_{\mathrm{tv}}(\mathrm{Ber}(p),\mathrm{Nor}(0,1))$.

(c) $d_{\mathrm{tv}}(\mathrm{Nor}(a,1),\mathrm{Nor}(b,1))$.

Exercise 9.13. Prove equality (9.11).

Chapter 10

Conditional probability

> *If there be two subsequent events, the probability of the second $\frac{b}{N}$ and the probability of both together $\frac{P}{N}$, and it being first discovered that the second event has happened, from hence I guess that the first event has also happened, the probability I am in the right is $\frac{P}{b}$.*
>
> —*Thomas Bayes 1763*

Conditional probabilities are intuitively easy but mathematically nontrivial to define with respect to events of infinitesimally (read: zero) probability. There are two approaches for doing this rigorously: (i) disintegration and probability kernels, and (ii) conditional expectations against a sigma-algebra. In this chapter we get introduced to these, with main focus being on the more concrete approach using probability kernels.

Key concepts: probability kernel, conditional distribution, Bayes formula

Learning outcomes:

- Learn to disintegrate joint probabilities using probability kernels
- Learn to determine posterior distributions in Bayesian context
- Learn to work with Markov chains in continuous state spaces

Prerequisites: Previous chapters.

10.1 Probability kernels

A *kernel* from a measurable space[1] S_1 into a measurable space S_2 is a function $K\colon S_1 \times \mathcal{S}_2 \to \bar{\mathbb{R}}_+$ such that for every point $x \in S_1$,

$$B \mapsto K(x,B) \text{ is a measure on } S_2,$$

and for every measurable set $B \subset S_2$,

$$x \mapsto K(x,B) \text{ is a measurable function.}$$

A *probability kernel* is a kernel such that $B \mapsto K(x,B)$ is a probability measure for every x. A kernel *on* a measurable space is a kernel from the space into itself.

> *A probability kernel corresponds to a collection of probability measures $K_x\colon B \mapsto K(x,B)$ on S_2 indexed by the points of S_1 in a measurable manner.*

Example 10.1 (Poisson kernel)**.** Define a map $K\colon \mathbb{R}_+ \times 2^{\mathbb{Z}_+} \to [0,1]$ by

$$K(x,B) \ = \ \begin{cases} \sum_{k \in B} e^{-x} \frac{x^k}{k!}, & x > 0, \\ \delta_0(B), & x = 0. \end{cases}$$

The the set function $B \mapsto K(x,B)$ can be recognised as the probability measure on $\mathbb{Z}_+$ corresponding to the Poisson distribution with mean x, with the natural extension that the Poisson distribution with mean zero equals the Dirac measure at zero. It is possible to verify that $x \mapsto K(x,B)$ is a continuous and thus measurable function from $\mathbb{R}_+$ into $[0,1]$ for every $B \subset \mathbb{Z}_+$. Hence K is a probability kernel from $\mathbb{R}_+$ into $\mathbb{Z}_+$.

Example 10.2 (Deterministic map)**.** Given an arbitrary measurable function $T\colon S_1 \to S_2$, define

$$K(x,B) \ = \ \delta_{T(x)}(B).$$

Then $K(x,B)$ as a function of B is just the Dirac measure at $y = T(x)$, so we see that $B \mapsto K(x,B)$ is a probability measure. After verifying that $x \mapsto K(x,B)$ is a measurable function (Exercise 10.12), we conclude that K is a kernel from S_1 into S_2.

[1] 'Measurable space S' means a pair $(S, \mathcal{S})$ where $\mathcal{S}$ is a sigma-algebra on S.

Every measurable function $T\colon S_1 \to S_2$ corresponds to a probability kernel from S_1 into S_2.

Example 10.3 (Random map)**.** Given a measurable function $\phi\colon S_1 \times [0,1] \to S_2$ and a uniformly in $[0,1]$ distributed random variable $U\colon \Omega \to [0,1]$, let us define

$$K(x,B) \ = \ \mathbb{P}\big(\phi(x,U) \in B\big).$$

For each x, the set function $B \mapsto K(x,B)$ is just the law of the random variable $Y = \phi(x,U)$, and hence a probability measure on S_2. Let us now verify that $K(x,B)$ is measurable as a function of x. By noting that the law of U is the Lebesgue measure on $[0,1]$, we see that

$$K(x,B) \ = \ \mathbb{E} 1_B(\phi(x,U)) \ = \ \int_{[0,1]} 1_B(\phi(x,u))\, \lambda(du).$$

Because $(x,u) \mapsto 1_B(\phi(x,u))$ is a measurable function from $S_1 \times [0,1]$ into $\bar{\mathbb{R}}_+$, Proposition 6.5 implies that $x \mapsto \int_{[0,1]} 1_B(\phi(x,u))\, \lambda(du)$ is a measurable function from S_1 into $\bar{\mathbb{R}}_+$. Therefore, $x \mapsto K(x,B)$ is measurable. We conclude that K is a probability kernel from S_1 into S_2.

The collection of random variables $\phi(x,U)$, $x \in S_1$, in Example 10.3 can be considered as mechanistic representation of a noisy function from S_1 into S_2. Indeed, *every* probability kernel can be represented as such as random map.

Theorem 10.4. *Every probability kernel K from S into $\mathbb{R}$ can be represented in terms of a random map according to*

$$K(x,B) \ = \ \mathbb{P}\big(\phi(x,U) \in B\big) \tag{10.1}$$

for some measurable function $\phi\colon S \times (0,1) \to \mathbb{R}$ and some uniformly in $(0,1)$ distributed random variable U.

Proof. For each $x \in S$, define a function $Q_x\colon (0,1) \to \mathbb{R}$ by

$$Q_x(u) = \sup\{t \in \mathbb{R} : K(x,(-\infty,t]) < u\}. \tag{10.2}$$

Then Q_x is the left-continuous quantile function of the probability measure $K_x\colon B \mapsto K(x,B)$, recall (8.3). Then $\operatorname{Law}(Q_x(U)) = K_x$ by Proposition 8.4. That is,

$$\mathbb{P}(Q_x(U) \in B) \ = \ K_x(B) \qquad \text{for all measurable } B \subset \mathbb{R}.$$

Hence (10.1) is valid for the function $\phi(x,u) = Q_x(u)$.

fixme We still need to verify that $x \mapsto Q_x(u)$ is measurable. Note that $x \mapsto K(x,(-\infty,t])$ is measurable for every t, and that the supremum in (10.2) can be restricted to the countable set $\mathbb{Q}$.

The set $\{(x,u)\colon K(x,(-\infty,t]) < u\}$ is measurable for every t. why?

□

10.2 Kernel products

The *product* of a probability measure μ on S_1 and a probability kernel K from S_1 into S_2 is a set function $\mu \otimes K \colon \mathcal{S}_1 \otimes \mathcal{S}_2 \to [0,1]$ defined by

$$(\mu \otimes K)(C) \;=\; \int_{S_1} \Big(\int_{S_2} 1_C(x,y)\, K(x,dy) \Big) \mu(dx). \tag{10.3}$$

This is quite similar to the definition of product measure in (6.4).

The product (10.3) is often abbreviated as $\mu(dx)K(x,dy)$.

Proposition 10.5. *$\mu \otimes K$ is a probability measure on $S_1 \times S_2$.*

The probability measure $\mu \otimes K$ corresponds to a pair of random variables (X_1, X_2) sampled in two stages as follows:

(i) Sample X_1 from probability distribution μ

(ii) Sample X_2 from probability distribution $K(X_1, \cdot)$.

Proof. To be sure that the right side of (10.3) is well defined, we need to verify that $x \mapsto \int_{S_2} 1_C(x,y)\, K(x,dy)$ is a measurable function. This needs some work, in a similar manner as in the Fubini theorem's proof. We omit this (see [Çın11, Proposition 6.9]).

(i) $(\mu \otimes K)(\emptyset) = 0$ because $1_\emptyset(x,y) = 0$ identically.

(ii) Let us verify that $\mu \otimes K$ is countably disjointly additive. Let $C_1, C_2, \dots$ be disjoint measurable sets in $S_1 \times S_2$. Then

$$1_{\cup_{i=1}^\infty C_i} \;=\; \sum_{i=1}^\infty 1_{C_i} \;=\; \lim_{n\to\infty} \sum_{i=1}^n 1_{C_i}$$

is a monotone increasing limits of nonnegative measurable functions 1_{C_i}. Therefore, by monotone continuity and linearity of nonnegative integration, we find that

$$\begin{aligned}(\mu \otimes K)\Big(\bigcup_{i=1}^{\infty} C_i\Big) &= \int_{S_1} \Big(\int_{S_2} 1_{\cup_{i=1}^{\infty} C_i}(x,y)\, K(x,dy) \Big) \mu(dx) \\ &= \int_{S_1} \Big(\int_{S_2} \lim_{n\to\infty} \sum_{i=1}^{n} 1_{C_i}(x,y)\, K(x,dy) \Big) \mu(dx) \\ &= \lim_{n\to\infty} \int_{S_1} \Big(\int_{S_2} \sum_{i=1}^{n} 1_{C_i}(x,y)\, K(x,dy) \Big) \mu(dx) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \int_{S_1} \Big(\int_{S_2} 1_{C_i}(x,y)\, K(x,dy) \Big) \mu(dx).\end{aligned}$$

By noting that the double integral on the right equals $(\mu \otimes K)(C_i)$, and that $\lim_{n\to\infty} \sum_{i=1}^{n} = \sum_{i=1}^{\infty}$ for nonnegative summands, we conclude that $\mu \otimes C$ is countably disjointly additive.

(iii) Because $1_{S_1 \times S_2}(x,y) = 1$ identically we find that

$$(\mu \otimes K)(S_1 \times S_2) \;=\; \int_{S_1} \Big(\int_{S_2} K(x,dy) \Big) \mu(dx).$$

By noting that $\int_{S_2} K(x,dy) = K(x,S_2) = 1$ for all x, and that $\int_{S_1} \mu(dx) = \mu(S_1) = 1$, we conclude that $(\mu \otimes K)(S_1 \times S_2) = 1$. □

The *pushforward* of a probability measure μ on S_1 by a probability kernel K from S_1 into S_2 is a set function $\mu K \colon \mathcal{S}_2 \to [0,1]$ defined by

$$(\mu K)(B) \;=\; \int_{S_1} K(x,B)\, \mu(dx). \tag{10.4}$$

Proposition 10.6. *The pushforward μK defined by (10.4) is a probability measure on S_2.*

The pushforward operation $\mu \mapsto \mu K$ maps probability measures into probability measures.

Proof. The proof is similar but easier than the proof of Proposition 10.5, and left to the reader (Exercise 10.13). □

Proposition 10.7. *The marginal distributions of the probability measure $\gamma = \mu \otimes K$ are given $\gamma \circ \pi_1^{-1} = \mu$ and $\gamma \circ \pi_2^{-1} = \mu K$.*

For a random vector with distribution $\mathrm{Law}(X_1, X_2) = \mu \otimes K$*, the coordinates are distributed according to* $\mathrm{Law}(X_1) = \mu$ *and* $\mathrm{Law}(X_2) = \mu K$*.*

Proof. Let us verify that

$$\gamma(A \times S_2) = \mu(A) \qquad \text{and} \qquad \gamma(S_1 \times B) = (\mu K)(B) \tag{10.5}$$

for any measurable sets $A \subset S_1$ and $B \subset S_2$. This will imply the claim because preimages of coordinate projections are given by $\pi_1^{-1}(A) = A \times S_2$ and $\pi_2^{-1}(B) = S_1 \times B$.

(i) By applying the equality $1_{A \times S_2}(x,y) = 1_A(x)$ and then the equality $\int_{S_2} K(x,dy) = K(x,S_2) = 1$, we find that

$$\begin{aligned}
(\mu \otimes K)(A \times S_2) &\overset{(10.3)}{=} \int_{S_1} \Big(\int_{S_2} 1_{A \times S_2}(x,y)\, K(x,dy) \Big) \mu(dx) \\
&= \int_{S_1} \Big(\int_{S_2} 1_A(x)\, K(x,dy) \Big) \mu(dx) \\
&= \int_{S_1} 1_A(x) \Big(\int_{S_2} K(x,dy) \Big) \mu(dx) \\
&= \int_{S_1} 1_A(x)\, \mu(dx) \\
&= \mu(A).
\end{aligned}$$

This confirms the first equality in (10.5).

The second equality in (10.5) follows by combining the equality $1_{S_1 \times B}(x,y) = 1_B(y)$ and the definition of μK to conclude that

$$\begin{aligned}
(\mu \otimes K)(S_1 \times B) &\overset{(10.3)}{=} \int_{S_1} \Big(\int_{S_2} 1_{S_1 \times B}(x,y)\, K(x,dy) \Big) \mu(dx) \\
&= \int_{S_1} \Big(\int_{S_2} 1_B(y)\, K(x,dy) \Big) \mu(dx) \\
&= \int_{S_1} K(x,B)\, \mu(dx) \\
&\overset{(10.4)}{=} (\mu K)(B).
\end{aligned}$$

□

10.3 Kernel densities

A measurable function $k\colon S_1 \times S_2 \to \mathbb{R}_+$ is called a *density function* of a probability kernel K with respect to a measure ν on S_2, if

$$K(x,B) \;=\; \int_B k(x,y)\,\nu(dy)$$

for all $x \in S_1$ and all measurable sets $B \subset S_2$.

In this we often write $K(x,dy) = k(x,y)\nu(dy)$.

Proposition 10.8. *If a probability measure μ_1 admits a density f_1 with respect to a measure ν_1, and a probability kernel K admits a density k with respect to a measure ν_2, then $\mu_1 \otimes K$ admits a density*

$$f(x,y) \;=\; f_1(x)k(x,y)$$

with respect to $\nu_1 \otimes \nu_2$.

In shorthand notation, the product of a probability measure $f_1(x)\nu_1(dx)$ and a probability kernel $k(x,y)\nu_2(dy)$ is given by $f_1(x)k(x,y)\nu_1(dx)\nu_2(dy)$.

Proof.

$$\begin{aligned}
&\int_A \biggl(\int_B f(x,y)\nu_2(dy) \biggr) \nu_1(dx) \\
&= \int_A \biggl(\int_B f_1(x)k(x,y)\nu_2(dy) \biggr) \nu_1(dx) \\
&= \int_A f_1(x) \biggl(\int_B k(x,y)\nu_2(dy) \biggr) \nu_1(dx) \\
&= \int_A f_1(x)K(x,B)\nu_1(dx) \\
&= \int_A K(x,B)\mu_1(dx).
\end{aligned}$$

Hence $\int_{A\times B} f d(\nu_1 \otimes \nu_2) = (\mu_1 \otimes K)(A \times B)$. Hence the probability measures $fd(\nu_1 \otimes \nu_2)$ and $\mu_1 \otimes K$ agree on sets $A \times B$. By Dynkin's identification theorem, they agree for all measurable sets. □

10.4 Markov chains

On a finite or countably infinite state space, a Markov chain is defined as a random sequence such that

$$\mathbb{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1}, \dots, X_0 = x_0) \;=\; \mathbb{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1})$$

for all $x_0, \dots, x_{t-1}$ such that the conditioning events occur with nonzero probabilities. Extending this definition to general uncountable state spaces is complicated because typically the probability of a Markov chain hitting any particular singleton set is zero. A rigorous definition can be formulated using probability kernels. We first need to extend the definition of the kernel product (10.3) into higher dimensions as follows. The *product* of a probability measure γ on $S_1 \times \cdots \times S_t$ and a probability kernel P from S_t into S_{t+1} is a probability measure on $S_1 \times \cdots \times S_{t+1}$ defined by

$$(\gamma \otimes P)(dx_1, \dots, dx_{t+1}) \;=\; \gamma(dx_1, \dots, dx_t) P(x_t, dx_{t+1}).$$

A *Markov chain* on a general measurable space S is defined as a random sequence $X_0, X_1, \dots$ such that for any integer $t \ge 1$ there exists a probability kernel P_t on S such that

$$\mathrm{Law}(X_0, \dots, X_t) \;=\; \mathrm{Law}(X_0, \dots, X_{t-1}) \otimes P_t. \tag{10.6}$$

The definition implies (Exercise 10.14) that

$$\mathbb{P}(X_t \in B \,|\, X_{t-1} = x) \;=\; P_t(x, B)$$

for all measurable sets B and all states x such that the conditioning event on the left occurs with a nonzero probability.

The probability kernel P_t represents the conditional distribution of a Markov chain at time t given its state at time $t-1$, and gives a rigorous meaning to the right side of

$$P_t(x_{t-1}, B) = \text{‘}\mathbb{P}(X_t \in B \,|\, X_{t-1} = x_{t-1})\text{’}$$

By iterating formula (10.6), we find that

$$\mathrm{Law}(X_0, \dots, X_t) \;=\; \mu_0 \otimes P_1 \otimes \cdots \otimes P_t,$$

where $\mu_0 = \mathrm{Law}(X_0)$ is the *initial distribution* of the Markov chain.

The finite-dimensional distributions $\mathrm{Law}(X_0, \dots, X_t)$, $t \geq 0$, *of a Markov chain are completely determined by the initial distribution* μ_0 *and transition probability kernels* $P_1, P_2, \dots$

We also find (Exercise 10.15) that the pushforward by P_t satisfies

$$\mu_t = \mu_{t-1} P_t, \tag{10.7}$$

so that the probability kernel P_t maps $\mu_{t-1} = \mathrm{Law}(X_{t-1})$ into $\mu_t = \mathrm{Law}(X_t)$. By iterating formula (10.7), we find that

$$\mathrm{Law}(X_t) \;=\; \mu_0 P_1 P_2 \cdots P_t.$$

A Markov chain is called *time-homogeneous* if $P_t = P$ for all t. In this case a random sequence satisfying (10.6) is called a Markov chain with *transition probability kernel* P. The joint laws of a time-homogeneous Markov chains can be computed by

$$\mathrm{Law}(X_0, \dots, X_t) \;=\; \mu_0 \otimes \underbrace{P \otimes \cdots \otimes P}_{t}.$$

and the law of the chain at time t by

$$\mathrm{Law}(X_t) \;=\; \mu_0 P^t.$$

A probability measure π is an *equilibrium distribution* for the probability kernel P if

$$\pi P \;=\; \pi.$$

Example 10.9 (Markov chain on a finite state space)**.** For a Markov chain on a finite state space $S = \{1, \dots, n\}$, with initial probability (row) vector $\mu \in \mathbb{R}^{1 \times n}$ and transition probability matrix $P \in \mathbb{R}^{n \times n}$, the initial distribution is the probability measure

$$\mu_0(B) \;=\; \sum_{i \in B} \mu_i, \qquad B \subset \{1, \dots, n\}$$

and the transition probability kernel is given by

$$P(i, B) \;=\; \sum_{j \in B} P_{i,j}. \qquad B \subset \{1, \dots, n\}.$$

Then $\mathrm{Law}(X_0, \dots, X_t)$ has probability mass function

$$(x_0, \dots, x_t) \;\mapsto\; \mu_0(x_0) P_{x_0, x_1} \cdots P_{x_{t-1}, x_t}.$$

10.5 Disintegration

Theorem 10.10 (Disintegration)**.** *For any probability measure γ on $S \times \mathbb{R}$ with with marginals μ and ν, there exists a probability kernel K from S into $\mathbb{R}$ such that $\gamma = \mu \otimes K$.*

For any random variables $X\colon \Omega \to S$ and $Y\colon \Omega \to \mathbb{R}$, there exists a probability kernel K such that $\operatorname{Law}(X,Y) = \operatorname{Law}(X) \otimes K$.

A probability kernel K is called a *regular conditional probability distribution* of Y given X, if $P_{X,Y} = P_X \otimes K$.

Proposition 10.11. *Assume that the law of (X,Y) admits a density function $f(x,y)$ with respect to $\nu_1 \otimes \nu_2$. Then:*

(i) $\operatorname{Law}(X)$ *admits a density function $f_1(x) = \int f(x,y)\nu_2(dy)$.*

(ii) $\operatorname{Law}(Y)$ *admits a density function $f_2(x) = \int f(x,y)\nu_1(dx)$.*

(iii) The probability kernel $K(dx,dy) = k(x,y)\nu_2(dy)$ with density function

$$k(x,y) \;=\; \begin{cases} \frac{f(x,y)}{f_1(x)}, & f_1(x) > 0, \\ f_2(x), & f_1(x) = 0, \end{cases}$$

with respect to ν_2 is a regular conditional probability distribution of Y given X.

Proof. Todo. □

10.6 Exercises

Exercise 10.12 (Deterministic map)**.** Prove that $K(x,B) = \delta_{T(x)}(B)$ in Exercise 10.2 is a probability kernel from S_1 to S_2.

Exercise 10.13 (Pushforward by a probability kernel)**.** Prove Proposition 10.6 by adapting and simplifying the proof of Proposition 10.5.

Exercise 10.14 (Transition probability kernel)**.** Let $X_0, X_1, \dots$ be a Markov chain with transition probability kernels $P_1, P_2, \dots$ as defined in (10.6).

(a) Prove that

$$\mathbb{P}(X_{t-1} \in A,\, X_t \in B) \;=\; \int_A P_t(x,B)\,\mu_{t-1}(dx),$$

where $\mu_{t-1} = \operatorname{Law}(X_{t-1})$.

(b) Prove that

$$\mathbb{P}(X_t \in B \,|\, X_{t-1} = x) \;=\; P_t(x, B)$$

for all $x \in S$ such that $\mathbb{P}(X_{t-1} = x) > 0$.

Exercise 10.15 (Markov pushforward)**.** Prove (10.7).

Exercise 10.16 (Forward and backward Bayes kernels)**.** Consider a random vector in $(0,1) \times \{0,\dots,n\}$ distributed according to $\operatorname{Law}(Z, X) = \pi_Z \otimes K_{X|Z}$, in which π_Z equals the uniform distribution on $(0,1)$ and $K_{X|Z}$ is a probability kernel from $(0,1)$ into $\{0,\dots,n\}$ defined by

$$K_{X|Z}(z, B) \;=\; \sum_{x \in B} f(x \,|\, z)$$

where $f(x \,|\, z) = \binom{n}{x}(1-z)^{n-x} z^x$.

(a) Determine the probability mass function of $\pi_X = \pi_Z K_{X|Z}$.
Hint: The formula $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ for the beta function $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ may be helpful.

(b) Determine a probability kernel $K_{Z|X}$ from $\{0,\dots,n\}$ into $(0,1)$ such that

$$\pi_Z \otimes K_{X|Z} \;=\; K_{Z|X} \otimes \pi_X,$$

and give a natural definition to the expression on the right.

(c) Determine a function $(z,x) \mapsto g(z \,|\, x)$ such that

$$K_{Z|X}(x, A) \;=\; \int_A g(z \,|\, x)\, dz.$$

(d) In a Bayesian model with data X, and parameter Z with prior distribution π_Z, can you identify what corresponds to 'likelihood function' and what to 'posterior distribution' in the above notation?

Exercise 10.17 (Kernel operations)**.** A probability kernel from a measurable space $(S_1, \mathcal{S}_1)$ into a measurable space $(S_2, \mathcal{S}_2)$ is a function $K\colon S_1 \times \mathcal{S}_2 \to \mathbb{R}_+$ such that

(i) $x \mapsto K(x, B)$ is $\mathcal{S}_1/\mathcal{B}(\mathbb{R}_+)$-measurable for all $B \in \mathcal{S}_2$

(ii) $B \mapsto K(x, B)$ is a probability measure on $(S_2, \mathcal{S}_2)$ for every $x \in S_1$.

Given a probability measure μ on $(S_1, \mathcal{S}_1)$, define set functions μK on $\mathcal{S}_2$ and $\mu \otimes K$ on $\mathcal{S}_1 \otimes \mathcal{S}_2$ by

$$\begin{aligned} (\mu K)(B) &= \int_{S_1} K(x,B)\,\mu(dx), \qquad B \in \mathcal{S}_2, \\ (\mu \otimes K)(C) &= \int_{S_1} \Big(\int_{S_2} 1_C(x,y)\,K(x,dy) \Big)\,\mu(dx), \qquad C \in \mathcal{S}_1 \otimes \mathcal{S}_2. \end{aligned}$$

(a) Verify that $\mu \otimes K$ is a probability measure on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$.

(b) Verify that $\mu \otimes K$ has marginal distributions μ and μK.

(c) Verify that μK is a probability measure on $(S_2, \mathcal{S}_2)$.

Exercise 10.18 (Deterministic map as a kernel)**.** Let $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ be measurable spaces and consider a measurable function $T \colon S_1 \to S_2$. Define $K(x,B) = (1_B \circ T)(x)$ for $x \in S_1$ and $B \in \mathcal{S}_2$.

(a) Prove that K is a probability kernel from $(S_1, \mathcal{S}_1)$ to $(S_2, \mathcal{S}_2)$.

(b) How is K related to the joint law of $(X, T(X)) \in S_1 \times S_2$ where $X \colon \Omega \to S_1$ is a random variable defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$?

(c) Is it necessary in (a) to assume that T is a measurable function?

Exercise 10.19 (Gaussian kernel)**.** Define a probability kernel on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by $K(x,B) = \int_B k(x,y)\,dy$ where

$$k(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2}.$$

Let μ_0 be the standard normal distribution with mean zero and variance one.

(a) Determine the probability measure $\mu_1 = \mu_0 K$. Is μ_1 a normal distribution? If yes, write down its mean and variance. If not, explain why.

(b) Determine the probability measure μ_n obtained recursively from μ_0 by $\mu_1 = \mu_0 K$, $\mu_2 = \mu_1 K$, and so on. Is μ_n a normal distribution? If yes, write down its mean and variance. If not, explain why.

(c) Determine a probability kernel K_n such that $\mu_n = \mu_0 K_n$. Can you express it as $K_n(x,B) = \int_B k_n(x,y)\,dy$ for some continuous function $k_n \colon \mathbb{R}^2 \to \mathbb{R}$?

Exercise 10.20 (Associativity)**.** The map $\mu \mapsto \mu K$ in Exercise 10.17 can be viewed as left multiplication of a probability kernel by a probability measure. We may also define right multiplication of the kernel K by a bounded $\mathcal{S}_2/\mathcal{B}(\mathbb{R})$-measurable function f as $(Kf)(x) = \int_{S_2} f(y)K(x,dy)$. Prove that the left and right multiplications are associative in the sense that

$$(\mu K)f \;=\; \mu(Kf).$$

Exercise 10.21 (Trivariate law)**.** For μ_0 and K as in Exercise 10.19, define a probability measure on $(\mathbb{R}^3, \mathcal{B}(\mathbb{R}^3))$ by

$$(\mu_0 \otimes K \otimes K)(C) \;=\; \int_{\mathbb{R}} \biggl(\Bigl(\int_{\mathbb{R}} 1_C(x,y,z)\, K(y,dz) \Bigr) \, K(x,dy) \biggr) \, \mu_0(dx).$$

Determine the three marginal distributions of $\mu_0 \otimes K \otimes K$.

Chapter 11

Stochastic limits

11.1 Almost sure convergence

A real-valued random sequence $X_1, X_2, \dots$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ *converges almost surely*[1] to a real-valued random variable X, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) \;=\; 1.$$

This is a strong form of convergence: The sequence $X_n(\omega)$ convergences to $X(\omega)$ for all outcomes ω of the randomness-generating mechanism, excluding an event of zero probability.

Obviously, any sequence of random variables converging pointwise (for every ω), also converges almost surely. A limit of an almost surely converging sequence is not unique. If $\tilde{X} : \Omega \to \mathbb{R}$ is any other random variable such that $\tilde{X} = X$ almost surely, then $X_n \xrightarrow{\text{a.s.}} \tilde{X}$ as well (Exercise 11.5). Allowing the convergence not to hold on an event of zero probability is needed because many important limit theorems (in particular, the strong law of large numbers) of probability theory are true for almost sure convergence, but not for pointwise convergence.

11.2 Convergence in probability

A real-valued random sequence $X_1, X_2, \dots$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ *converges in probability*[2] to a real-valued random variable X, de-

[1] In Finnish: ‘suppenee melkein varmasti’

[2] In Finnish: ‘suppenee stokastisesti’

noted $X_n \xrightarrow{\mathbb{P}} X$, if

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \epsilon\Big) \;=\; 0 \qquad \text{for all } \epsilon > 0.$$

It can be proved that $X_n \xrightarrow{\text{a.s.}} X$ implies $X_n \xrightarrow{\mathbb{P}} X$ (see [Çın11, Theorem 3.3]).

Proposition 11.1. $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X$.

Proof. Fix a number $\epsilon > 0$ and define a random variable $Y_n = 1_{A_n \cap \Omega_0}$, where

$$\begin{aligned} A_n &= \big\{\omega\colon |X_n(\omega) - X(\omega)| > \epsilon\big\}, \\ \Omega_0 &= \big\{\omega\colon \lim_{n\to\infty} X_n(\omega) \to X(\omega)\big\}. \end{aligned}$$

Because $X_n(\omega) \to X(\omega)$ for every $\omega \in \Omega_0$, we find that $Y_n \to 0$ pointwise. Because $|Y_n| \le 1$ pointwise, it follows by the bounded continuity of expectation (Theorem 4.15) that

$$0 \;=\; \mathbb{E}(\lim_{n\to\infty} Y_n) \;=\; \lim_{n\to\infty} \mathbb{E}Y_n \;=\; \lim_{n\to\infty} \mathbb{P}(A_n \cap \Omega_0).$$

Because $\mathbb{P}(\Omega_0) = 1$, the insensitivity of integration (Theorem 5.3) implies that

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) \;=\; \lim_{n\to\infty} \mathbb{P}(A_n) \;=\; \lim_{n\to\infty} \mathbb{P}(A_n \cap \Omega_0) \;=\; 0.$$

Hence $X_n \xrightarrow{\mathbb{P}} X$. □

11.3 Borel–Cantelli theorem

The Borel[3] –Cantelli[4] theorem is a powerful tool for analysing whether or not certain events may occur infinitely often. An *event* is a measurable set $A \subset \Omega$ in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Events A_j, $j \in J$, are called *stochastically independent* if the corresponding indicator random variables 1_{A_j}, $j \in J$, are stochastically independent.

If $\mathbb{P}(A_n) \to 0$ sufficiently fast so that the sum $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ is finite, then with probability one, only finite many events among $A_1, A_2, \ldots$ may occur.

[3] Same Émile Borel who invented the modern set-theoretic approach to integration and supervised Lebesgue's doctoral thesis.

[4] Francesco Paolo Cantelli, 1875–1966, PhD 1899 @ University of Palermo.

If $\mathbb{P}(A_n)$ does not converge to zero, or if $\mathbb{P}(A_n) \to 0$ so slowly that the sum $\sum_{n=1}^\infty \mathbb{P}(A_n)$ is infinite, then with probability one, infinitely many events among independent events $A_1, A_2, \dots$ will occur.

Theorem 11.2 (Borel–Cantelli)**.** *For any events $A_1, A_2, \dots,$ the random variable $N = \sum_{j=1}^\infty 1_{A_j}$ satisfies*

$$\sum_{j=1}^\infty \mathbb{P}(A_j) < \infty \quad \implies \quad N < \infty \text{ almost surely.} \tag{11.1}$$

Furthermore, if the events $A_1, A_2, \dots$ are stochastically independent, then

$$\sum_{j=1}^\infty \mathbb{P}(A_j) = \infty \quad \implies \quad N = \infty \text{ almost surely.} \tag{11.2}$$

Proof. (i) Define $N_k = 1_{A_1} + \cdots + 1_{A_k}$. Then by the linearity of expectation, and noting that $\mathbb{E}1_{A_j} = \mathbb{P}(A_j)$, we see that

$$\mathbb{E}N_k \;=\; \sum_{j=1}^k \mathbb{P}(A_j).$$

Because $N_k \uparrow N$ pointwise, the monotone continuity of expectation implies that

$$\mathbb{E}N \;=\; \lim_{k\to\infty} \mathbb{E}N_k \;=\; \lim_{k\to\infty} \sum_{j=1}^k \mathbb{P}(A_j) \;=\; \sum_{j=1}^\infty \mathbb{P}(A_j).$$

If we assume that $\sum_{j=1}^\infty \mathbb{P}(A_j) < \infty$, then $\mathbb{E}N < \infty$, and hence (Theorem 5.2) implies that $N < \infty$ almost surely. This confirm the implication (11.1).

(ii) To verify (11.2), assume now that $\sum_{j=1}^\infty \mathbb{P}(A_j) = \infty$. Then the above computation shows that $\mathbb{E}N = \infty$, but this does not automatically imply that $N = \infty$ almost surely. We will impose the extra assumption that $A_1, A_2, \dots$ are independent. Then the indicator variables $1_{A_1}, 1_{A_2}, \dots$ are independent. Because $\mathbb{E}(1_{A_j}^2) = \mathbb{E}(1_{A_j}) = \mathbb{P}(A_j)$, we see that

$$\operatorname{Var}(1_{A_j}) \;=\; \mathbb{E}(1_{A_j}^2) - (\mathbb{E}1_{A_j})^2 \;\le\; \mathbb{P}(A_j),$$

and it follows by independence that

$$\operatorname{Var}(N_k) \;=\; \sum_{j=1}^k \operatorname{Var}(1_{A_j}) \;\le\; \sum_{j=1}^k \mathbb{P}(A_j) \;=\; \mathbb{E}N_k.$$

Chebyshev's inequality then implies that for any real number $t > 0$,

$$\mathbb{P}(N_k \le \mathbb{E}N_k - t) \ \le \ \mathbb{P}(|N_k - \mathbb{E}N_k| \ge t) \ \le \ \frac{\mathrm{Var}(N_k)}{t^2} \ \le \ \frac{\mathbb{E}N_k}{t^2}.$$

Because $N_k \le N$, this further implies that

$$\mathbb{P}(N \le \mathbb{E}N_k - t) \ \le \ \frac{\mathbb{E}N_k}{t^2} \qquad \text{for all } k \ge 1 \text{ and } t > 0.$$

By denoting $s_k = \mathbb{E}N_k$ and substituting $t = s_k^{2/3}$, it follows that

$$\mathbb{P}\Big(N \le s_k - s_k^{2/3}\Big) \ \le \ s_k^{-1/3}.$$

Differentiation shows that the function $s \mapsto s - s^{2/3}$ is strictly increasing on the interval $\big((\frac{2}{3})^3, \infty\big)$ and converges to infinity as $s \to \infty$. Because $s_k \uparrow \infty$, we conclude that the sequence $k \mapsto s_k - s_k^{2/3}$ is nondecreasing for $k \ge k_0$ and converges to infinity as $k \to \infty$, when we select a large enough integer k_0 such that $s_{k_0} > (\frac{2}{3})^3$. As a consequence, we see that the events $F_k = \Big\{N < s_k - s_k^{2/3}\Big\}$ satisfy

$$F_{k_0} \subset F_{k_0+1} \subset F_{k_0+1} \subset \cdots \qquad \text{and} \qquad \bigcup_{k=k_0}^{\infty} F_k \ = \{N < \infty\}.$$

The monotone continuity of measures now implies that

$$\mathbb{P}(N < \infty) \ = \ \lim_{k\to\infty} \mathbb{P}(F_k) \ \le \ \lim_{k\to\infty} s_k^{-1/3} \ = \ 0.$$

Hence $\mathbb{P}(N < \infty) = 0$, confirming implication (11.2). □

Example 11.3 (Records)**.** Let $X_1, X_2, \dots$ be independent real-valued random variables distributed according to a probability measure μ. Assume that μ is *diffuse* in the sense that $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$. We assume that X_n represents the n-th observation in a sequence (e.g. sports score, temperature measurement). Let

$$A_n \ = \ \big\{X_n = \max(X_1, \dots, X_n)\big\}$$

denote the event that the n-th observation is a record in the sequence. Then

$\mathbb{P}(A_1) = 1$. An application of Fubini's theorem implies that

$$\begin{aligned} \mathbb{P}(X_1 = X_2) &= \int_{\mathbb{R}^2} 1_D(x,y)\,(\mu \otimes \mu)(dx,dy) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_{\{x\}}(y)\,\mu(dy) \right) \mu(dx) \\ &= \int_{\mathbb{R}} \mu(\{x\})\,\mu(dx) \\ &= 0. \end{aligned}$$

Because the probability of a tie break equals zero, we find that

$$\begin{aligned} \mathbb{P}(X_2 = \max(X_1,X_2)) &= \mathbb{P}(X_2 \geq X_1) \\ &= \mathbb{P}(X_2 > X_1) + \mathbb{P}(X_2 = X_1) \\ &= \mathbb{P}(X_2 > X_1). \end{aligned}$$

By symmetry, we see that $\mathbb{P}(X_2 > X_1) = \mathbb{P}(X_1 > X_2)$. Therefore,

$$\begin{aligned} \mathbb{P}(X_2 > X_1) &= \frac{1}{2}\Big(\mathbb{P}(X_2 > X_1) + \mathbb{P}(X_1 > X_2) \Big) \\ &= \frac{1}{2}\Big(\underbrace{\mathbb{P}(X_2 > X_1) + \mathbb{P}(X_1 > X_2) + \mathbb{P}(X_1 = X_2)}_{1} \Big) \;=\; \frac{1}{2}. \end{aligned}$$

Hence $\mathbb{P}(A_2) = \frac{1}{2}$. Similarly, one can verify that $\mathbb{P}(A_n) = \frac{1}{n}$ for all n. We conclude that $\sum_{n=1}^\infty \mathbb{P}(A_n) = \sum_{n=1}^\infty \frac{1}{n} = \infty$. It is also possible (Exercise 11.10) to verify that the events $A_1, A_2, \dots$ are independent. For example, because tie breaks among X_1, X_2, X_3 have zero probability, it follows, again by symmetry, that

$$\begin{aligned} \mathbb{P}(A_3, A_2) &= \mathbb{P}(A_3,\, X_2 > X_1) \\ &= \frac{1}{2}\Big(\mathbb{P}(A_3,\, X_2 > X_1) + \mathbb{P}(A_3,\, X_1 > X_2) \Big) \\ &= \frac{1}{2}\Big(\mathbb{P}(A_3,\, X_2 > X_1) + \mathbb{P}(A_3,\, X_1 > X_2) + \underbrace{\mathbb{P}(A_3,\, X_1 = X_2)}_{0} \Big) \\ &= \frac{1}{2}\mathbb{P}(A_3) \\ &= \mathbb{P}(A_3)\mathbb{P}(A_2). \end{aligned}$$

By similar computations, one may verify that

$$\mathbb{P}(A_1, A_2, \dots, A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\cdots\mathbb{P}(A_n).$$

After extending this to joint events involving complements of A_j it follows that $A_1, A_2, \ldots$ are independent. Hence by Borel–Cantelli (Theorem 11.2), the record counter $N = \sum_{n=1}^\infty 1_{A_n}$ satisfies $N = \infty$ with probability one. We conclude that almost surely, *new records will be made infinitely many times.*

Example 11.4 (Consecutive records)**.** Continuing Example 11.3, let $\tilde A_n = A_n \cap A_{n-1}$ denote the event that a record occurs twice in a row. We saw that the events A_n and A_{n-1} are independent, and that $\mathbb{P}(A_n) = \frac{1}{n}$. Therefore,

$$\sum_{n=2}^\infty \mathbb{P}(\tilde A_n) \ = \ \sum_{n=2}^\infty \frac{1}{n(n-1)} \ \le \ \sum_{n=2}^\infty \frac{1}{(n-1)^2} \ = \ \sum_{n=1}^\infty \frac{1}{n^2} \ < \ \infty.$$

The events $\tilde A_2, \tilde A_3, \ldots$ are *not* independent, but may nevertheless apply the first implication of the Borel–Cantelli theorem (Theorem 11.2) to conclude that the consecutive record counter $\tilde N = \sum_{n=1}^\infty 1_{\tilde A_n}$ is finite almost surely.

11.4 Transforms

The *moment generating function* of a probability measure μ on $\bar{\mathbb{R}}$ is a function $M_\mu \colon \mathbb{R} \to \bar{\mathbb{R}}_+$ defined by

$$M_\mu(t) \ = \ \int_{\bar{\mathbb{R}}} e^{tx} \mu(dx).$$

The moment generating function M_X of a random variable X is defined as that of the law of X, so that

$$M_X(t) \ = \ \mathbb{E} e^{tX}.$$

The *characteristic function* of a probability measure μ on $\mathbb{R}$ is a function $\phi_\mu \colon \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_\mu(t) \ = \ \int_{\mathbb{R}} \cos(tx)\, \mu(dx) + i \int_{\mathbb{R}} \sin(tx)\, \mu(dx).$$

The characteristic function ϕ_X of a random variable X is defined as that of the law of X, so that

$$\phi_X(t) \ = \ \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX).$$

This can be also written as $\phi_X(t) = \mathbb{E} e^{itX}$ when we extend the definition of integral to complex-valued functions with integrable real and complex part by defining $\int f\, d\mu = \int \mathrm{Re}(f)\, d\mu + i \int \mathrm{Im}(f)\, d\mu$.

11.5 Exercises

Exercise 11.5 (Almost sure limits are nonunique)**.** Assume that $X_n \xrightarrow{\text{a.s.}} X$ and $\tilde{X} = X$ almost surely. Prove that $X_n \xrightarrow{\text{a.s.}} \tilde{X}$.

Exercise 11.6 (Escaping mass)**.** Let $X_1, X_2, \ldots$ be real-valued random variables defined some some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, distributed according to $\mathrm{Law}(X_n) = \mu_n$, where

$$\mu_n \;=\; (1 - \frac{1}{n})\delta_0 + \frac{1}{n}\delta_n.$$

Are the following statements true or false? If true, describe what the limit is. If false, explain why the sequence does not converge.

(a) $X_n \xrightarrow{\text{a.s.}} X$ for some limiting random variable X.

(b) $X_n \xrightarrow{\mathbb{P}} X$ for some limiting random variable X.

(c) $\mu_n \to \mu$ in the total variation metric for some probability measure μ.

(d) $\mu_n \to \mu$ in the Wasserstein-1 metric for some probability measure μ.

(e) $\mu_n \to \mu$ in the Wasserstein-2 metric for some probability measure μ.

Exercise 11.7 (Expectations vs tail integrals)**.** Prove that for any nonnegative random variable:

(a) $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)\, dt$.

(b) $\mathbb{E}X^\alpha = \int_0^\infty \mathbb{P}(X > t^{1/\alpha})\, dt$ for any $\alpha > 0$.

Exercise 11.8 (Noise peaks)**.** Let $\xi, \xi_1, \xi_2, \ldots$ be independent and identically distributed nonnegative random variables. Denote $\xi_n = O(n^\beta)$ if there exists a constant c such that $\xi_n \le cn^\beta$ for all n starting from some n_0 onwards.

(a) Show that for all $\alpha > 0$ and $c > 0$, the following are equivalent:

- $\xi_n \le cn^{1/\alpha}$ starting from some n onwards, almost surely.
- $\mathbb{E}\xi^\alpha < \infty$.

Hint: Exercise 11.7 and Borel–Cantelli.

(b) Show that if $\mathbb{E}\xi^\alpha < \infty$, then $\xi_n = O(n^{1/\alpha})$ almost surely.

(c) Show that if $\mathbb{E}\xi^\alpha = \infty$, then $\limsup_{n\to\infty} n^{-1/\alpha}\xi_n = \infty$ almost surely.

(d) Based on the above, can you deduce something about the behaviour of $M_n = \max(\xi_1, \dots, \xi_n)$ for large n?

Exercise 11.9 (Cramér meets Bernoulli)**.** Let $S_n = X_1 + \cdots + X_n$ be a sum of independent $\mathrm{Ber}(p)$-distributed random integers.

(a) Determine the moment generating function $M_1(t) = \mathbb{E} e^{tX_1}$.

(b) Determine the moment generating function $M_n(t) = \mathbb{E} e^{tS_n}$.

(c) Fix $a \in (p, 1)$. Prove that $\mathbb{P}(\frac{1}{n} S_n \ge a) \le e^{-ant} M_n(t)$ for all $t \ge 0$.
Hint: Investigate what Markov's inequality tells about e^{tS_n}.

(d) Determine a value $t_* = t$ that yields the sharpest bound in (c).

(e) With the help of (d), prove that

$$\mathbb{P}\left(\frac{1}{n} S_n \ge a\right) \; \le \; e^{-nD(a\|p)},$$

where $D(a\|p) = (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p}$ equals the Kullback–Leibler divergence of $\mathrm{Ber}(p)$ from $\mathrm{Ber}(a)$.

(f) Apply (e) to compute an upper bound for the probability that the relative share of heads among 350000 fair coin flips[5] is at least 51%.

Exercise 11.10 (Records)**.** Prove that the events $A_1, A_2, \dots$ in Example 11.3 are stochastically independent.

Exercise 11.11 (Ky Fan distance)**.** For real-valued random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ define the *Ky Fan distance* [Fan43] by

$$d_{\mathrm{KF}}(X,Y) \; = \; \inf\left\{\epsilon > 0 \colon \mathbb{P}(|X-Y| \ge \epsilon) < \epsilon\right\}.$$

(a) Prove that $d_{\mathrm{KF}}(X,Y) = 0$ if and only if $X = Y$ almost surely.

(b) Prove the triangle inequality $d_{\mathrm{KF}}(X,Z) \le d_{\mathrm{KF}}(X,Y) + d_{\mathrm{KF}}(Y,Z)$.

[5] See `https://arxiv.org/abs/2310.04153` for an experimental study.

Chapter 12

Central limit theorems

The central limit theorem, arguably the most famous result of probability theory, states that under mild regularity conditions, the law of a properly normalised sum of independent random variables is approximated by the *standard normal distribution*, the probability measure

$$B \mapsto \int_B \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx \qquad \text{on } (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

There are many versions of the central limit theorem. The Lindeberg–Lévy version establishes weak convergence, and Tanaka's version convergence in a stronger Wasserstein-2 metric. We will discuss these convergence concepts, and take a glimpse of what happens for heavy-tailed random variables with infinite second moments.

Key concepts: weak convergence, Wasserstein distance

Learning outcomes:

- Learn what convergence in distribution means, and how to analyse it using characteristic functions.
- Get introduced to the concept of coupling and optimal transportation.
- Learn to properly normalise a sum of random variables to get a limiting distribution.
- Learn to identify situations in which a sum of random variables can or cannot be approximation by a normal distribution.

Prerequisites: Previous chapters.

12.1 Weak law of large numbers

Theorem 12.1 (Weak law of large numbers)**.** *Let $S_n = X_1 + \cdots X_n$ be a sum of independent and identically distributed square-integrable random variables with a common mean $m = \mathbb{E}X_1$. Then*

$$\frac{1}{n} S_n \xrightarrow{\mathbb{P}} m \qquad \text{as } n \to \infty.$$

Proof. Because $\mathbb{E}S_n = mn$, we find that by Chebyshev's inequality that

$$\mathbb{P}(|\tfrac{1}{n}S_n - m| > \epsilon) \;=\; \mathbb{P}(|S_n - \mathbb{E}S_n| > \epsilon n) \;\le\; \frac{\operatorname{Var}(S_n)}{(\epsilon n)^2}.$$

Fix a number $\epsilon > 0$. Because $\operatorname{Var}(S_n) = n \operatorname{Var}(X_1)$ due to independence, it follows that

$$\mathbb{P}(|\tfrac{1}{n}S_n - m| > \epsilon) \;\le\; \frac{\operatorname{Var}(X_1)}{\epsilon^2 n}.$$

We conclude that $\mathbb{P}(|\frac{1}{n}S_n - m| > \epsilon) \to 0$. Hence $\frac{1}{n}S_n \xrightarrow{\mathbb{P}} m$. □

12.2 Weak convergence of probability measures

A sequence of probability measures on $\mathbb{R}$ *converges weakly*[1], denoted $\mu_n \xrightarrow{w} \mu$, if $\int_{\mathbb{R}} \phi \, d\mu_n \to \int_{\mathbb{R}} \phi \, d\mu$ for all bounded continuous functions $\phi \colon \mathbb{R} \to \mathbb{R}$. The theory of weak convergence of probability measures on a metric space is a broad topic, worthy of a lecture course of its own. There are several equivalent characterisations for weak convergence. Below are the most important

Theorem 12.2. *The following are equivalent for any probability measures $\mu_1, \mu_2, \dots, \mu$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with characteristic functions $\phi_1, \phi_2, \dots, \phi$:*

(i) $\mu_n \xrightarrow{w} \mu$.

(ii) $\phi_{X_n} \to \phi_X$ *pointwise.*

(iii) $\mu_n(A) \to \mu(A)$ *for all* $A \in \mathcal{B}(\mathbb{R})$ *such that* $\mu(\partial A) = 0$ *where* ∂A *is the boundary of* A.

[1] Finnish: 'suppenee heikosti'

Proof. The proof is omitted but well documented in standard textbooks (for example [Bil99], [Dud02], [Çın11]). The equivalence of (i)–(ii) is called the *Lévy continuity theorem*. □

12.3 Gaussian central limit theorems

The following famous result is attributed to Lindeberg[2] and Lévy[3] who proved it in a series of articles published in 1922. The result says that the law of a properly normalised sum of n independent and identically distributed random variables with finite second moments converges to the standard normal distribution as $n \to \infty$ in the following sense.

Theorem 12.3 (Lindeberg–Lévy central limit theorem)**.** *Let $S_n = X_1 + \cdots + X_n$ be a sum of independent and identically distributed square-integrable random variables. Then*

$$\operatorname{Law}\left(\frac{S_n - \mathbb{E}S_n}{\sqrt{\operatorname{Var}(S_n)}}\right) \xrightarrow{w} \operatorname{Nor}(0,1) \qquad \text{as } n \to \infty.$$

The normal distribution is universal. No matter what the shape of the law of X is (e.g. binomial, Poisson, exponential — the normalised limiting distribution is always the same standard normal distribution, as long as the summands are independent and square integrable.

Proof. (i) Denote by X a generic random variable having the same distribution as $X_1, X_2, \ldots$ Let us assume that $\mathbb{E}X = 0$ and $\operatorname{Var}(X) = 1$. Then $\mathbb{E}S_n = 0$ and $\operatorname{Var}(S_n) = n$ by independence. Then we need to show that

$$\operatorname{Law}(S_n/\sqrt{n}) \xrightarrow{w} \operatorname{Law}(Z),$$

where Z is a generic random variable having the standard normal distribution.

Observe that the characteristic function of $S_n = \sum_{j=1}^n X_j$ equals

$$\phi_{S_n}(\theta) \;=\; \mathbb{E}e^{i\theta \sum_{j=1}^n X_j} \;=\; \mathbb{E}\prod_{j=1}^n e^{i\theta X_j} \;\overset{\text{indep}}{=}\; \prod_{j=1}^n \mathbb{E}e^{i\theta X_j} \;=\; \phi_X(\theta)^n.$$

[2] Jarl Lindeberg, 1876–1932, PhD 1901 @ U Helsinki for E Lindelöf.

[3] Paul Lévy, 1886–1971, PhD 1911 @ Paris for J Hadamard.

As a consequence,

$$\phi_{S_n/\sqrt{n}}(\theta) \;=\; \mathbb{E}e^{i\theta S_n/\sqrt{n}} \;=\; \phi_{S_n}(\theta/\sqrt{n}) \;=\; \phi_X(\theta/\sqrt{n})^n.$$

(ii) When X is square-integrable, it is possible (with the help of the dominated continuity of expectation) to verify that $\phi(\theta)$ is twice differentiable with Taylor expansion

$$\phi_X(\theta) \;=\; \phi_X(0) + \phi'(0)\theta + \frac{1}{2}\phi''(0) + r(\theta), \tag{12.1}$$

in which $r(\theta)/\theta^2 \to 0$ as $\theta \to 0$, and

$$\begin{aligned}\phi'(\theta) \;&=\; \frac{d}{d\theta}\mathbb{E}e^{i\theta X} \;=\; \mathbb{E}\left(\frac{d}{d\theta}e^{i\theta X}\right) \;=\; \mathbb{E}\left(iXe^{i\theta X}\right),\\ \phi''(\theta) \;&=\; \frac{d}{d\theta}\mathbb{E}\left(iXe^{i\theta X}\right) \;=\; \mathbb{E}\left(iX\frac{d}{d\theta}e^{i\theta X}\right) \;=\; \mathbb{E}\left((iX)^2e^{i\theta X}\right).\end{aligned}$$

In particular, $\phi'(0) = i\mathbb{E}X$ and $\phi''(0) = -\mathbb{E}X^2$. Because $\mathbb{E}X = 0$ and $\mathrm{Var}(X) = 1$, we see that $\mathbb{E}X^2 = 1$, and the Taylor approximation (12.1) becomes

$$\phi_X(\theta) \;=\; 1 - \frac{1}{2}\theta^2 + r(\theta).$$

Then

$$\phi_X(\theta/\sqrt{n}) \;=\; 1 - \frac{1}{2}\theta^2/n + r(\theta/\sqrt{n}),$$

and it follows (with some work) that for any $\theta \in \mathbb{R}$,

$$\lim_{n\to\infty}\phi_{S_n/\sqrt{n}}(\theta) \;=\; \lim_{n\to\infty}\left(1 - \frac{\theta^2/2}{n} + r(\theta/\sqrt{n})\right)^n \;=\; e^{-\theta^2/2}.$$

(iii) We note (homework) that the characteristic function of the standard normal distribution equals

$$\phi_Z(\theta) \;=\; e^{-\theta^2/2}.$$

By (ii), we conclude that

$$\lim_{n\to\infty}\phi_{S_n/\sqrt{n}}(\theta) \;=\; \phi_Z(\theta).$$

Lévy's continuity theorem (Theorem 12.2) implies that $\mathrm{Law}(S_n/\sqrt{n}) \xrightarrow{w} \mathrm{Law}(Z)$. □

12.4 Wasserstein distances

The *Wasserstein distance* of order $p \in [1, \infty)$ between probability measures on $\mathbb{R}$ is defined by

$$W_p(\mu, \nu) \;=\; \inf_{(X,Y) \in \Gamma(\mu,\nu)} \Big(\mathbb{E}|X - Y|^p \Big)^{1/p}, \tag{12.2}$$

where $\Gamma(\mu, \nu)$ denotes the set of *couplings* of μ and ν, that is, the set of random vectors (X, Y) such that $\operatorname{Law}(X) = \mu$ and $\operatorname{Law}(Y) = \nu$. Wasserstein distances are named after Leonid Vaserstein[4], and also many other such as Dall'Aglio, Gini, Kantorovich, Mallows, Rubinstein. Also the term *earth mover's distance* is used.

Wasserstein distances are similar in spirit to total variation distances. In light of Proposition 9.7, we see that

$$d_{\mathrm{tv}}(\mu, \nu) \;=\; \inf_{(X,Y) \in \Gamma(\mu,\nu)} \mathbb{P}(X \neq Y). \tag{12.3}$$

It is possible, but nontrivial, to show that the infima in (12.2) and (12.3) are always attained by some (usually not equal) couplings. Geometrically, the minimum corresponds to optimal transportation, where the task is to transport a unit of mass with supply in $\mathbb{R}$ distributed according to μ into new locations with demand distributed according to ν, where transportation cost from x to y equals $|x - y|^p$ in (12.2) and $1(x \neq y)$ in[5] (12.3). A probability measure $\gamma = \operatorname{Law}(X, Y)$ on $\mathbb{R}^2$ corresponds to a transportation plan in which $\gamma(dx, dy)$ is the amount of mass transported from x to y.

We denote by $\mathcal{P}(\mathbb{R})$ the collection of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and by

$$\mathcal{P}_p(\mathbb{R}) \;=\; \Big\{ \mu \in \mathcal{P}(\mathbb{R}) \colon \int_{\mathbb{R}} x^p \, \mu(dx) < \infty \Big\}$$

the collection of probability measures on $\mathbb{R}$ with finite p-th moments.

Proposition 12.4. *The Wasserstein distance of order $p \in [1, \infty)$ is a metric on the space $\mathcal{P}_p(\mathbb{R})$ of probability measures on $\mathbb{R}$ with finite p-th moments.*

Proof. This is harder to than what one might expect. A common proof of the triangle inequality is based on two things: (i) verify that the Wasserstein

[4] Leonid N Vaserstein (PhD 1969 @ Moscow State University) is a professor at Penn State University, USA. His last name is spelled in many ways: Vasershtein, Wasserstein, Waserstein, Vasserstein, Vaseršteǐn, Vassershtejn, Vasershtejn,

[5] We denote $1(A) = 1$ when A is valid and $1(A) = 0$ when not.

minimisation problem always admits an optimal solution (optimal coupling), and (ii) verify that two optimal couplings can be further coupled to a suitable trivariate random vector (see [Vil09] for details). There might be simpler proof in [Dud02]. □

Proposition 12.5. *The Wasserstein distance of order $p \in [1, \infty)$ for probability measures on $\mathbb{R}$ can be computed as*

$$W_p(\mu_1, \mu_2) \;=\; \left(\mathbb{E}(Q_1(U) - Q_2(U))^p\right)^{1/p}, \tag{12.4}$$

where Q_1, Q_2 are quantile functions of μ_1, μ_2, and U is a uniformly distributed random variable in $(0, 1)$.

A quantile coupling $(Q_1(U), Q_2(U))$ is an optimal coupling for Wasserstein distance of order p.

Proof. Major [Maj78, Theorem 8.1] proves this in a rather simple manner. Check that the proof has no gaps. Maybe a similar proof can be done along the lines in [Dud02]. □

Example 12.6 (People transportation)**.** Fix a number $x > 0$ and consider probability measures on $\mathbb{R}$ defined by $\mu = \frac{1}{2}\delta_{-x} + \frac{1}{2}\delta_0$ and $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_x$. We may interpret this as people transportation in which

- 50% of supply is located in Vaasa (location $-x$) and 50% in Oulu (location 0).
- 50% of demand is located in Oulu (location 0) and 50% in Sodankylä (location x).

Let (X_0, Y_0) and (X_1, Y_1) be random vectors in $\mathbb{R}^2$ distributed according to

$$\begin{aligned}
\gamma_0 &= \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(-x,x)}, \\
\gamma_1 &= \frac{1}{2}\delta_{(-x,0)} + \frac{1}{2}\delta_{(0,x)}.
\end{aligned}$$

We see that both (X_0, Y_0) and (X_1, Y_1) are couplings of μ and ν. Their laws correspond to transportation plans:

- In plan γ_0, people from Oulu are transported to Oulu, and people from Vaasa to Sodankylä.

- In plan γ_1, people from Vaasa are transported to Oulu, and people from Oulu to Sodankylä.

For these couplings, we see that

$$\begin{aligned} \mathbb{P}(X_0 \neq Y_0) &= \frac{1}{2} & \qquad \left(\mathbb{E}|X_0 - Y_0|^2\right)^{1/2} &= \sqrt{2}x \\ \mathbb{P}(X_1 \neq Y_1) &= 1 & \qquad \left(\mathbb{E}|X_1 - Y_1|^2\right)^{1/2} &= x. \end{aligned}$$

By Proposition 9.4, we see that $d_{\mathrm{tv}}(\mu,\nu) = \frac{1}{2}$. In light of (12.3), we see that (X_0, Y_0) is an optimal coupling for the total variation distance. One may also verify that the γ_1 equals the law of the quantile coupling in (12.4). Because $\gamma_1 = \mathrm{Law}(X_1, Y_1)$, we conclude that (X_1, Y_1) is an optimal coupling for W_2, and $W_2(\mu,\nu) = x$.

The quantile coupling is always optimal for Wasserstein distances, but not necessarily for the total variation distance.

12.5 Tanaka central limit theorem

Te following stronger version of the Gaussian central limit theorem was given by Hiroshi Tanaka[6] [Tan73] and later independently in [JS05].

Theorem 12.7 (Tanaka central limit theorem). *Let $S_n = X_1 + \cdots X_n$ be a sum of independent and identically distributed square-integrable random variables. Then*

$$\mathrm{Law}\left(\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}(S_n)}}\right) \xrightarrow{W_2} \mathrm{Nor}(0,1) \qquad \textit{as } n \to \infty.$$

Proof. A student-friendly proof does not yet exist, but hopefully appears here sooner or later. A short (3 pages) but rather demanding proof is available in Tanaka's original research article [Tan73, Section 2]. A longer, but equally demanding proof is in [JS05]. □

12.6 Non-Gaussian central limit theorems

What if the summands in $S_n = X_1 + \cdots X_n$ have infinite second moments (not square-integrable)? In this case we cannot normalise S_n to have a finite

[6] Hiroshi Tanaka, PhD 1958 @ Kyushu University for G Maruyama.

variance, we cannot normalise S_n to obtain a Gaussian limit in distribution. Nevertheless, under mild regularity about the tails of X_1, it is possible to prove that with a proper scaling, different from $n^{1/2}$, there exists a non-Gaussian limit distribution. Such limiting distributions are called *stable distributions*.

Theorem 12.8 (Stable central limit theorem). *Let $S_n = X_1 + \cdots X_n$ be a sum of independent and identically distributed and symmetric random variables such that $\mathbb{P}(|X| > t) \sim ct^{-\alpha}$ as $t \to \infty$, for some $\alpha \in (0, 2)$. Then*

$$\mathrm{Law}\left(\frac{S_n}{n^{1/\alpha}}\right) \xrightarrow{w} G_\alpha \qquad \text{as } n \to \infty,$$

for some probability measure G_α on $\mathbb{R}$.

For heavy-tailed summands with symmetric power-law tails of tail exponent $\alpha \in (0, 2)$, $\frac{S_n}{n^{1/2}}$ does not converge in law. The fluctuations of S_n are of larger order $n^{1/\alpha} \gg n^{1/2}$ compared to the usual Gaussian fluctuations. The properly normalised sums $\frac{S_n}{n^{1/\alpha}}$ converge in law to a limiting distribution G_α, a so-called α-stable distribution.

For summands with symmetric power-law tails of tail exponent $\alpha = 3$, one might guess that the fluctuations of S_n are of order $n^{1/3} \ll n^{1/2}$, and that $\frac{S_n}{n^{1/3}}$ would converge in law to some limiting distribution. This guess is wrong. Because the summands now have finite second moments, the proper normalisation is $\frac{S_n}{n^{1/2}}$. For power laws, there is huge difference in behaviour depending on whether or not $\alpha < 2$ (non-Gaussian behaviour) or $\alpha > 2$ (Gaussian domain).

Proof. Here is a proof sketch, for details see [NWZ22]. Let X be a generic random variable representing the common law of $X_1, X_2, \ldots$ Because $\mathbb{E}X^2 = \infty$, we cannot use a second-order Taylor expansion for the characteristic function $\phi_X(\theta)$ as in the proof of the Gaussian CLT. Instead, a so-called Tauberian theorem (relating the tails of X to the behaviour of ϕ_X near zero) can be applied to conclude that for some constant b,

$$\phi_X(\theta) \;=\; 1 - b|\theta|^\alpha(1 + o(1)), \qquad \text{as } t \to 0.$$

Now recall that $\phi_{S_n}(\theta) = \phi_X(\theta)^n$. Then

$$\phi_{S_n/n^{1/\alpha}}(\theta) \;=\; \phi_{S_n}(\theta/n^{1/\alpha}) \;=\; \phi_X(\theta/n^{1/\alpha})^n,$$

so that

$$\phi_{S_n/n^{1/\alpha}}(\theta) \;=\; \Big(1 - \frac{b|\theta|^\alpha}{n}(1+o(1))\Big)^n \;\to\; e^{-b|\theta|^\alpha}.$$

Lévy's continuity theorem implies that Law $\left(\frac{S_n}{n^{1/\alpha}}\right) \xrightarrow{w} G_\alpha$ where the limit is a probability measure with characteristic function $e^{-b|\theta|^\alpha}$. □

Epilogue

This short course in probability theory aimed to give a rapid introduction to the modern theory of probability, in its rigorous measure-theoretic format. Due to time and space limitations, several important things had to be left out. For example, conditional expectations with respect sub-sigma-algebras, concentration inequalities of probability theory, large deviations theory, stochastic ordering techniques, random measures and point patterns, continuous-time stochastic processes. Nevertheless, I hope that this course has given you a solid foundation in the foundations of probability theory and raised your curiosity to pursue more advanced topics in probability and statistics.

Here are some recommendations for further reading in the field of probability theory:

- E Çınlar. *Probability and Stochastics* [Çın11]. This is an excellent textbook on probability theory and stochastic processes, designed for a long (30–32 weeks) course in Princeton University. Its first three chapters roughly correspond to the scope of this course. The later chapters give a nice introduction to the theory of martingales, Poisson point patterns, and Markov and Lévy processes, including Brownian motion.

- R Dudley. *Real Analysis and Probability.* [Dud02] This excellent textbook has been used in probability theory and real analysis courses at MIT. Besides probability theory, the book contains loads of clearly written material on real analysis and topology. This is one of the rare probability textbooks also presenting results on Wasserstein distances.

- O Kallenberg. *Foundations of Modern Probability.* This famous reference book has three editions: 1997, 2002, 2021. My favourite is the second volume [Kal02]. This book contains a dazzling amount of theory packed into its 600+ pages, with elegant proofs written in a very concise manner. Warning: The proofs are short and elegant, but may demand some independent effort from the reader to open up. If you know what you are looking, you will probably find it here. The book assumes a solid background in metric spaces.

- R Vershynin. *High-Dimensional Probability* [Ver18]. Despite its still rather young age, this book can already be called a classic. The textbook gives a nice introduction to the analysis of random vectors and random matrices, with a focus on high-dimensional settings.

- RL Schilling, FT Kühn. *Counterexamples in Measure and Integration* [SK21]. This curious book offers loads of fascinating counterexamples in probability theory and measure theory, many of which are truly counterintuitive.

Appendix A

Extended half-line

The *extended half-line* is a set $[0,\infty] = \mathbb{R}_+ \cup \{\infty\}$, where $\mathbb{R}_+$ denotes the nonnegative real numbers, and ∞ is an element not in $\mathbb{R}_+$. The extended half-line has a natural algebraic, order-theoretic, and topological structure. The open sets of the topology further generate a Borel sigma-algebra on this space.

A.1 Algebraic structure

The sum and product on $\mathbb{R}_+$ are extended to $[0,\infty]$ by defining

$$x + \infty \;=\; \infty + x \;=\; \infty \qquad \text{for } x \geq 0,$$

and

$$x \cdot \infty \;=\; \infty \cdot x \;=\; \begin{cases} 0 & \text{for } x = 0, \\ \infty & \text{for } x > 0. \end{cases}$$

The set $[0,\infty]$ equipped with these operations is a semi-ring[1] with additive identity 0 and multiplicative identity 1.

Addition and multiplication on $[0,\infty]$ satisfy the usual arithmetic rules of addition and multiplication on $\mathbb{R}_+$.

A.2 Order

We define a relation $\leq$ on $[0,\infty]$ by saying that $x \leq y$ if either $y = \infty$, or $x, y \in \mathbb{R}_+$ and $x \leq y$ in the usual ordering on the real line. We denote $x < y$

[1] https://en.wikipedia.org/wiki/Semiring

whenever $x \le y$ and $x \ne y$. Then set $[0, \infty]$ then becomes totally ordered[2], and a complete lattice in the sense that $\inf(A), \sup(A) \in [0, \infty]$ for every nonempty $A \subset [0, \infty]$. We denote intervals with endpoints $a, b \in [0, \infty]$ by (a, b), $(a, b]$, $[a, b)$, and $[a, b]$ as usual.

A.3 Topology

A subset of $[0, \infty]$ is called *open* if it can be expressed as a union of *open intervals*

$$(a, b) = \{x \colon a < x < b\} \qquad \text{with } a, b \in \mathbb{R}_+$$

and *open rays*

$$[0, a) = \{x \colon x < a\} \quad \text{and} \quad (a, \infty] = \{x \colon x > a\} \qquad \text{with } a \in \mathbb{R}_+.$$

A subset of $[0, \infty]$ is called *closed* if its complement is open. The family of open sets $\mathcal{T}([0, \infty])$ is called the *topology*[3] of $[0, \infty]$.

One may verify that $F \colon [0, \infty] \to [0, 1]$ defined by

$$F(x) \;=\; \begin{cases} 1 - e^{-x}, & 0 \le x < \infty, \\ 1, & x = \infty \end{cases}$$

is an increasing and continuous bijection with an increasing and continuous inverse

$$F^{-1}(x) \;=\; \begin{cases} \log \frac{1}{1-x}, & 0 \le x < 1, \\ \infty, & x = 1. \end{cases}$$

Therefore F serves as an order isomorphism and a topology isomorphism (homeomorphism) between $[0, \infty]$ and $[0, 1]$. Especially, we find that $[0, \infty]$ is a compact and connected topological space. We can express the topology of the extended half line as $\mathcal{T}([0, \infty]) = F^{-1}(\mathcal{T}([0, 1]))$.

The sets $[0, \infty]$ and $[0, 1]$ are topologically and order-theoretically equivalent.

[2] A partial order is a relation $\le$ that is reflexive ($x \le x$), antisymmetric ($x \le y$, $y \le x \implies x = y$), and (transitive $x \le y$, $y \le z \implies x \le z$). A total order is a partial order such for every x, y, either $x \le y$ or $y \le x$.

[3] A topology is a set family that contains $\emptyset, S$ and is closed under arbitrary unions and finite intersections.

Proposition A.1. *Every open set in $[0, \infty]$ can be expressed as a countable union of open intervals and open rays with rational endpoints.*

Open intervals and open rays with rational endpoints constitute a countable basis for the topology of $[0, \infty]$.

Proof. Fix an arbitrary nonempty open set $U \subset [0, \infty]$. By definition, U can be expressed as a union

$$U \;=\; \Bigg(\bigcup_{a \in I} (a, \infty] \Bigg) \cup \Bigg(\bigcup_{b \in J} [0, b) \Bigg) \cup \Bigg(\bigcup_{(a,b) \in K} (a, b) \Bigg)$$

for some $I, J \subset \mathbb{R}_+$ and some set $K \subset \mathbb{R}_+ \times \mathbb{R}_+$ of pairs (a, b) with $a < b$. Denote by $\mathbb{Q}_+$ the set of nonnegative rational numbers. For any $a, b \in \mathbb{R}_+$ we may fix sequences $a_n, b_n \in \mathbb{Q}_+$ such that $a_n \downarrow a$ and $b_n \uparrow b$. Then

$$U \;=\; \Bigg(\bigcup_{a \in I} \bigcup_{n=1}^{\infty} (a_n, \infty] \Bigg) \cup \Bigg(\bigcup_{b \in J} \bigcup_{n=1}^{\infty} [0, b_n) \Bigg) \cup \Bigg(\bigcup_{(a,b) \in K} \bigcup_{n=1}^{\infty} (a_n, b_n) \Bigg). \tag{A.1}$$

We may also write

$$U \;=\; \Bigg(\bigcup_{a \in I'} (a, \infty] \Bigg) \cup \Bigg(\bigcup_{b \in J'} [0, b) \Bigg) \cup \Bigg(\bigcup_{(a,b) \in K'} (a, b) \Bigg), \tag{A.2}$$

where

$$\begin{aligned} I' &= \{a_n \colon n \ge 1,\ a \in I\}, \\ J' &= \{b_n \colon n \ge 1,\ b \in J\}, \\ K' &= \{(a_n, b_n) \colon n \ge 1,\ (a, b) \in K\} \end{aligned}$$

represent the sets of unique values appearing as endpoints in (A.1). Because the sets $\mathbb{Q}_+$ and $\mathbb{Q}_+ \times \mathbb{Q}_+$ are countable, so are the sets $I', J' \subset \mathbb{Q}_+$ and $K' \subset \mathbb{Q}_+ \times \mathbb{Q}_+$. Therefore, (A.2) is a representation of U as a desired type of countable union. □

A.4 Borel sigma-algebra

The *Borel sigma-algebra* on $[0, \infty]$ is defined as $\mathcal{B}([0, \infty]) = \sigma(\mathcal{T}([0, \infty]))$, the smallest sigma-algebra containing the open sets of $[0, \infty]$.

Proposition A.2. *The family of closed lower rays* $\{[0,t] \colon t \in \mathbb{R}_+\}$ *is a generator of the Borel sigma-algebra* $\mathcal{B}([0,\infty])$*.*

Proof. It suffices to verify that $\mathcal{C} = \{[0,t] \colon t \in \mathbb{R}_+\}$ satisfies

$$\mathcal{C} \subset \sigma(\mathcal{T}) \tag{A.3}$$

and

$$\mathcal{T} \subset \sigma(\mathcal{C}), \tag{A.4}$$

where $\mathcal{T} = \mathcal{T}([0,\infty])$ is the family of open sets in $[0,\infty]$.

Verifying (A.3) is easy because each closed ray $[0,t]$, being the complement of an open ray $(t,\infty]$, belongs to $\mathcal{T} \subset \sigma(\mathcal{T})$. To verify (A.4), we proceed in three steps.

(i) First we observe that $(a,b] \in \sigma(\mathcal{C})$ for all $a,b \in [0,\infty]$, because $(a,b] = [0,b] \cap [0,a]^c$: when $b < \infty$, both $[0,a]$ and $[0,b]$ belong to $\mathcal{C}$; when $b = \infty$, $(a,b] = [0,a]^c$ is the complement of a set in $\mathcal{C}$.

(ii) By applying (i), we see that $(a,b) \in \sigma(\mathcal{C})$ for all $a,b \in [0,\infty]$, because

$$(a,b) \;=\; \begin{cases} \cup_{n\in\mathbb{N}}(a, b-\frac{1}{n}], & b < \infty, \\ \cup_{n\in\mathbb{N}}(a,n], & b = \infty. \end{cases}$$

(iii) By applying (ii), we see that $[a,b) \in \sigma(\mathcal{C})$ for all $a,b \in [0,\infty]$, because

$$\begin{aligned} [a,b) &= [0,b) \cap [0,a)^c \\ &= \Big([0,0] \cup (0,b)\Big) \cap \Big([0,0] \cup (0,a)\Big)^c. \end{aligned}$$

Property (A.4) follows from the above observations, because every open set in $\mathcal{T}$ can be expressed as a countable union (see Proposition A.1) of intervals of form (a,b) and $[0,a)$ and $(a,\infty]$ with $a,b \in [0,\infty]$. □

The following result provides another generator for $\mathcal{B}([0,\infty])$ which may be useful in some contexts.

Proposition A.3. $\mathcal{B}([0,\infty])$ *is the smallest sigma-algebra containing the Borel sets of* $\mathbb{R}_+$ *and the set* $\{\infty\}$*, that is, the set family* $\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\}$ *is a generator of* $\mathcal{B}([0,\infty])$*.*

Proof. The open sets of $\mathbb{R}_+ = [0,\infty)$ are those that can be represented as unions of open intervals (a,b) and open rays $[0,a)$. Hence any open set in $\mathbb{R}_+$ is also open as a subset of $[0,\infty]$. Therefore, $\mathcal{T}(\mathbb{R}_+) \subset \mathcal{T}([0,\infty])$. This implies that $\sigma(\mathcal{T}(\mathbb{R}_+)) \subset \sigma(\mathcal{T}([0,\infty]))$, or equivalently

$$\mathcal{B}(\mathbb{R}_+) \subset \mathcal{B}([0,\infty]). \tag{A.5}$$

We also note that $\{\infty\} = [0,\infty)^c$ is a closed set in $[0,\infty]$, and therefore $\{\infty\} \in \mathcal{B}([0,\infty])$. This means that the one-set family $\{\{\infty\}\}$ is contained in $\mathcal{B}([0,\infty])$, and together with (A.5) we conclude that

$$\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\} \subset \mathcal{B}([0,\infty]),$$

which by the definition of a generated sigma-algebra implies that

$$\sigma(\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\}) \subset \mathcal{B}([0,\infty]). \tag{A.6}$$

Observe next that all open intervals (a,b) of $[0,\infty]$ are also open sets of $\mathbb{R}_+$. Therefore, all such open intervals belong to $\mathcal{T}(\mathbb{R}_+) \subset \mathcal{B}(\mathbb{R}_+)$. The same argument implies all open rays of form $[0,a)$ also belong to $\mathcal{B}(\mathbb{R}_+)$. Furthermore, by writing $(a,\infty] = (a,\infty) \cup \{\infty\}$, we find that all open rays of form $(a,\infty]$ are contained in $\sigma(\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\})$. We conclude that all open intervals and open rays of $[0,\infty]$ are contained in $\sigma(\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\})$. Because every open set in $[0,\infty]$ be be expressed as a countable union of open intervals and open rays (Proposition A.1), it follows that

$$\mathcal{T}([0,\infty]) \subset \sigma(\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\}),$$

which by the definition of a generated sigma-algebra implies that

$$\mathcal{B}([0,\infty]) \subset \sigma(\mathcal{B}(\mathbb{R}_+) \cup \{\{\infty\}\}). \tag{A.7}$$

The claim follows by combining (A.6)–(A.7).

□

A.5 Measurable functions

Let $(S,\mathcal{S})$ be a measurable space. A function $f\colon S \to [0,\infty]$ is called measurable if it is $\mathcal{S}/\mathcal{B}([0,\infty])$-measurable, that is, $f^{-1}(B) \in \mathcal{S}$ for all $B \in \mathcal{B}([0,\infty)]$.

Proposition A.4. *A function $f\colon S \to [0,\infty]$ is measurable if and only if $\{x\colon f(x) \le t\} \in \mathcal{S}$ for all $t \in \mathbb{R}_+$.*

Proof. By Proposition A.2, the set family $\{[0,t] \colon t \in \mathbb{R}_+\}$ is a generator of $\mathcal{B}([0,\infty])$. The claim follows by Proposition 3.3. □

A.6 Sequences

A *sequence* in $[0,\infty]$ is a function from $\mathbb{N}$ into $[0,\infty]$, often denoted $(x_1, x_2, \dots)$ or abbreviated as (x_n). We write $x_n \to x$ and say that (x_n) *converges* to x if for every open set U containing x there exists an integer $m \ge 1$ such that $x_n \in U$ for all $n \ge m$. In this case x is called a *limit* of the sequence.

Lemma A.5. *If a sequence in* $[0,\infty]$ *has a limit, it is unique.*

Proof. We will show that $[0,\infty]$ is Hausdorff space, that is, for any distinct x, y there exist disjoint open sets U, V containing x, y. By relabelling the points if necessary, we may assume that $x < y$.

(i) If $x < \infty$ and $y < \infty$ are such that $x < y$, then we may choose $U = [0, x+\epsilon)$ and $V = (y-\epsilon, \infty]$ where $\epsilon = (y-x)/2$.

(ii) If $x < \infty$ and $y = \infty$ then we may choose $U = [0, x+1)$ and $V = (x+1, \infty]$.

Next we note that in a Hausdorff space, all sequences have can have at most one limit. Assume that x and y are limits of (x_n), and $x \neq y$. Then fix neighbourhoods U of x and V of y which are disjoint. Then there exists $m_1, m_2 \ge 1$ such that $x_n \in U$ and $x_n \in V$ for all $n \ge m_1 \vee m_2$. This is a contradiction. Therefore, $x = y$. □

Lemma A.6. *For any sequence in* $[0,\infty]$,

(i) $x_n \to x \in [0,\infty)$ *iff for any* $\epsilon > 0$ *there exists an integer* $m \ge 1$ *such that* $|x_n - x| < \epsilon$ *for all integers* $n \ge m$.

(ii) $x_n \to \infty$ *iff for any* $0 < M < \infty$ *there exists an integer* $m \ge 1$ *such that* $x_n > M$ *for all integers* $n \ge m$.

Proof. (ia) Assume that $x_n \to 0$. Fix $0 < \epsilon < \infty$. Consider the neighbourhood $U = [0,\epsilon)$ of 0. Fix an integer $m \ge 1$ such that $x_n \in U$ for all $n \ge m$. Then $|x_n - x| = x_n < \epsilon$ for all $n \ge m$.

For the converse, let U be a neighbourhood of 0. Being open, U can be written as a union of open intervals of form (a,b) and open rays of form $[0,a)$ and $(b,\infty]$. Because U contains 0, we see that at least one of the open rays of form $[0, a)$ is contained in the union. Let $\epsilon = a$. Then fix $m \geq 1$ such that $x_n < \epsilon$ for all $n \geq m$. Then $x_n \in [0, a) \subset U$ for all $n \geq m$. Hence $x_n \to 0$.

(ib) Assume that $x_n \to x \in (0, \infty)$. Fix $\epsilon > 0$, and define $\epsilon' = \epsilon \wedge x$. Consider the neighbourhood $U = (x-\epsilon', x+\epsilon')$ of x. Because $x_n \to x$, we may fix an integer $m \geq 1$ such that $x_n \in U$ for all $n \geq m$. Then $|x_n - x| < \epsilon' \leq \epsilon$ for all $n \geq m$.

For the converse, let U be a neighbourhood of x. Then U can be written as a union of open intervals of form (a, b) and open rays of form $[0, a)$ and $(b, \infty]$. Then at least one such interval or ray must contain x:

- If (a, b) contains x, then we may choose a small $\epsilon > 0$ such that $(x - \epsilon, x + \epsilon) \subset (a, b)$. Then $(x - \epsilon, x + \epsilon) \subset U$.

- If $[0, a)$ contains x, then we may choose a small $\epsilon > 0$ such that $(x - \epsilon, x + \epsilon) \subset [0, a)$. Then $(x - \epsilon, x + \epsilon) \subset U$.

- If $(b, \infty]$ contains x, then we may choose a small $\epsilon > 0$ such that $(x - \epsilon, x + \epsilon) \subset (b, \infty]$. Then $(x - \epsilon, x + \epsilon) \subset U$.

In each of the cases we found a number $\epsilon > 0$ such that $(x-\epsilon, x+\epsilon) \subset U$. For this ϵ, we select m large enough and we conclude that $x_n \in (x - \epsilon, x + \epsilon) \subset U$ for all $n \geq m$. Therefore, $x_n \to x$.

(ii) Assume that $x_n \to \infty$. Fix $0 < M < \infty$. Let $U = (M, \infty]$. The there exists $m \geq 1$ such that $x_n \in U$ for all $n \geq m$.

For the converse, assume that for any $0 < M < \infty$ there exists an integer $m \geq 1$ such that $x_n > M$ for all $n \geq m$. Fix a neighbourhood U of ∞. Such a neighbourhood can be written as a union of open intervals of form (a, b) and open rays of form $[0, a)$ and $(b, \infty]$. Because U contains ∞, we see that U contains at least one open ray of form $(b, \infty]$. Choose $M = b$. Then we may select an integer $m \geq 1$ such that $x_n > M$ for all $n \geq m$. Then $x_n \in (b, \infty] \subset U$ for all $n \geq m$. Hence $x_n \to \infty$. □

Lemma A.7. *Any nondecreasing sequence in* $[0, \infty]$ *converges according to* $x_n \uparrow x$ *with* $x = \sup_n x_n$.

Proof. Assume that $x = \infty$. Fix a number $0 < M < \infty$. Because M is not an upper bound of (x_n), we see that $x_m > M$ for some $m \geq 1$. Because (x_n) is nondecreasing, it follows that $x_n > M$ for all $n \geq m$. It follows by Lemma A.6 that $x_n \to \infty$.

(ii) Assume that $0 < x < \infty$. Fix a number $\epsilon > 0$, and denote $\epsilon' = \epsilon \wedge x$. Because x is an upper bound of the sequence, we see that $x_n \leq x$ for all n. Because $x - \epsilon'$ is not an upper bound of the sequence, we see that $x_m > x - \epsilon$

for some integer $m \geq 1$. Therefore, $x - \epsilon' < x_n < x + \epsilon$ for all $n \geq m$. In particular $|x_n - x| < \epsilon$ for all $n \geq m$. It follows by Lemma A.6 that $x_n \to x$.

(iii) The case with $x = 0$ is trivial because $x_n = 0$ for all n. □

Lemma A.8. *For any nondecreasing sequences in* $[0, \infty]$*,*

(i) $\lim_{n\to\infty}(x_n + y_n) = \lim_{n\to\infty} x_n + \lim_{n\to\infty} y_n$.

(ii) $\lim_{n\to\infty}(cx_n) = c \lim_{n\to\infty} x_n$ *for all* $c \in [0, \infty]$.

Proof. (i) If (x_n) and (y_n) are bounded sequences, then the first claim follows from standard results on sequences in $\mathbb{R}$. Hence it suffices to consider the case in which (x_n) or (y_n) is bounded. Assume that (x_n) is unbounded. Then Lemma A.7 tells us that $x_n \uparrow \infty$. By noting that $z_n = x_n + y_n \geq x_n$ for all n, we find that $z_n \uparrow \infty$, and the claim follows. By symmetry, the claim follows similarly also in a case where (y_n) is unbounded.

(ii) Denote $x = \sup_n x_n$ and note that $x_n \uparrow x$ by Lemma A.7. Consider the following cases:

- If $x \wedge c = 0$, then both limits in (ii) are equal to 0.
- If $x \wedge c > 0$ and $x \vee c = \infty$, then both limits in (ii) are equal to ∞.
- If $x \wedge c > 0$ and $x \vee c < \infty$, then (x_n) is bounded nondecreasing sequence in $\mathbb{R}_+$ and $c \in (0, \infty)$, so the claim follows from standard results on sequences in $\mathbb{R}$ (or is easy to check directly).

□

A.7 Sums

The sum of a sequence $(x_1, x_2, \dots)$ in $[0, \infty]$ is defined by

$$\sum_{n=1}^{\infty} x_n = \lim_{N\to\infty} \sum_{n=1}^{N} x_n.$$

The above limit is well defined (Lemma A.7) because the partial sums $S_N = \sum_{n=1}^N x_n$ form a nondecreasing sequence in $[0, \infty]$.

For nonnegative functions on countable sets we shall employ the following *abstract sum notation*. An *enumeration* of a countably infinite set A is a sequence of distinct elements such that $A = \{x_1, x_2, \dots\}$. The *sum* of a function $f\colon A \to [0,\infty]$ over a countably infinite set A is defined by

$$\sum_{x \in A} f(x) \ = \ \sum_{n=1}^{\infty} f(x_n),$$

in which the sequence $x_1, x_2, \dots$ is an arbitrary enumeration of A. The following result confirms that the value of the sum is insensitive to the choice of the enumeration, so that the notation on the left side above makes sense.

Lemma A.9. *For any $f\colon A \to [0,\infty]$, any enumerations $A = \{x_1, x_2, \dots\}$ and $A = \{y_1, y_2, \dots\}$,*

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

Proof. Fix an integer $M \geq 1$. Because $\{x_1, \dots, x_M\} \subset A = \{y_1, y_2, \dots\}$, we see that $\{x_1, \dots, x_M\} \subset \{y_1, \dots, y_N\}$ for some $N \geq M$. Therefore,

$$\sum_{n=1}^{M} f(x_n) \ \leq \ \sum_{n=1}^{N} f(y_n) \ \leq \ \sum_{n=1}^{\infty} f(y_n).$$

Hence $\sum_{n=1}^{M} f(x_n) \leq \sum_{n=1}^{\infty} f(y_n)$ for all M, and it follows that

$$\sum_{n=1}^{\infty} f(x_n) \ = \ \lim_{M \to \infty} \sum_{n=1}^{M} f(x_n) \ \leq \ \sum_{n=1}^{\infty} f(y_n).$$

We conclude that $\sum_{n=1}^{\infty} f(x_n) \leq \sum_{n=1}^{\infty} f(y_n)$. A symmetric argument shows that $\sum_{n=1}^{\infty} f(y_n) \leq \sum_{n=1}^{\infty} f(x_n)$, and hence the claim follows. □

Proposition A.10. *For any $f, g\colon S \to [0,\infty]$ defined on a countable set S,*

$$\sum_{x \in S} (af(x) + bg(x)) = a \sum_{x \in S} f(x) + b \sum_{x \in S} g(x) \qquad \text{for all } a, b \in [0,\infty].$$

Proof. Fix an enumeration $S = \{x_1, x_2, \dots\}$ and denote $F_n = \sum_{k=1}^{n} f(x_k)$ and $G_n = \sum_{k=1}^{n} g(x_k)$. These are nondecreasing sequences in $[0,\infty]$, so we find that

$$\begin{aligned}\sum_{x\in S}(af(x)+bg(x)) &= \lim_{n\to\infty}\left(\sum_{k=1}^{n}(af(x_k)+bg(x_k))\right)\\ &= \lim_{n\to\infty}(aF_n+bG_n)\\ &\overset{\text{(Lemma A.8)}}{=} a\lim_{n\to\infty}F_n+b\lim_{n\to\infty}G_n\\ &= a\sum_{x\in S}f(x)+b\sum_{x\in S}g(x).\end{aligned}$$

□

A.8 Double sums

Lemma A.11. *For any $a_{i,j}\in[0,\infty]$,*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{i,j} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}a_{i,j}.$$

Proof. Fix integers $m,n\geq 1$. Note that

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{i,j} \geq \sum_{i=1}^{m}\sum_{j=1}^{n}a_{i,j} = \sum_{j=1}^{n}\sum_{i=1}^{m}a_{i,j}. \tag{A.8}$$

Because $\lim_{m\to\infty}\sum_{i=1}^{m}a_{i,j} = \sum_{i=1}^{\infty}a_{i,j}$ for any j, we see by taking limits $m\to\infty$ in (A.8) that

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{i,j} \geq \sum_{j=1}^{n}\sum_{i=1}^{\infty}a_{i,j} = \sum_{j=1}^{n}b_j, \tag{A.9}$$

where $b_j=\sum_{i=1}^{\infty}a_{i,j}$. Because $\lim_{n\to\infty}\sum_{j=1}^{n}b_j=\sum_{j=1}^{\infty}b_j$, we see by taking limits $n\to\infty$ in (A.9) that

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{i,j} \geq \sum_{j=1}^{\infty}b_j = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}a_{i,j}.$$

We have thus shown that $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j} \geq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{i,j}$. By repeating a similar reasoning with the roles of i and j interchanged, we may verify that $\sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{i,j} \geq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{i,j}$. Hence the claim follows. □

Appendix B

Monotone class theorem

This section contains the proof of the monotone class theorem (Theorem 2.8) that is repeated below. Recall that a set family on S is called a Dynkin class if it contains S and is closed under subset difference and increasing set limit.

Theorem (Monotone class theorem)**.** *If $\mathcal{C}$ is a set family that is closed under pairwise intersection and generates a sigma-algebra $\mathcal{S}$, then every Dynkin class containing $\mathcal{C}$ also contains $\mathcal{S}$.*

Lemma B.1. *Any intersection of Dynkin classes on S is a Dynkin class on S.*

Proof. The proof is similar to proving that the intersection of sigma-algebras is a sigma-algebra (Proposition 2.3), and left to the reader (Exercise B.3). □

Lemma B.2. *Any Dynkin class on S that is closed under pairwise intersections is a sigma-algebra on S.*

Proof. Let $\mathcal{D}$ be a Dynkin class on S that is closed under pairwise intersection. We will verify that $\mathcal{D}$ is actually a sigma-algebra.

(i) By definition, $S \in \mathcal{D}$. Because $\mathcal{D}$ is closed under subset difference, we see that $\emptyset = S \setminus S \in \mathcal{D}$ as well.

(ii) Assume that $A \in \mathcal{D}$. Because $S \in \mathcal{D}$ and $\mathcal{D}$ is closed under subset difference, it follows that $A^c = S \setminus A \in \mathcal{D}$.

(iii) Assume that $A_1, A_2 \in \mathcal{D}$. Because $\mathcal{D}$ is closed under pairwise intersection, we see with the help of (i) that $A_1 \cup A_2 = (A_1^c \cap A_2^c)^c \in \mathcal{D}$. Therefore, $\mathcal{D}$ is closed under *pairwise* union as well, and by induction, we see that $\mathcal{D}$ is closed under *finite* union. To verify that $\mathcal{D}$ is closed under *countable* union, consider a list of sets $A_1, A_2, \dots \in \mathcal{D}$, and denote

$B_n = A_1 \cup \cdots A_n$. Then $B_1 \subset B_2 \subset \cdots$ forms an increasing sequence of sets in $\mathcal{D}$ with limit $\cup_n B_n$. Because $\mathcal{D}$ is closed under increasing set limit, it follows that $\cup_n A_n = \cup_n B_n \in \mathcal{D}$.

(iv) By (ii) and (iii), we see that $\cap_n A_n = (\cup_n A_n^c)^c \in \mathcal{D}$ for any sequence of sets $A_1, A_2, \cdots \in \mathcal{D}$. □

Proof of the monotone class theorem. Let $\mathcal{C}$ be a set family that is closed under pairwise intersection and generates a sigma-algebra $\mathcal{S}$. Let $\mathcal{D}$ be the smallest Dynkin class that contains $\mathcal{C}$, which by Lemma B.1 is well defined as the intersection of all Dynkin classes containing $\mathcal{C}$ (this intersection is nonempty because 2^S is such a Dynkin class). We will prove below that $\mathcal{D}$ is closed under pairwise intersection:

$$A, B \in \mathcal{D} \quad \Longrightarrow \quad A \cap B \in \mathcal{D}. \tag{B.1}$$

Thereafter (B.1) combined with Lemma B.2 implies that $\mathcal{D}$ is a sigma-algebra. We conclude that $\mathcal{D}$ is a sigma-algebra containing $\mathcal{C}$. By the definition of a generator, $\mathcal{S}$ is the *smallest* sigma-algebra containing $\mathcal{C}$, and therefore

$$\mathcal{S} \subset \mathcal{D}. \tag{B.2}$$

Assume now that $\mathcal{E}$ is an arbitrary Dynkin class containing $\mathcal{C}$. The definition of $\mathcal{D}$ then implies that $\mathcal{D} \subset \mathcal{E}$. By combining this with (B.2), we conclude that $\mathcal{S} \subset \mathcal{E}$, and this yields the claim of the monotone class theorem.

Let us now finish the proof by verifying (B.1). Because $\mathcal{C}$ is closed under pairwise intersection and $\mathcal{C} \subset \mathcal{D}$, we see that

$$A \in \mathcal{C},\ B \in \mathcal{C} \quad \Longrightarrow \quad A \cap B \in \mathcal{D}. \tag{B.3}$$

We will next extend this property so that we can replace $\mathcal{C}$ on the left side of (B.3) by $\mathcal{D}$. This extension will be carried out in two stages.

(i) First, fix a set $B \in \mathcal{C}$ and define

$$\mathcal{A}_B \ = \ \{A \subset S \colon A \cap B \in \mathcal{D}\}.$$

Property (B.3) now implies that $\mathcal{C} \subset \mathcal{A}_B$. We also see that $\mathcal{A}_B$ is a Dynkin class because:

- $A_1, A_2 \in \mathcal{A}_B$ with $A_1 \subset A_2$ implies that $A_1 \cap B, A_2 \cap B \in \mathcal{D}$, so that because $\mathcal{D}$ is a Dynkin class, the set $(A_2 \setminus A_1) \cap B = (A_2 \cap B) \setminus (A_1 \cap B)$ also belongs to $\mathcal{D}$, and this means that $A_2 \setminus A_1 \in \mathcal{A}_B$.

- $A_n \in \mathcal{A}_B$, $A_n \uparrow A$ implies that $A_n \cap B \in \mathcal{D}$, $A_n \cap B \uparrow A \cap B$, so that because $\mathcal{D}$ is a Dynkin class, also $A \cap B \in \mathcal{D}$, which means that $A \in \mathcal{A}_B$.

We may hence conclude that $\mathcal{A}_B$ is a Dynkin class that contains $\mathcal{C}$. Because $\mathcal{D}$ is the smallest Dynkin class with this property, we conclude that $\mathcal{D} \subset \mathcal{A}_B$. By recalling that $B \in \mathcal{C}$ was arbitrarily chosen, and recalling the definition of $\mathcal{A}_B$, we conclude that

$$A \in \mathcal{C},\ B \in \mathcal{D} \quad \implies \quad A \cap B \in \mathcal{D}. \tag{B.4}$$

This extends (B.3).

(ii) Next, fix a set $A \in \mathcal{D}$ and define

$$\mathcal{B}_A \ = \ \{B \subset S \colon A \cap B \in \mathcal{D}\}.$$

Property (B.4) then implies that $\mathcal{C} \subset \mathcal{B}_A$. We also see that $\mathcal{B}_A$ is a Dynkin class by repeating the argument in the first part of the proof. We may hence conclude that $\mathcal{B}_A$ is a Dynkin class that contains $\mathcal{C}$. Because $\mathcal{D}$ is the smallest Dynkin class with this property, we conclude that $\mathcal{D} \subset \mathcal{B}_A$. By recalling that $A \in \mathcal{D}$ was arbitrarily chosen, and recalling the definition of $\mathcal{B}_A$, we conclude that

$$A \in \mathcal{D},\ B \in \mathcal{D} \quad \implies \quad A \cap B \in \mathcal{D}.$$

This is equivalent to (B.1). □

Exercise B.3. Prove Lemma B.1. (Hint: See the proof of Proposition 2.3).

English–Finnish dictionary

almost everywhere melkein kaikkialla

almost surely melkein varmasti

Bernoulli distribution Bernoulli-jakauma

binomial distribution binomijakauma

Borel set Borel-joukko

Borel sigma-algebra Borelin sigma-algebra

central limit theorem keskeinen raja-arvolause

characteristic function karakteristinen funktio

closed set suljettu joukko

complement komplementti

concentration inequality keskittymisepäyhtälö

countable numeroituva

countably disjointly additive numeroituvasti erilleen additiivinen

countably infinite numeroituvasti ääretön

counting measure laskurimitta

coupling kytkentä

converge in distribution supeta jakaumassa

converge in probability supeta stokastisesti

convergence in distribution jakaumasuppeneminen

convergence in probability stokastinen suppeneminen

cumulative distribution function kertymäfunktio

discrete measurable space diskreetti mitallinen avaruus

discrete sigma-algebra diskreetti sigma-algebra

Dirac measure Dirac-mitta

disintegration disintegrointi

disjointly additive erilleen additiivinen

finite äärellinen

generating family viritysperhe

disjoint sets erilliset joukot

distance etäisyys

distribution jakauma

earth mover's distance maansiirtoetäisyys

expectation odotusarvo

extended halfline laajennettu puoliakseli

extended real line laajennettu reaaliakseli

function funktio, kuvaus

generator virittäjä

indicator function indikaattorifunktio

infinite ääretön

intersection leikkaus

kernel ydin

law jakauma

law of large numbers suurten lukujen laki

Lebesgue measure Lebesguen mitta

list lista

map kuvaus, funktio

marginal distribution reunajakauma

measure mitta

measurable mitallinen

measurable space mitallinen avaruus

measure space mitta-avaruus

metric metriikka

moment generating function momentit generoiva funktio

negligible mitätön

normal distribution normaalijakauma

open set avoin joukko

optimal coupling optimikytkentä

optimal solution optimiratkaisu

optimization optimointi

partition ositus

Poisson distribution Poisson-jakauma

preimage alkukuva

probability distribution todennäköisyysjakauma

probability kernel todennäköisyysydin

probability mass function pistemassafunktio

probability measure todennäköisyysmitta

product measure tulomitta

product sigma-algebra tulosigma-algebra

pushforward measure työntömitta

quantile kvantiili

quantile function kvantiilifunktio

random graph satunnaisverkko

random matrix satunnaismatriisi

random number satunnaisluku

random sequence satunnaisjono

random variable satunnaismuuttuja

random vector satunnaisvektori

range kuvajoukko

set joukko

set difference joukkoerotus

set family joukkoperhe

set function joukkofunktio

set limit rajajoukko

sequence jono

shift invariant siirtoinvariantti

sigma-algebra sigma-algebra

stochastic process stokastinen prosessi

subset osajoukko

total variation distance kokonaisvaihteluetäisyys

topology topologia

uncountable ylinumeroituva

uniform distribution tasajakauma

union yhdiste

Wassertein distance Wasserstein-etäisyys

weak convergence heikko suppeneminen

weak topology heikko topology

Bibliography

[Bil99] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, second edition, 1999.

[Çın11] Erhan Çınlar. *Probability and Stochastics.* Springer, 2011.

[Dud02] Richard M. Dudley. *Real Analysis and Probability.* Cambridge University Press, second edition, 2002.

[Dyn61] Eugene B. Dynkin. *Theory of Markov Processes.* Prentice Hall and Pergamon Press, 1961.

[Fan43] Ky Fan. Entfernung zweier zufälliger Größen und die Konvergenz nach Wahrscheinlichkeit. *Mathematische Zeitschrift*, 49:685–683, 1943.

[JS05] Oliver Johnson and Richard J. Samworth. Central limit theorem and convergence to stable laws in Mallows distance. *Bernoulli*, 11(5):829–845, 2005.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability.* Springer, second edition, 2002.

[Kol33] Andrey Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, 1933.

[Maj78] Péter Major. On the invariance principle for sums of independent identically distributed random variables. *Journal of Multivariate Analysis*, 8(4):487–517, 1978.

[NWZ22] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The Fundamentals of Heavy Tails.* Cambridge University Press, 2022.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* McGraw–Hill, third edition, 1976.

[Sie28] Wacław Sierpiński. Une théorème générale sur les familles d'ensemble. *Fundamenta Mathematicae*, 12:206–210, 1928.

[SK21] René L. Schilling and Franziska T. Kühn. *Counterexamples in Measure and Integration.* Cambridge University Press, 2021.

[SV06] Glenn Shafer and Vladimir Vovk. The sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 21(1):70–98, 2006.

[Tan73] Hiroshi Tanaka. An inequality for a functional of probability distributions and its application to Kac's one-dimensional model of a Maxwellian gas. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 27(1):47–52, 1973.

[Ver18] Roman Vershynin. *High-Dimensional Probability.* Cambridge University Press, 2018.

[Vil09] Ćedric Villani. *Optimal Transport: Old and New.* Springer, 2009.

Subject index