Probability Theory

Billy Fang

Fall 2014

The following is a collection of notes for a course on measure-theoretic probability taught by Prof. Patrick Cheridito. Most of the material is drawn from [1] and [2]. Some proofs have been omitted because they are homework questions.

Contents

1 Finite probability spaces **1**

2 Countable probability spaces **5**

3 General probability spaces **5**

3.1 σ-algebras 5
3.2 Measurable functions and random variables 10
3.3 Extension theorems 11

4 The Lebesgue integral and expectation **14**

4.1 The Lebesgue integral and convergence theorems 14
4.2 Probability measures and modes of convergence 20
4.3 Uniform integrability 23
4.4 Jensen's inequality 26
4.5 Weak law of large numbers 27
4.6 Types of distributions 27
4.7 The characteristic function 29
4.8 Normal distributions 35
4.9 Gaussian processes 38

5 Martingales **39**

5.1 Conditional expectation 39
5.2 Martingale definitions 41
5.3 Stopping times and Doob's optional stopping theorem 42
5.4 Doob's decomposition theorem 45
5.5 An example: one-dimensional random walk 45
5.6 Doob's upcrossing inequality 47
5.7 Convergence theorems 49
5.8 Strong law of large numbers 53
5.9 Maximal inequalities 54

6 Markov chains **56**

6.1 Kernels 56
6.2 Definition and basic properties 57
6.3 Class structure 58
6.4 Hitting times and absorption probabilities 58
6.5 Strong Markov property 60

6.6 Recurrence and transience 60
6.7 Recurrence and transience of random walks 63
6.8 Invariant distributions 63
6.9 Convergence to equilibrium 66
6.10 Ergodic theorem 67

7 Poisson processes **69**

1 Finite probability spaces

Definition 1.1.

- A **finite probability space** is a finite set $\Omega = \{\omega_1, \dots, \omega_N\}$ with numbers $p_1, \dots, p_N \in [0,1]$ such that $\sum_n p_n = 1$. We will only consider finite probability spaces in this section.
- An event is a subset A of Ω.
- The set of all events is denoted 2^Ω.
- The **probability measure** on Ω corresponding to $p_1, \dots, p_N$ is the map $P : 2^\Omega \to [0,1]$ defined by $P(A) := \sum_{n:\omega_n \in A} p_n$.
- A [real] **random variable** is a function $X : \Omega \to \mathbb{R}$.
- A collection of events $\mathcal{F}$ is an **algebra** if
  1) $\Omega \in \mathcal{F}$,
  2) $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$, and
  3) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.
- A random variable X is **$\mathcal{F}$-measurable** if $\{X = x\} \in \mathcal{F}$ for any $x \in \mathbb{R}$. Here, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$.

Example 1.2.

- If $\mathcal{F} := \{\varnothing, \Omega\}$, then the $\mathcal{F}$-measurable random variables are the constants.
- If $\mathcal{F} := 2^\Omega$, then every random variable is $\mathcal{F}$-measurable.

Example 1.3. Consider the act of making two coin tosses, and let $\Omega := \{hh, ht, th, tt\}$. Algebras can represent various states of knowledge. For instance, the algebra $\mathcal{F}_1 := \{\varnothing, \{hh, ht\}, \{th, tt\}, \Omega\}$ represents [in some sense] the knowledge of one coin flip, while $\mathcal{F}_2 := 2^\Omega$ refines this algebra and represents the knowledge of both coin flips. This will be made clearer when we introduce the concept of atoms.

We list some properties of P.

- $P(\varnothing) = 0$.
- $P(\Omega) = 1$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A^c) = 1 - P(A)$.

Definition 1.4.

- We say $X = Y$ a.s. (**equal almost surely**) if $P(X = Y) = 1$.
- We say $X \overset{d}{=} Y$ (**equal in distribution**) if $P(X = x) = P(Y = x)$ for all $x \in \mathbb{R}$.

Note that almost sure equality implies equality in distribution. The following example shows that the converse is false.

Example 1.5. Let $\Omega := \{hh, ht, th, hh\}$ as before, and let $X(hh) = 2$, $X(ht) = X(th) = 0$, and $X(tt) = -2$. If $Y := -X$, then $X \overset{d}{=} Y$ but $P(X = Y) = 1/2$.

Definition 1.6. The **expected value** of a random variable X is

$$\mathrm{E}[X] := \sum_{n=1}^{N} p_n X(\omega_n).$$

Note that a random variable can be interpreted to be the element $(X(\omega_1), \dots, X(\omega_N))^T \in \mathbb{R}^n$.

The expectation has the following properties.

- **Linearity.** $\mathrm{E}[aX + Y] = a\,\mathrm{E}[X] + \mathrm{E}[Y]$ for any $c \in \mathbb{R}$ and random variables X, Y.
- **Positivity.** If $X \geq 0$ a.s., then $\mathrm{E}[X] \geq 0$.
- **Continuity.** If $X_k(\omega) \to X(\omega)$ a.s., then $\mathrm{E}[X_k] \to \mathrm{E}[X]$.

One might remark that formalisms like "almost surely" are meaningless in the discrete case, but they become relevant when we transition to the continuous case; consider taking the limit of the model of N coin flips as $N \to \infty$.

Definition 1.7. The **covariance** of two random variables X, Y is defined by

$$\mathrm{Cov}(X, Y) := \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y].$$

The **variance** of a random variable is

$$\mathrm{Var}(X) := \mathrm{Cov}(X, X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

The covariance and variance satisfy these properties.

- **Symmetry.** $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- **Bilinearity.** $\mathrm{Cov}(aX + Y, Z) = a\,\mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.
- **Shift invariance.** $\mathrm{Cov}(X + a, Y) = \mathrm{Cov}(X, Y)$.
- $\mathrm{Var}(X) \geq 0$.
- $\mathrm{Var}(X) = 0 \implies X = \mathrm{E}[X]$ a.s.
- **Cauchy-Schwarz.** $|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$, with equality if and only if $X - \mathrm{E}[X]$ is a multiple of $Y - \mathrm{E}[Y]$ a.s.

The **correlation** is a scale-invariant version of covariance, defined to be

$$\rho(X, Y) := \begin{cases} \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} & \text{if } \mathrm{Var}(X)\,\mathrm{Var}(Y) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

By Cauchy-Schwarz, the $-1 \leq \rho(X, Y) \leq 1$.

Definition 1.8. Events $A_1, \dots, A_M$ are **independent** if for any $m \in \{1, \dots, M\}$ and $1 \leq j_1 < j_2 < \cdots < j_m \leq M$, we have

$$P\left(\bigcap_{k=1}^{m} A_{j_k}\right) = \prod_{k=1}^{m} P(A_{j_k}).$$

Random variables $X_1, \dots, X_M$ are **independent** if for any $x_1, \dots, x_M \in \mathbb{R}$, the events $\{X_1 = x_1\}$, ..., $\{X_M = x_M\}$ are independent.

Proposition 1.9. *For random variables $X_1, \dots, X_M$ on a finite probability space, the following are equivalent.*

(i) *$X_1, \dots, X_M$ are independent.*

(ii) *For any $f_1, \dots, f_M : \mathbb{R} \to \mathbb{R}$,*

$$\mathrm{E}\left[\prod_{m=1}^{M} f_m(X_m)\right] = \prod_{m=1}^{M} \mathrm{E}[f_m(X_m)].$$

(iii) For any $u \in \mathbb{R}^M$, and letting $X = (X_1, \dots, X_M)$,

$$\mathrm{E}[\exp(iu^T x)] = \prod_{m=1}^{M} \mathrm{E}[\exp(iu_m X_m)].$$

Corollary 1.10. *If $X_1, \dots, X_M$ are independent random variables, then $g_1(X_1), \dots, g_M(X_M)$ are independent for any $g_1, \dots, g_M : \mathbb{R} \to \mathbb{R}$.*

Corollary 1.11. *If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. However, the converse does not hold.*

Definition 1.12. A probability measure P is **absolutely continuous** with respect to another probability measure Q if $Q(A) = 0$ implies $P(A) = 0$ for any subset A of Ω. We denote this $P \ll Q$.

We say P is **equivalent** to Q if $P \ll Q$ and $Q \ll P$. We denote this $P \sim Q$.

Definition 1.13. The **indicator function** of a set A is a function defined by

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

Lemma 1.14. *Let Z be a random variable that is nonnegative almost surely and satisfies $\mathrm{E}_P[Z] = 1$. Then*

$$Q(A) := \mathrm{E}_p[\mathbf{1}_A Z]$$

is a probability measure with elementary probabilities $q_n := p_n Z(\omega_n)$. Moreover, $Q \ll P$.

Proof. It is clear that $q_n := Q(\{\omega_n\}) = \mathrm{E}_P[\mathbf{1}_{\{\omega_n\}} Z] = p_n Z(\omega_n)$ and that $q_n \geq 0$ whenever $p_n > 0$. Moreover, $Q(\Omega) = \sum_{n=1}^{N} p_n Z(\omega_n) = \mathrm{E}_P[Z] = 1$. This shows that $q_n \leq 1$ whenever $p_n > 0$. □

Theorem 1.15 (Elementary version of the Radon-Nikodym derivative)**.** *Let P and Q be probability measures on Ω, with $Q \ll P$. Then there exists a random variable Z that is nonnegative almost surely satisfying $\mathrm{E}_P[Z] = 1$ and $Q(A) = \mathrm{E}_P[\mathbf{1}_A Z]$ for any subset A of Ω.*

Proof. Let $Z(\omega_n) := q_n / p_n$ whenever $p_n > 0$. [The values of $Z(\omega_n)$ when $p_n = 0$ are irrelevant.] □

In the following we assume $p_1, \dots, p_N > 0$.

Definition 1.16. An **atom** A of an algebra $\mathcal{F}$ is a set in $\mathcal{F} \setminus \{\varnothing\}$ such that $\varnothing$ and A are the only subsets of A in $\mathcal{F}$. In other words, A is “indivisible” in $\mathcal{F}$.

Note that for every algebra $\mathcal{F}$ there exist finitely many atoms $A_1, \dots, A_M$ such that $\Omega = \bigcup_{m=1}^{M} A_m$ and $A_i \cap A_j = \varnothing$ when $i \neq j$. $\mathcal{F}$ consists of $\varnothing$ and all unions of the A_m.

Definition 1.17. Let $X : \Omega \to \mathbb{R}$ be a random variable taking the values $x_1, \dots, x_M \in \mathbb{R}$, $M \leq N$. The algebra **generated** by X, denoted $\alpha(X)$, is the algebra with atoms $\{X = x_1\}, \dots, \{X = x_m\}$. It is the coarsest algebra on Ω with respect to which X is measurable.

This is how algebras encode knowledge or information. In $\alpha(X)$, knowledge of the value of $X(\omega)$ determines the unique atom of the algebra $\alpha(X)$ that contains ω. In finer algebras, it is not always possible to identify the atom that contains ω by only observing $X(\omega)$.

Definition 1.18.

- Let $A, B \in 2^\Omega$ and $P(B) > 0$. The **conditional probability** of A given B is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

- Let X be a random variable, and let $\mathcal{F}$ be an algebra with atoms $A_1, \ldots, A_M$. The **conditional expectation** of X with respect to $\mathcal{F}$ is the random variable defined by

$$\mathrm{E}[X \mid \mathcal{F}](\omega) := \frac{1}{P(A_m)} \sum_{\omega_j \in A_m} X(\omega_j) p_j \quad \text{where } A_m \text{ is the [unique] atom containing } \omega.$$

  Note that $\mathrm{E}[X \mid \mathcal{F}]$ is constant on each atom and is therefore $\mathcal{F}$-measurable.

- Let X and Y be random variables. We define

$$\mathrm{E}[X \mid Y] := \mathrm{E}[X \mid \alpha(Y)].$$

- Let $A \in 2^\Omega$ and let $\mathcal{F}$ be an algebra. We define

$$P(A \mid \mathcal{F}) := \mathrm{E}[\mathbf{1}_A \mid \mathcal{F}].$$

- If $A \in 2^\Omega$ and let $\mathcal{F}$ be an algebra. We define

$$P(A \mid Y) := \mathrm{E}[\mathbf{1}_A \mid Y].$$

Proposition 1.19. *Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ be a finite probability space with $p_n = P[\omega_n] > 0$ for all n, X a random variable on Ω and $\mathcal{F}$ an algebra of subsets of Ω with atoms $A_1, \ldots, A_M$.*

a) $\mathrm{E}[X \mid \mathcal{F}]$ *is $\mathcal{F}$-measurable.*

b) $\mathrm{E}[X \mid \mathcal{F}] = \mathrm{E}[X]$ *if $\mathcal{F} = \{\varnothing, \Omega\}$.*

c) $\mathrm{E}[X \mid \mathcal{F}] = X$ *if X is $\mathcal{F}$-measurable.*

d) $\mathrm{E}[X \mid \mathcal{F}] \geq 0$ *if $X \geq 0$.*

e) $\mathrm{E}[XY + Z \mid \mathcal{F}] = X\,\mathrm{E}[Y \mid \mathcal{F}] + \mathrm{E}[Z \mid \mathcal{F}]$ *if X is $\mathcal{F}$-measurable.*

f) $\mathrm{E}[\mathrm{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathrm{E}[X \mid \mathcal{G}]$ *for every sub-algebra $\mathcal{G}$ of $\mathcal{F}$.*

g) $\mathrm{E}[X \mid \mathcal{F}] = \mathrm{E}[X]$ *if X is independent of $\mathcal{F}$.*

h) $\mathrm{E}[X \mid \mathcal{F}]$ *is the unique minimizer of the quadratic optimization problem*

$$\textit{minimize } \mathrm{E}[(X - Y)^2)] \textit{ over all } \mathcal{F}\textit{-measurable random variables } Y.$$

That is, $\mathrm{E}[X \mid \mathcal{F}]$ *is the projection of X to the space of $\mathcal{F}$-measurable random variables with respect to the norm $\|X\|_2 = \mathrm{E}[X^2]^{1/2}$, or in other words,* $\mathrm{E}[X \mid \mathcal{F}]$ *is the best least-squares estimate of X given the information contained in $\mathcal{F}$.*

Definition 1.20. Let (Ω, P) and (Ω', P') be [finite] probability spaces. Then $(\Omega \times \Omega', P \otimes P')$ is a probability space, where $\Omega \times \Omega' := \{(\omega, \omega') : \omega \in \Omega, \omega' \in \Omega'\}$ is the Cartesian product and where

$$(P \otimes P')((\omega, \omega')) := P(\omega)P'(\omega').$$

Note that this construction gives "independence" to each component of the space and preserves the probabilities when embedding from the original space. More explicitly, let $A \subset \Omega$, $B \subset \Omega'$, $X : \Omega \to \mathbb{R}$, and $Y : \Omega' \to \mathbb{R}$. The random events $\widehat{A} := A \times \Omega'$ and $\widehat{B} := \Omega \times B$ are independent in $(\Omega \times \Omega', P \otimes P')$. Moreover, $(P \otimes P')(\widehat{A}) = P(A)$ and $(P \otimes P')(\widehat{B}) = P'(B)$. The random variables $\widehat{X}((\omega, \omega')) := X(\omega)$ and $\widehat{Y}((\omega, \omega')) := Y(\omega')$ are independent on $(P \otimes P')(\widehat{A}) = P(A)$. Moreover, $X \overset{d}{=} \widehat{X}$ and $Y \overset{d}{=} \widehat{Y}$.

Example 1.21. Let $\Omega := \{0, 1\}$, $P(0) := q$, $P(1) := p$, and $p + q = 1$. Then the random variable defined by $\xi(0) := 0$ and $\xi(1) := 1$ is called a **Bernoulli** random variable with parameter p.

Example 1.22. We consider the product of the previous probability space with itself n times. Let $\Omega := \{0, 1\}^n$, and let $P((a_1, \ldots, a_n)) := p^{\sum_i a_i} q^{1 - \sum_i a_i}$. If we define $\xi_i((a_1, \ldots, a_n)) := a_i$, then $\xi_1, \ldots, \xi_n$ are independent Bernoulli random variables with parameter p.

Example 1.23. Let $S := \sum_{i=1}^n \xi_i$, where ξ_i are as defined in the previous example. Then $P(S = k) = \binom{n}{k} p^k q^{n-k}$ for $k = 0, \ldots, n$. S is called a **binomial** random variable with parameters n and p. Its expectation is $\mathrm{E}[S] = \sum_{i=1}^n \mathrm{E}[\xi_i] = np$.

2 Countable probability spaces

Definition 2.1. A countable probability space consists of a countable set $\Omega = \{\omega_1, \omega_2, \ldots, \}$ with numbers $p_1, p_2, \ldots \geq 0$ such that $\sum_{n\geq 1} p_n = 1$. The probability measure P, events, and random variables can be defined analogously from the case of finite probability spaces.

Example 2.2 (Poisson distribution)**.** Let $\Omega = \{0, 1, \ldots, \}$, and let $P(n) = e^{-\lambda}\frac{\lambda^n}{n!}$, where $\lambda \geq 0$. The random variable defined by $X(n) := n$ is said to follow the **Poisson** distribution with parameter λ.

Example 2.3 (Geometric distribution)**.** Let $\Omega = \{1, 2, \ldots\}$, and let $P(n) = (1-p)^{n-1}p$. The random variable $X(n) := n$ is said to follow the **geometric** distribution with parameter p.

3 General probability spaces

3.1 σ-algebras

Definition 3.1. A system $\mathcal{F}$ of subsets of a nonempty set Ω is a **σ-algebra** if it is an algebra (see Definition 1.1) satisfying $\bigcup_{n\geq 1} A_n \in \mathcal{F}$ for every sequence $A_1, A_2, \ldots \in \mathcal{F}$.

Consequently, the intersection of a sequence of sets in $\mathcal{F}$ is also in $\mathcal{F}$.

Definition 3.2. A pair $(\Omega, \mathcal{F})$ of a nonempty set Ω and an associated σ-algebra $\mathcal{F}$ is called a **measurable space**.

Definition 3.3. A **measure** on a measurable space is a mapping $\mu : \mathcal{F} \to [0, \infty]$ (infinity is included) such that $\mu(\varnothing) = 0$ and

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n), \quad \text{for any sequence of [pairwise] disjoint sets } A_1, A_2, \ldots \in \mathcal{F}.$$

This last property is called **countable additivity** or **σ-additivity**.

Definition 3.4. Let μ be a measure on Ω. It is **probability measure** if $\mu(\Omega) = 1$. It is a **finite measure** if $\mu(\Omega) < \infty$. It is **σ-finite** if there exist a sequence $\Omega_1, \Omega_2, \ldots \in \mathcal{F}$ such that $\mu(\Omega_n) < \infty$ for all n and $\bigcup_{n\geq 1} \Omega_n = \Omega$.

Definition 3.5. A **measure space** $(\Omega, \mathcal{F}, \mu)$ is a measurable space with a measure. It is **complete** if for any $B \in \mathcal{F}$ such that $\mu(B) = 0$, any subset $A \subset B$ satisfies $A \in \mathcal{F}$ and $\mu(A) = 0$.

Every measure space $(\Omega, \mathcal{F}, \mu)$ can be **completed** by defining the σ-algebra

$$\overline{\mathcal{F}} := \{A \cup C : C \in \mathcal{F}, A \subset B, B \in \mathcal{F}, \mu(B) = 0\}$$

with the measure $\overline{\mu}(A \cup C) := \mu(C)$.

Definition 3.6. We call a set function $\mu : \mathcal{F} \to [0, \infty]$ on an algebra $\mathcal{F}$ of subsets of Ω a **finitely additive measure** if $\mu(\varnothing) = 0$ and $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint $A, B \in \mathcal{F}$.

Theorem 3.7. *Let μ be a finitely-additive measure on an algebra (not σ-algebra) $\mathcal{F}$ of subsets of a set Ω such that $\mu(\Omega) < \infty$. Then the following are equivalent.*

1) μ is σ-additive:

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n)$$

for any sequence of [pairwise] disjoint sets $A_1, A_2, \ldots \in \mathcal{F}$ such that $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.

2) μ is continuous from below:

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \lim_{n\to\infty} \mu(A_n)$$

for any increasing sequence $A_1 \subset A_2 \subset \cdots$ of sets in $\mathcal{F}$ such that $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.

3) μ is continuous from above:

$$\mu\left(\bigcap_{n\geq 1} A_n\right) = \lim_{n\to\infty} \mu(A_n)$$

for any decreasing sequence $A_1 \supset A_2 \supset \cdots$ of sets in $\mathcal{F}$ such that $\bigcap_{n\geq 1} A_n \in \mathcal{F}$.

4) μ is continuous at $\varnothing$:

$$\lim_{n\to\infty} \mu(A_n) = 0$$

for any decreasing sequence $A_1 \supset A_2 \supset \cdots$ of sets in $\mathcal{F}$ such that $\bigcap_{n\geq 1} A_n = \varnothing$.

Definition 3.8. A **monotone class** $\mathcal{M}$ is a collection of subsets of a set Ω such that

1) $\bigcup_{n\geq 1} A_n \in \mathcal{M}$ for any increasing sequence $A_1 \subset A_2 \subset \cdots$ of sets $A_n \in \mathcal{M}$, and

2) $\bigcap_{n\geq 1} A_n \in \mathcal{M}$ for any decreasing sequence $A_1 \supset A_2 \supset \cdots$ of sets $A_n \in \mathcal{M}$.

Note that any σ-algebra is also a monotone class.

Definition 3.9. A **Dynkin system** $\mathcal{D}$ is a collection of subsets of a set Ω such that

1) $\Omega \in \mathcal{D}$.

2) $\bigcup_{n\geq 1} A_n \in \mathcal{D}$ for every sequence of pairwise disjoint sets $A_1, A_2, \ldots$ in $\mathcal{D}$.

3) $A^c \in \mathcal{D}$ for every $A \in \mathcal{D}$.

Note that any σ-algebra is also a Dynkin system.

Lemma 3.10. *The following conditions are equivalent to the above three conditions.*

1') $\Omega \in \mathcal{D}$.

2') $B \setminus A \in \mathcal{D}$ for all $A, B \in \mathcal{D}$ such that $A \subset B$.

3') $\bigcup_{n\geq 1} A_n \in \mathcal{D}$ for every increasing sequence of sets $A_1 \subset A_2 \subset \cdots$ in $\mathcal{D}$.

Lemma 3.11. *The arbitrary [not necessarily countable] intersection of σ-algebras is also a σ-algebra. The same statement holds after replacing "σ-algebra" with either "Dynkin systems" or with "monotone classes."*

Definition 3.12. Let $\mathcal{E}$ be a nonempty collection of subsets of Ω. Then $\sigma(\mathcal{E})$ denotes the intersection of all σ-algebras containing $\mathcal{E}$, and is called the σ-agebra **generated** by $\mathcal{E}$. It is the "smallest" σ-algebra containing $\mathcal{E}$, in that any σ-algebra containing $\mathcal{E}$ must also contain $\sigma(\mathcal{E})$. We let $\delta(\mathcal{E})$ and $\mu(\mathcal{E})$ denote the Dynkin system **generated** by $\mathcal{E}$ and the monotone class **generated** by $\mathcal{E}$ respectively, which are defined analogously.

Lemma 3.13. *An algebra $\mathcal{A}$ is a σ-algebra if and only if it is a monotone class.*

Proof. All σ-algebras are monotone classes, so we need only show the other direction. Let $\mathcal{A}$ be an algebra that is a monotone class. We only need to show that $\mathcal{A}$ is closed under countable unions. Let $A_1, A_2, \ldots$ be a sequence of subsets in $\mathcal{A}$. The sets $B_n := \bigcup_{k=1}^n A_k$ are in $\mathcal{A}$ because $\mathcal{A}$ is an algebra. Then, using the fact that B_n form an increasing sequence, we have

$$\bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} B_n \in \mathcal{A}.$$

□

The following principle is a tautology, but is the key technique that appears multiple times in the proof of the monotone class theorem.

Lemma 3.14 (Principle of good sets)**.** *The statement "all elements of H satisfy property P" is equivalent to "$H \subset \{a : a$ satisfies $P\}$."*

Theorem 3.15 (Monotone class theorem)**.** *If $\mathcal{A}$ is an algebra, then*

$$\mu(\mathcal{A}) = \sigma(\mathcal{A}).$$

Proof. Because σ-algebras are monotone classes, we have $\mu(\mathcal{A}) \subset \sigma(\mathcal{A})$. To show the reverse inclusion, note that by Lemma 3.13, it suffices to show that $\mu(\mathcal{A})$ is an algebra, since then $\mu(\mathcal{A})$ would be a σ-algebra containing $\mathcal{A}$, and thus must contain $\sigma(\mathcal{A})$ by definition.

1) $\Omega \in \mathcal{A} \subset \mu(A)$ because $\mathcal{A}$ is an algebra.

2) Fix $S \in \mu(\mathcal{A})$. We would like to show that $S^c \in \mu(\mathcal{A})$. It suffices to show that

$$\mu(\mathcal{A}) \subset B := \{S \subset \Omega : S^c \in \mu(\mathcal{A})\}.$$

We claim B is a monotone class containing $\mathcal{A}$ as a subset.

- Because $\mathcal{A}$ is an algebra, $S \in \mathcal{A}$ implies $S^c \in \mathcal{A}$, so $\mathcal{A} \subset B$.
- We now show B is a monotone class. Suppose $A_1, A_2, \ldots$ is an increasing sequence of sets in B. To show that $\bigcup_{n\geq 1} A_n$ is also in B, note that because $A_n^c \in \mu(\mathcal{A})$ (by definition of B) and because $\mu(\mathcal{A})$ is a monotone class, we know

$$\left(\bigcup_{n\geq 1} A_n\right) = \bigcap_{n\geq 1} A_n^c$$

is in $\mu(\mathcal{A})$. A similar argument shows why $\bigcap_{n\geq 1} A_n$ is in B if $A_1 \supset A_2 \supset \cdots$ in B.

By definition of $\mu(\mathcal{A})$ being the "smallest" monotone class containing $\mathcal{A}$, we must have $\mu(\mathcal{A}) \subset B$ as desired.

3) Fix $S \in \mu(\mathcal{A})$ we would like to show that for any $T \in \mu(\mathcal{A})$, we have $S \cup T \in \mu(\mathcal{A})$. Define

$$N_S := \{T \subset \Omega : S \cup T \in \mu(\mathcal{A})\}.$$

We would like to show $\mu(\mathcal{A}) \subset N_S$. We claim N_S is a monotone class containing $\mathcal{A}$.

- We show N_S is a monotone class. If $A_1, A_2, \ldots$ is an increasing sequence of sets in N_S, then $(S \cup A_n)_{n\geq 1}$ is an increasing sequence of sets in $\mu(\mathcal{A})$. Since $\mu(\mathcal{A})$ is a monotone class, $\bigcup_{n\geq 1}(S \cup A_n) = S \cup \bigcup_{n\geq 1} A_n$ is in $\mu(\mathcal{A})$, implying $\bigcup_{n\geq 1} A_n$ is in N_S. A similar argument shows that the intersection of a decreasing sequence of sets in N_S is also in N_S.
- We show $\mathcal{A} \subset N_S$. This will be implied if we show the stronger statement that $U \cup V \in \mu(\mathcal{A})$ for any $U, V \in \mathcal{A}$. Fix $U \in \mathcal{A}$ and let $N_U := \{V \subset \Omega : U \cup V \in \mu(\mathcal{A})\}$.[1] We would like to show $\mu(\mathcal{A}) \subset N_U$. We claim N_U is a monotone class containing $\mathcal{A}$.
  - $\mathcal{A} \subset N_U$ is clear because $\mathcal{A}$ is an algebra.
  - N_U is a monotone class by the same argument we used for N_S above.

  By the definition of $\mu(\mathcal{A})$ being the "smallest" monotone class containing $\mathcal{A}$, we have $\mu(\mathcal{A}) \subset N_U$, and thus $\mathcal{A} \subset N_U$ as desired.

By the definition of $\mu(\mathcal{A})$ being the "smallest" monotone class containing $\mathcal{A}$, we have $\mu(\mathcal{A}) \subset N_S$, proving that $S \cup T \in \mu(\mathcal{A})$ for any $T \in \mu(\mathcal{A})$.

[1] This is the identical to the definition of N_S, but here $U \in \mathcal{A}$, which is in some sense an improvement over $S \in \mu(\mathcal{A})$.

To summarize, we have shown that the monotone class $\mu(\mathcal{A})$ is an algebra, so by Lemma 3.13, it is a σ-algebra containing $\mathcal{A}$, and thus contains $\sigma(\mathcal{A})$. □

Lemma 3.16. *A Dynkin system $\mathcal{D}$ that is closed under finite intersection is a σ-algebra.*

Proof.

1) $\Omega \in \mathcal{D}$ because $\mathcal{D}$ is a Dynkin system.

2) If $A \in \mathcal{D}$, then $A^c \in \mathcal{D}$ because $\mathcal{D}$ is a Dynkin system.

3) Let $A_1, A_2, \dots$ be a sequence of sets in $\mathcal{D}$. Then $A_n^c \in \mathcal{D}$ for all n (definition of Dynkin system) and the sets $B_n := A_n \cap \bigcap_{k=1}^{n-1} A_k^c$ are in $\mathcal{D}$ as well (closure under finite intersection). Since the B_n are disjoint, we have
$$\bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} B_n \in \mathcal{D}.$$

□

The following theorem is a version of the monotone class theorem for Dynkin systems, and the proof is identical in spirit to the previous one.

Theorem 3.17 (Variant of the monotone class theorem)**.** *If $\mathcal{E}$ is a nonempty collection of subsets of Ω that is closed under finite intersection, then*
$$\delta(\mathcal{E}) = \sigma(\mathcal{E}).$$

Proof. Because all σ-algebras are Dynkin systems, we have $\delta(\mathcal{E}) \subset \sigma(\mathcal{E})$ because $\delta(\mathcal{E})$ is contained in any Dynkin system containing $\mathcal{E}$. To show the reverse inclusion, note that by Lemma 3.16, it suffices to show that $\delta(\mathcal{E})$ is closed under finite intersection, since then $\delta(\mathcal{E})$ would be a σ-algebra containing $\mathcal{E}$, and thus must contain $\sigma(\mathcal{E})$ by definition.

Fix $A \in \delta(\mathcal{E})$. We would like to show that $A \cap B \in \delta(\mathcal{E})$ for any $B \in \delta(\mathcal{E})$. It suffices to show that
$$\delta(\mathcal{E}) \subset N_A := \{B \subset \Omega : A \cap B \in \delta(\mathcal{E})\}.$$

- We claim N_A is a Dynkin system.

  1) $\Omega \in N_A$ because $A \cap \Omega = A \in \delta(\mathcal{E})$.
  2) Given a sequence $A_1, A_2, \dots$ of disjoint sets in N_A, we claim $\bigcup_{n\geq 1} A_n$ is also in N_A. Indeed, since $(A \cap A_n)_{n\geq 1}$ is a sequence of disjoint sets in the Dynkin system $\delta(\mathcal{E})$, we have
  $$A \cap \bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} (A \cap A_n) \in \mathcal{D}.$$
  3) If $B \in N_A$, then $A \cap B \in \delta(\mathcal{E})$. Thus,
  $$A \cap B^c = (A^c \cup B)^c = (A^c \cup (A \cap B))^c \in \delta(\mathcal{E})$$
  because A^c and $A \cap B$ are disjoint sets in the Dynkin system $\delta(\mathcal{E})$. So, $B^c \in N_A$ as well.

- We also claim that $\mathcal{E} \subset N_A$. This will be implied if we show the stronger statement that $U \cap V \in \delta(\mathcal{E})$ for any $U, V \in \mathcal{E}$. Fix $U \in \mathcal{E}$, and let $N_U := \{V \subset \Omega : U \cap V \in \delta(\mathcal{E})\}$. We would like to show $\delta(\mathcal{E}) \subset N_U$. We claim N_U is a Dynkin system containing $\mathcal{E}$.

  - $\mathcal{E} \subset N_U$ is clear because $\mathcal{E}$ is closed under finite intersection by assumption.
  - N_U is a Dynkin system by the same argument used for N_A above.

  Since N_U is a Dynkin system containing $\mathcal{E}$, so by definition $\delta(\mathcal{E}) \subset N_U$.

We have shown that N_A is a Dynkin system containing $\mathcal{E}$, so by definition $\delta(\mathcal{E}) \subset N_A$, i.e., $\delta(\mathcal{E})$ is closed under finite intersection. By Lemma 3.16, $\delta(\mathcal{E})$ is a σ-algebra containing $\mathcal{E}$, so $\sigma(\mathcal{E}) \subset \delta(\mathcal{E})$ by definition. □

Corollary 3.18. *Let P, Q be probability measures on $(\Omega, \sigma(\mathcal{E}))$ where $\mathcal{E}$ is a collection of subsets of Ω that is closed under finite intersection. If $P = Q$ on $\mathcal{E}$, then $P = Q$ on $\sigma(\mathcal{E})$.*

Proof. Let $\mathcal{D} := \{A \in \sigma(\mathcal{E}) : P(A) = Q(A)\}$. By assumption, $\mathcal{E} \subset \mathcal{D}$. Moreover, $\mathcal{D}$ is a Dynkin system, due to the definition of a probability measure. So, $\delta(\mathcal{E}) \subset \mathcal{D}$ by definition. By Theorem 3.17, $\delta(\mathcal{E}) = \sigma(\mathcal{E})$, so we have $\sigma(\mathcal{E}) \subset \mathcal{D}$, that is, $P(A) = Q(A)$ for any $A \in \sigma(\mathcal{E})$. □

Definition 3.19. The **Borel σ-algebra** on a topological space is the σ-algebra generated by the collection of open sets.

Lemma 3.20. *The σ-algebras on $\mathbb{R}$ generated by the following collections are the same: the Borel σ-algebra on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$.*

1) The open sets in $\mathbb{R}$.

2) The half-open intervals $(a, b]$ where $a < b$, $a, b \in \mathbb{R}$.

3) The intervals $(-\infty, x]$ for $x \in \mathbb{R}$.

Proof. Let $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ be the respective σ-algebras generated by the three collections. To show containment, it suffices to show that one σ-algebra contains the generators of another. We remark that any open set in $\mathbb{R}$ can be written as the countable union of disjoint open intervals.

- $\mathcal{B}_3 \subset \mathcal{B}_2$ because $(-\infty, x] = \bigcup_{n \geq 1} (x - n, x]$.
- $\mathcal{B}_2 \subset \mathcal{B}_3$ because $(a, b] = (-\infty, b] \cap (-\infty, a]^c$.
- $\mathcal{B}_1 \subset \mathcal{B}_2$ because $(a, b) = \bigcup_{n \geq 1} (a, b - 1/n]$.
- $\mathcal{B}_2 \subset \mathcal{B}_1$ because $(a, b] = \bigcap_{n \geq 1} (a, b + 1/n)$.

□

Corollary 3.21. *A probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is uniquely determined by its **cumulative density function (cdf)** defined by*

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

Proof. Suppose P and Q have the same cdf. Let $\mathcal{E} := \{(-\infty, x] : x \in \mathbb{R}\}$ and note that it is closed under finite intersection. Since P and Q agree on $\mathcal{E}$, they also agree on $\sigma(\mathcal{E})$ by Corollary 3.18. To conclude, note that Lemma 3.20 implies $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$. □

Proposition 3.22. *The cdf F of a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies the following properties.*

1) F is nondecreasing.

2) F is right continuous, that is,

$$\lim_{x \searrow x_0} F(x) = F(x_0).$$

3)

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

Proof. The first and third properties are clear from the definition of a probability measure. For the second, suppose we have a sequence $(x_n)_{n \geq 1}$ that converges monotonically to x from above. Then

$$\lim_{n \to \infty} F(x_n) - F(x) = \lim_{n \to \infty} P((-\infty, x_n]) - P((-\infty, x]) = \lim_{n \to \infty} P((x, x_n]) = 0.$$

□

3.2 Measurable functions and random variables

Definition 3.23. A function $f : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ is **measurable** if $f^{-1}(A)$ is in $\mathcal{F}$ for any $A \in \mathcal{E}$.

Lemma 3.24. *Fix $f : \Omega \to E$.*

- *If $\mathcal{E}$ is a σ-algebra on E, then $\sigma(f) := \{f^{-1}(A) : A \in \mathcal{E}\}$ is the smallest σ-algebra on Ω such that $f : (\Omega, \sigma(f)) \to (E, \mathcal{E})$ is measurable.*
- *If $\mathcal{F}$ is a σ-algebra on Ω, then $\widehat{\sigma}(f) := \{A \subset E : f^{-1}(A) \in \mathcal{F}\}$ is the finest (largest) σ-algebra on E such that $f(\Omega, \mathcal{F}) \to (E, \widehat{\sigma}(f))$ is measurable.*

Proof. The fact that $\sigma(f)$ is a σ-algebra follows easily because the pre-image transfers the properties of the σ-algebra $\mathcal{E}$.

1) $\sigma(f)$ contains Ω because $\Omega = f^{-1}(E)$ and $E \in \mathcal{E}$.

2) If $B \in \sigma(f)$, then $B = f^{-1}(A)$ for some $A \in \mathcal{E}$. Then $B^c = f^{-1}(A^c)$ is also in $\sigma(f)$.

3) If $B_1, B_2, \dots$ is a sequence of sets in $\sigma(f)$, then there exists a sequence of sets $A_1, A_2, \dots$ in $\mathcal{E}$ such that $B_n = f^{-1}(A_n)$ for all n. Then
$$\bigcup_{n\geq 1} B_n = \bigcup_{n\geq 1} f^{-1}(A_n) = f^{-1}\left(\bigcup_{n\geq 1} A_n\right)$$
is in $\sigma(f)$.

By definition of measurability, any σ-algebra for Ω that makes f measurable must contain $\sigma(f)$.

Verifying that $\widehat{\sigma}(f)$ is a σ-algebra is also simple.

1) $\widehat{\sigma}(f)$ contains E because $f^{-1}(E) = \Omega$.

2) If A is contained in $\widehat{\sigma}(f)$, then $f^{-1}(A) \in \mathcal{F}$. Thus, A^c is also contained, since $f^{-1}(A^c) = (f^{-1}(A))^c \in \mathcal{F}$.

3) Let $A_1, A_2, \dots$ be a sequence of sets in $\widehat{\sigma}(f)$, which implies the sets $f^{-1}(A_n)$ form a sequence of sets in $\mathcal{F}$. Thus,
$$f^{-1}\left(\bigcup_{n\geq 1} A_n\right) = \bigcup_{n\geq 1} f^{-1}(A_n)$$
is contained in $\widehat{\sigma}(f)$.

By definition of measurability, any σ-algebra for E that makes f measurable must be contained in $\widehat{\sigma}(f)$. □

Lemma 3.25. *Let $f : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ and $g : (E, \mathcal{E}) \to (G, \mathcal{G})$ both be measurable. Then $g \circ f$ is measurable.*

Lemma 3.26. *Let $f : \Omega \to E$ be a function and let $\mathcal{E}$ be a collection of subsets of E (not necessarily an algebra). Then*
$$\sigma(\{f^{-1}(A) : A \in \mathcal{E}\}) = \{f^{-1}(A) : A \in \sigma(\mathcal{E})\}.$$

Proof. Let $\mathcal{F}$ (and $\mathcal{G}$) be the left-hand (and right-hand) side of the above equality. To show $\mathcal{F} \subset \mathcal{G}$, it suffices to show that $\mathcal{G}$ is a σ-algebra containing the generator $\{f^{-1}(A) : A \in \mathcal{E}\}$ of $\mathcal{F}$. The containment is clear because $\mathcal{E} \subset \sigma(\mathcal{E})$, and the fact that $\mathcal{G}$ is a σ-algebra follows from the proof of Lemma 3.24.

To show the reverse containment let $\mathcal{H} := \{A \subset E : f^{-1}(A) \in \mathcal{F}\}$. We claim it is a σ-algebra containing $\mathcal{E}$. Indeed, the proof of Lemma 3.24 shows that it is a σ-algebra, and the containment of $\mathcal{E}$ follows from the definition of $\mathcal{F}$. Thus, $\mathcal{H} \supset \sigma(\mathcal{E})$, which in turn shows $\mathcal{G} \subset \mathcal{F}$. □

Definition 3.27. A **random variable** is a measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Corollary 3.28. *A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable if $\{X \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.*

Proof. Note that $\{X \leq x\} = X^{-1}((-\infty, x])$. Let $\mathcal{E} := \{(-\infty, x] : x \in \mathbb{R}\}$, and recall that $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$ by Lemma 3.20. By assumption, $\sigma(X^{-1}(A) : A \in \mathcal{E}\}) \subset \mathcal{F}$, but by Lemma 3.26, this implies $\{f^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\} \subset \mathcal{F}$, i.e., X is measurable. □

We remark that we can replace the "$\leq$" of $\{X \leq x\}$ in the above corollary with any one of $<$, $>$, or $\geq$, and the result will still hold. These correspond to other collections that generate the Borel σ-algebra on $\mathbb{R}$.

Proposition 3.29. *Let $X, Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be random variables. Then $X + Y$, $X - Y$, and XY are random variables. If $Y(\omega) \neq 0$ for all $\omega \in \Omega$, then X/Y is a random variable as well.*

Proof. To show $X + Y$ is measurable, note that

$$\{X + Y \geq x\} = \bigcup_{q \in \mathbb{Q}} (\{X > q\} \cap \{Y > x - q\}.$$

The other cases can be shown similarly, although it is a bit tedious. □

Definition 3.30. An **extended random variable** is a function $X(\Omega, \mathcal{F}) \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ such that $\{X \leq x\} \in \mathcal{F}$ for any $x \in \overline{\mathbb{R}}$.

Proposition 3.31. *If $(X_n)_{n\geq 1}$ is a sequence of extended random variables on $(\Omega, \mathcal{F})$, then $\sup_n X_n$ and $\inf_n X_n$ are extended random variables.*

Proof. This follows if we note $\{\sup_n X_n \leq x\} = \bigcap_{n\geq 1}\{X_n \leq x\}$ and $\{\inf_n X_n \geq x\} = \bigcap_{n\geq 1}\{X_n \geq x\}$. □

Lemma 3.32. *If $(X_n)_{n\geq 1}$ is a monotonically increasing or decreasing sequence of extended random variables, then their limit X is an extended random variable.*

Proof. If the sequence is increasing, then note that $\{X \leq x\} = \bigcap_{n\geq 1}\{X_n \leq x\}$. If the sequence is decreasing, then note that $\{X \geq x\} = \bigcap_{n\geq 1}\{X_n \geq x\}$. □

Proposition 3.33. *If $(X_n)_{n\geq 1}$ is a sequence of extended random variables, then $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are extended random variables.*

Proof. Note that $\limsup_{n\to\infty} X_n = \lim_{n\to\infty} \sup_{k\geq n} X_k$ and $\liminf_{n\to\infty} X_n = \lim_{n\to\infty} \inf_{k\geq n} X_k$. Since $(\sup_{k\geq n} X_k)_{n\geq 1}$ and $(\inf_{k\geq n} X_k)_{n\geq 1}$ are monotonic sequences of extended random variables (Proposition 3.31), their limits are also extended random variables by Lemma 3.32. [Alternatively, we could have noted that $\limsup_{n\to\infty} X_n = \inf_{n\geq 1} \sup_{k\geq n} X_k$ and $\liminf_{n\to\infty} X_n = \sup_{n\geq 1} \inf_{k\geq n} X_k$.] □

Corollary 3.34. *If $(X_n)_{n\geq 1}$ is a sequence of extended random variables such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$, then X is an extended random variable.*

3.3 Extension theorems

Definition 3.35. A collection $\mathcal{S}$ of subsets of a nonempty set Ω is a **semiring** if

1) $\varnothing \in \mathcal{S}$,

2) $A \cap B \in \mathcal{S}$ if $A, B \in \mathcal{S}$, and

3) for any $A, B \in \mathcal{S}$, there exist finitely many pairwise disjoint sets $C_1, \ldots, C_n \in \mathcal{S}$ such that $A \setminus B = C_1 \cup \cdots \cup C_n$.

Definition 3.36. A **measure** on a semiring $\mathcal{S}$ is a function $\mu : \mathcal{S} \to [0, \infty]$ (includes ∞) such that

1) $\mu(\varnothing) = 0$,

2) for any sequence $A_1, A_2, \ldots$ of pairwise disjoint sets in $\mathcal{S}$ such that $\bigcup_{n\geq 1} A_n \in \mathcal{S}$, we have

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n).$$

A **finitely additive measure** is a function μ that satisfies the above properties, but with the countable collection in 2) replaced by a finite collection.

Definition 3.37. A measure μ is **σ-finite** if there exists a sequence $\Omega_1, \Omega_2, \dots$ of sets in $\mathcal{S}$ such that $\mu(\Omega_n) < \infty$ for all n and $\Omega = \bigcup_{n \geq 1} \Omega_n$.

The following lemma is a useful tool to check if the conditions for the Carathéodory Extension Theorem hold.

Lemma 3.38. *Let $\mathcal{S}$ be a semiring on Ω that contains Ω. A finitely additive measure μ on $\mathcal{S}$ such that $\mu(\Omega) < \infty$ is σ-additive if and only if $\mu(A_n) \to 0$ holds for any decreasing sequence $A_1 \supset A_2 \supset \cdots$ of sets in $\mathcal{S}$ such that $\bigcap_{n \geq 1} A_n = \varnothing$.*

Theorem 3.39 (Carathéodory Extension Theorem)**.** *A measure μ on a semiring $\mathcal{S}$ can be extended to a measure on $\sigma(\mathcal{S})$. If μ is σ-finite, then the extension is unique.*

Example 3.40. Let $\mathcal{S} = \{(a, b] \cap \mathbb{R} : -\infty \leq a, b \leq \infty\}$. (Note that a and b can be infinite. Also, $\mathbb{R} \in \mathcal{S}$.) This is a semiring on $\mathbb{R}$. Every nondecreasing right-continuous (gives continuity at $\varnothing$) function $F : \mathbb{R} \to \mathbb{R}$ induces a σ-additive measure μ_F on $\mathcal{S}$ given by $\mu_F((a, b]) := F(b) - F(a)$. [If a, b are infinite, take the limit.] Then μ_F has a unique extension to $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{S})$.

If $F(x) := x$, then μ_F is the Lebesgue measure.

If $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, then μ_F is a probability measure on the real line, and F is its cdf.

Example 3.41. The family of hypercubes $(a, b] := (a_1, b_1] \times \cdots (a_n, b_n] \cap \mathbb{R}^n$ forms a semiring on $\mathbb{R}^n$. Let $F : \mathbb{R}^n \to [0, 1]$ satisfy the following.

1) $\Delta_{a_1, b_1} \cdots \Delta_{a_n, b_n} F(x_1, \dots, x_n) \geq 0$ for all $a_1 \leq b_1, \dots, a_n \leq b_n$, where the operator Δ_{a_i, b_i} maps
$$F(x_1, \dots, x_n) \mapsto F(x_1, \dots, x_{i-1}, b_i, x_{i+1}, \dots, x_n) - F(x_1, \dots, x_{i-1}, a_i, x_{i+1}, \dots, x_n).$$
[This is the analogue of nondecreasing in higher dimensions and ensures that areas will have have non-negative measure in the induced measure.]

2) $F(x^{(k)}) \to F(x)$ if $x_i^{(k)} \searrow x_i$ for all $i = 1, \dots, n$. [This is the analogue of right-continuity.]

3) $F(x) \to 1$ as $x_1 \to \infty, \dots, x_n \to \infty$.

4) $F(x) \to 0$ as $x_1 \searrow y_1, \dots, x_n \searrow y_n$ for any $y_1, \dots, y_n$ such that $y_i = -\infty$ for at least one i.

Then $P(a, b) = \Delta_{a_1, b_1} \cdots \Delta_{a_n, b_n} F(x_1, \dots, x_n)$ is a probability measure on $\mathcal{S}$ with a unique extension to $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{S})$.

An example of such a function F is $F(x) := F_1(x_1) \cdots F_n(x_n)$, a product of one-dimensional cdfs.

Definition 3.42. If $\mathcal{F}$ and $\mathcal{F}'$ are σ-algebras on Ω and Ω' respectively, then $\mathcal{S} := \{A_1 \times A_2 : A_1 \in \mathcal{F}, A_2 \in \mathcal{F}'\}$ is a semiring on $\Omega \times \Omega'$. We define their **tensor product** by
$$\mathcal{F} \otimes \mathcal{F}' := \sigma(\mathcal{S}).$$

Example 3.43. Let $(\Omega, \mathcal{F}, P)$ and $(\Omega', \mathcal{F}', P')$ be two probability spaces. The probability measure $(P \otimes P')(A_1 \times A_2) := P(A_1) P'(A_2)$ uniquely extends to a measure on $\mathcal{F} \otimes \mathcal{F}' = \sigma(S)$.

The definition of tensor product allows us to define the Borel σ-algebra on $\mathbb{R}^n$, which we denote by $\mathcal{B}(\mathbb{R})^{\otimes n}$ or $\mathcal{B}(\mathbb{R}^n)$. We would like to generalize from the finite exponent n to arbitrary exponents.

Definition 3.44. Let I be an arbitrary nonempty set (possibly uncountable); we will use it as an index set. We define $\mathbb{R}^I := \{(\omega_i)_{i \in I} : \omega_i \in \mathbb{R}\}$. Equivalently, we can view the elements of this space as functions rather than I-tuples, that is, $\mathbb{R}^I := \{f : I \to \mathbb{R}\}$. For each $i \in I$ we let $q_i : \mathbb{R}^I \to \mathbb{R}$ be the projection defined by $q_i(f) := f(i)$.

The product σ-algebra $\mathcal{B}(\mathbb{R})^{\otimes I}$ defined to be the smallest σ-algebra on $\mathbb{R}^I$ such that every projection q_i is measurable. That is,

$$\mathcal{B}(\mathbb{R})^I = \sigma(\{q_i^{-1}(B) : i \in I, B \in \mathcal{B}(\mathbb{R})\}).$$

[This is analogous to the construction of the so-called product topology, which is the coarsest topology for which the projections are continuous.]

Definition 3.45. Let I be a nonempty set. Given any finite tuple $(i_1, \dots, i_n)$ with entries in I, let $P^{i_1,\dots,i_n}$ be a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. The family of all such measures is **consistent** if the following hold.

1) **Permutation invariance.** For any finite I-tuple $(i_1, \dots, i_n)$, any permutation $\pi \in S_n$, and any subset $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$, we have

$$P^{i_1,\dots,i_n}(A_1 \times \cdots A_n) = P^{i_{\pi(1)},\cdots,i_{\pi(n)}}(A_{\pi(i)} \times \cdots \times A_{\pi(n)}).$$

2) **Projection invariance.** For any I-tuple $(i_1, \dots, i_n)$ with $n \geq 2$ and any subsets $A_1, \dots, A_{n-1} \in \mathcal{B}(\mathbb{R})$, we have

$$P^{i_1,\dots,i_{n-1}}(A_1 \times \cdots \times A_{n-1}) = P^{i_1,\dots,i_n}(A_1 \times \cdots A_{n-1} \times \mathbb{R}).$$

Theorem 3.46 (Kolmogorov extension theorem)**.** *Let $\{P^{i_1,\dots,i_n} :$ finite I-tuples $(i_1, \dots, i_n)\}$ be a consistent family of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then there exists a unique probability measure P on $(\mathbb{R}^I, \mathcal{B}(\mathbb{R})^{\otimes i})$ such that for any finite I-tuple $(i_1, \dots, i_n)$ and subset $B \in \mathcal{B}(\mathbb{R})^{\otimes n}$, the probability of the cylinder defined by B coincides with the marginal $P^{i_1,\dots,i_n}$. Explicitly,*

$$P(\{\omega \in \mathbb{R}^I : (\omega_{i_1}, \dots, \omega_{i_n}) \in B\}) = P^{i_1,\dots,i_n}(B).$$

This theorem is relevant in the study of stochastic processes: I represents the time space, and the i_n represent fixed times.

Corollary 3.47. *Let $P_1, P_2, \dots$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $(\mathbb{R}^2, \mathcal{B}(\mathbb{R})^{\otimes 2})$, and so on, such that $P_{n+1}(B \times \mathbb{R}) = P_n(B)$ for any $B \in \mathcal{B}(\mathbb{R})$. Then there exists a unique probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{N}})$ such that*

$$P(\{\omega \in \mathbb{R}^{\mathbb{N}} : (\omega_1, \dots, \omega_n) \in B\}) = P_n(B)$$

for any $n \geq 1$ and any subset $B \in \mathcal{B}(\mathbb{R})^{\otimes n}$.

Proof. It suffices to form the relevant consistent family and apply Theorem 3.46. Given an I-tuple $(i_1, \dots, i_n)$, we define

$$P^{i_1,\dots,i_n}(A_{i_1} \times \cdots \times A_{i_n}) := P_{i^*}(\{\omega \in \mathbb{R}^{i^*} : (\omega_{i_1}, \dots, \omega_{i_n}) \in A_{i_1} \times \cdots \times A_{i_n}\}),$$

where $i^* := \max\{i_1, \dots, i_n\}$ and $A_{i_1}, \dots, A_{i_n} \in \mathcal{B}(\mathbb{R})$. This family is permutation invariant; the projection invariance follows as a result of the assumption $P_{n+1}(B \times \mathbb{R}) = P_n(B)$. □

Note that to specify a measure on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{N}})$, it is not sufficient to specify the one-dimensional distributions for each component; there needs to be a specification of how the components interact. For example, specifying the one-dimensional distributions and also specifying that the components are independent would suffice.

Example 3.48 (Sequence of coin flips)**.** There exists a unique probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{N}})$ such that

$$P(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_1 = j_1, \dots, \omega_n = j_n\}) = 2^{-n}$$

for any $n \in \mathbb{N}$ and any $(j_1, \dots, j_n) \in \{0, 1\}^n$. Then the random variables $\xi_n(\omega) := \omega_n$ are independent Bernoulli random variables. Note that identifying the ξ_n with coefficients of a dyadic expansion $\sum_{n=1}^{\infty} \xi_n 2^{-n}$ shows that this models the uniform distribution on $[0, 1]$.

4 The Lebesgue integral and expectation

4.1 The Lebesgue integral and convergence theorems

The Riemann integral is defined for functions with countably many points of discontinuity, but cannot handle functions like $\mathbf{1}_{\mathbb{Q}\cap[0,1]}$.

Definition 4.1. A simple function on a measurable space $(\Omega, \mathcal{F})$ is of the form

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

for $a_i \in \mathbb{R}$, $A_i \in \mathcal{F}$, and $n \in \mathbb{N}$.

Note that the form (1) is not necessarily unique for a given function f.

Definition 4.2. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function.

1) If $f = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$ with $a_i \in \mathbb{R}_+$ (nonnegative), $A_i \in \mathcal{F}$, and $n \in \mathbb{N}$, then we define

$$\int f \, d\mu := \sum_{i=1}^{n} a_i \mu(A_i),$$

which will take a value in $[0, \infty]$. [Note that this definition is independent of the choice of representation (1) for f.]

2) If f is nonnegative, then we define

$$\int f \, d\mu := \sup_{\substack{g \text{ simple} \\ 0 \le g \le f}} \int g \, d\mu,$$

which will take a value in $[0, \infty]$.

3) For any other [measurable] f, we define $f^+ := f \vee 0$ and $f^- := (-f) \vee 0$. These two auxiliary functions are measurable and nonnegative. If either $\int f^+ \, d\mu < \infty$ or $\int f^- \, d\mu < \infty$, then we define

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

which will take a value in $[-\infty, \infty]$.

Definition 4.3.

- If $\int f^+ < \infty$ or $\int f^- < \infty$, then we say the **integral of f exists**.
- We say f is **integrable** if any of the following equivalent conditions hold.
  - ○ $\int f^+ < \infty$ and $\int f^- < \infty$
  - ○ $\int |f| < \infty$
- If $A \in \mathcal{F}$, we define $\int_A f \, d\mu := \int \mathbf{1}_A f \, d\mu$.

Proposition 4.4. *If the integrals of f and g exist, then the following hold.*

- $\mu(f \neq g) = 0 \implies \int f \, d\mu = \int g \, d\mu$.
- $\mu(f < g) = 0 \implies \int f \, d\mu \ge \int g \, d\mu$.
- $\int |f| \, d\mu = 0 \implies \mu(f \neq 0) = 0$.

Proposition 4.5. *Let f and g be simple integrable functions. Then $\int f + g = \int f + \int g$.*

Lemma 4.6. *Let f be a measurable nonnegative function. Then there exists a sequence of nonnegative simple measurable functions f_n that are pointwise increasing to f.*

Proof. Let

$$f_n := n\mathbf{1}_{\{f\geq n\}} + \sum_{i=1}^{n2^n-1} \frac{i}{2^n}\mathbf{1}_{\{\frac{i}{2^n}\leq f<\frac{i+1}{2^n}\}}.$$

□

Lemma 4.7. *Let g be a simple function and f a measurable function such that $0 \leq g \leq f$. If $(g_n)_{n\geq 1}$ is a sequence of simple nonnegative functions increasing pointwise to f, then*

$$\lim_{n\to\infty} \int g_n \, d\mu \geq \int g \, d\mu \, .$$

Proof. Let $x_1, \ldots, x_m \in \mathbb{R}_+ \setminus \{0\}$ be the values g takes, excluding zero. Then $g = \sum_{i=1}^m x_i \mathbf{1}_{\{g=x_i\}}$.

Case 1. Suppose $\mu(g = x_i) = \infty$ for some i. Let $A_n := \{g = x_i\} \cap \{g_n \geq x_i/2\}$. Then the A_n form an increasing sequence of sets whose union is $\{g = x_i\}$. By the σ-additivity of μ, applying Theorem 3.7 shows that $\mu(A_n) \to \mu(g = x_i) = \infty$ as $n \to \infty$. Then,

$$\int g_n \, d\mu \geq \frac{x_i}{2}\mu(A_n) \to \infty$$

as $n \to \infty$, proving the lemma.

Case 2. Otherwise, $\mu(g = x_i) < \infty$ for all $i = 1, \ldots, m$. For a fixed i, choose ϵ such that $0 < \epsilon < x_i$. Let $A_n := \{g = x_i\} \cap \{g_n \geq x_i - \epsilon\}$; again, this is an increasing sequence of sets whose union is $\{g = x_i\}$. By σ-additivity, we have $\mu(A_n) \nearrow \mu(g = x_i)$ as $n \to \infty$. Then,

$$\int_{\{g=x_i\}} g_n \, d\mu \geq (x_i - \epsilon)\mu(A_n) \nearrow (x_i - \epsilon)\mu(g = x_i)$$

as $n \to \infty$. Summing this result over $i = 1, \ldots, m$ gives

$$\lim_{n\to\infty} \int g_n \, d\mu \geq \int g \, d\mu \, .$$

□

Theorem 4.8 (Poor man's Beppo Levi's Monotone Convergence Theorem)**.** *Let $f \geq 0$ be measurable and let $(f_n)_{n\geq 1}$ be a sequence of nonnegative simple functions that increases to f. Then*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu \, .$$

Proof. One direction is clear by the definition of the integral for nonnegative functions.

$$\int f_n \, d\mu \leq \sup_{\substack{g \text{ simple} \\ 0\leq g\leq f}} \int g \, d\mu = \int f \, d\mu \, .$$

We consider the other direction.

Case 1. $\int f \, d\mu < \infty$. For any $\epsilon > 0$, there exists a simple function h such that $0 \leq h \leq f$ and $\int h \, d\mu \geq \int f \, d\mu - \epsilon$. By the previous lemma, $\lim_{n\to\infty} \int f_n \, d\mu \geq \int h \, d\mu$, so $\lim_{n\to\infty} \int f_n \, d\mu \geq \int f \, d\mu$.

Case 2. $\int f \, d\mu = \infty$. For any $k \in \mathbb{N}$, there exists a simple function h such that $0 \leq h \leq f$ and $\int h \, d\mu \geq k$. Again, using the previous lemma, $\lim_{n\to\infty} \int f_n \, d\mu \geq \int h \, d\mu \geq k$, so $\lim_{n\to\infty} \int f_n \, d\mu = \infty$.

□

Lemma 4.9. *Let $f, g \geq 0$ be measurable. Then $\int f + g \, d\mu = \int f \, d\mu + \int g \, d\mu$.*

Proof. Let $(f_n)_{n\geq 1}$ and $(g_n)_{n \geq 1}$ be increasing sequences of nonnegative simple functions that increase to f and g respectively. Then $f_n + g_n \nearrow f + g$. Thus,

$$\int f + g \, d\mu = \lim_{n\to\infty} \int f_n + g_n \, d\mu = \lim_{n\to\infty} f_n \, d\mu + \lim_{n\to\infty} \int g_n \, d\mu = \int f \, d\mu + \int g \, d\mu \, .$$

□

Lemma 4.10. *Let f be integrable and let g be measurable and such that either $\int g^+ \, d\mu < \infty$ or $\int g^- \, d\mu < \infty$ holds. Then*

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

Proof. Assume without loss of generality that $\int g^- \, d\mu < \infty$. Then

$$\int (f+g)^- \, d\mu \leq \int f^- + g^- \, d\mu = \int f^- \, d\mu + \int g^- \, d\mu < \infty.$$

Then noting that

$$(f+g)^+ - (f+g)^- = f + g = f^+ - f^- + g^+ - g^-$$

gives

$$\begin{aligned} \int (f+g)^+ \, d\mu + \int f^- \, d\mu + \int g^- \, d\mu &= \int (f+g)^+ + f^- + g^- \, d\mu \\ &= \int (f+g)^- + f^+ + g^+ \, d\mu \end{aligned}$$

and finally, using the fact that $\int (f+g)^- \, d\mu$, $\int f^-$, and $\int g^-$ are finite, we may rearrange the above equality to get

$$\begin{gathered} \int (f+g)^+ \, d\mu - \int (f+g)^- \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu + \int g^+ \, d\mu - \int g^- \, d\mu \\ \int f + g \, d\mu = \int f + \int g. \end{gathered}$$

□

Theorem 4.11 (Beppo Levi's Monotone Convergence Theorem)**.** *Let $g, f, (f_n)_{n\geq 1}$ be measurable functions such that $\int |g| \, d\mu < \infty$, $g \leq f_n$ almost everywhere for each n, and $f_n \nearrow f$ almost everywhere. Then*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu \, .$$

Proof. We first assume $g = 0$, and handle the general case later. By assumption, there exists $N \in \mathcal{F}$ with $\mu(N) = 0$ such that $\widetilde{f}_n \nearrow \widetilde{f}$, where $\widetilde{f}_n := \mathbf{1}_{N^c} f_n$ and $\widetilde{f} := \mathbf{1}_{N^c} f$. Note that $\widetilde{f}_1 \geq 0$. Then we have $h_n \nearrow \widetilde{f}$, where

$$h_n := n \mathbf{1}_{\{\widetilde{f}_n \geq n\}} + \sum_{i=1}^{n2^n - 1} \frac{i}{2^n} \mathbf{1}_{\{\frac{i}{2^n} \leq \widetilde{f}_n < \frac{i+1}{2^n}\}}.$$

By the previous theorem, $\lim_{n\to\infty} \int h_n \, d\mu = \int \widetilde{f} \, d\mu$. Since $h_n \leq \widetilde{f}_n$ for each n, we have

$$\lim_{n\to\infty} \int f_n \, d\mu = \lim_{n\to\infty} \int \widetilde{f}_n \, d\mu = \int \widetilde{f} \, d\mu = \int f \, d\mu,$$

and we are finished for this case.

For general g, we can apply the above result for $f_n - g \nearrow f - g$ to get $\lim_{n\to\infty} \int f_n - g \, d\mu = \int f - g \, d\mu$. Adding both sides by $\int g \, d\mu$ (which is justified because $\int |g| \, d\mu < \infty$ by assumption) gives the result. □

Lemma 4.12. *If f is integrable, then for each $\epsilon > 0$ there exists $\delta > 0$ such that*

$$\int_A |f| \, d\mu < \epsilon$$

for any measurable set A satisfying $P(A) < \delta$.

Proof. By the monotone convergence theorem (Theorem 4.11),

$$\lim_{k\to\infty} \int_{\{|f|\le k\}} |f| \, d\mu = \int |f| \, d\mu,$$

so there exists a large K such that

$$\int |f| \, d\mu - \int_{\{|f|\le K\}} |f| \, d\mu < \epsilon/2.$$

Then for measurable set A satisfying $P(A) < \delta := \epsilon/(2K)$, we have

$$\begin{aligned} \int_A |f| \, d\mu &= \left(\int_A |f| \, d\mu - \int_{A\cap\{|f|\ge K\}} |f| \, d\mu \right) + \int_{A\cap\{|f|\ge K\}} |f| \, d\mu \\ &< \epsilon/2 + K \cdot P(A) \\ &< \epsilon. \end{aligned}$$

□

We now consider sequences of functions that are not necessarily monotone. In general, we cannot push the limit under the integral. Consider $f_n := n\mathbf{1}_{(0,1/n]}$. The functions converge to $f := 0$, but

$$\lim_{n\to\infty} \int f_n \, d\mu = 1 \neq 0 = \int f \, d\mu \, .$$

However, the following result does hold.

Theorem 4.13 (Fatou's Lemma)**.** *Let g and $(f_n)_{n\ge1}$ be measurable functions such that $\int |g| \, d\mu < \infty$ and $g \le f_n$ almost everywhere. Then*

$$\int \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int f_n \, d\mu \, .$$

Proof. Let $h_n := \inf_{m\ge n} f_m$. Then $h_n \nearrow \liminf_{n\to\infty} f_n$ and $h_n \ge g$ almost everywhere for all n. Thus,

$$\begin{aligned} \int \liminf_{n\to\infty} f_n \, d\mu &= \int \lim_{n\to\infty} h_n \, d\mu \\ &= \lim_{n\to\infty} \int h_n \, d\mu && \text{monotone convergence theorem} \\ &\le \lim_{n\to\infty} \inf_{m\ge n} \int f_m \, d\mu \\ &= \liminf_{n\to\infty} \int f_n \, d\mu \, . \end{aligned}$$

□

Corollary 4.14. *If g and $(f_n)_{n\ge1}$ are measurable functions such that $\int |g| \, d\mu$ and $g \ge f_n$ almost everywhere for all n, then*

$$\int \limsup_{n\to\infty} f_n \, d\mu \ge \limsup_{n\to\infty} \int f_n \, d\mu \, .$$

Theorem 4.15 (Lebesgue's Dominated Convergence Theorem)**.** *Let g, f, and $(f_n)_{n\geq 1}$ be measurable functions such that $\int |g|\,d\mu < \infty$, $|f_n| \leq |g|$ almost everywhere for each n, and $f_n \to f$ almost everywhere. Then*

1)

$$\int |f|\,d\mu \leq \int |g|\,d\mu < \infty.$$

2)

$$\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu\,.$$

3)

$$\lim_{n\to\infty} \int |f_n - f|\,d\mu = 0.$$

Proof. Since $|f_n| \leq |g|$ almost everywhere and $f_n \to f$ almost everywhere, we have $|f| \leq |g|$ almost everywhere. Thus $\int |f|\,d\mu \leq \int |g|\,d\mu < \infty$.

For the second result, note that

$$\begin{aligned}
\int f\,d\mu &\leq \liminf_{n\to\infty} \int f_n\,d\mu && \text{Fatou's lemma}\\
&\leq \limsup_{n\to\infty} \int f_n\,d\mu\\
&\leq \int \limsup_{n\to\infty} f_n\,d\mu && \text{Fatou's lemma (corollary, use } -g)\\
&= \int f\,d\mu,
\end{aligned}$$

so all inequalities above are equalities.

Finally, for the final result, note that $|f_n - f| \leq 2|g|$ and $\lim_{n\to\infty} |f_n - f| = 0$ almost everywhere. Applying the second result produces the third result. □

Theorem 4.16 (Fubini's theorem and Tonelli's theorem)**.** *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces and let $f : (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be measurable.*

a) ***Fubini's theorem.*** *If*

$$\int_{\Omega_1\times\Omega_2} |f|\,d(\mu_1 \otimes \mu_2) < \infty,$$

then

$$\int_{\Omega_1\times\Omega_2} f\,d(\mu_1 \otimes \mu_2) = \int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\,d\mu_2(\omega_2)\,d\mu_1(\omega_1) = \int_{\Omega_2}\int_{\Omega_1} f(\omega_1,\omega_2)\,d\mu_1(\omega_1)\,d\mu_2(\omega_2) \in \mathbb{R}.$$

b) ***Tonelli's theorem.*** *If $f \geq 0$ and μ_1 and μ_2 are both σ-finite, then*

$$\int_{\Omega_1\times\Omega_2} f\,d(\mu_1 \otimes \mu_2) = \int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\,d\mu_2(\omega_2)\,d\mu_1(\omega_1) = \int_{\Omega_2}\int_{\Omega_1} f(\omega_1,\omega_2)\,d\mu_1(\omega_1)\,d\mu_2(\omega_2) \in [0,\infty].$$

Definition 4.17. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We define

$$\mathcal{L}^0 := \{\text{measurable } f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}.$$

Let $\sim$ be the equivalence relation on $\mathcal{L}^0$ given by $f \sim g$ whenever $f = g$ almost everywhere. Then we define

$$L^0 := \mathcal{L}^0/\sim\,.$$

For $f \in \mathcal{L}^0$ and $p \in [1, \infty)$, we define the ***p*-norm**

$$\|f\|_p := \left(\int \|f\|^p \, d\mu \right)^{1/p},$$

along with the function spaces

$$\begin{aligned} \mathcal{L}^p &:= \{ f \in \mathcal{L}^0 : \|f\|_p < \infty \} \\ L^p &:= \mathcal{L}^p / \sim \end{aligned}$$

The p-norm is a **norm** on L^p, i.e., it satisfies

1) $\|cf\| = |c| \|f\|$ for $c \in \mathbb{R}$, $f \in L^p$,
2) $\|f + g\| \leq \|f\| + \|g\|$ for $f, g \in L^p$, and
3) $\|f\| = 0$ implies that $f = 0$ [note that $0 \in L^p$ is the equivalence class of all functions that are zero almost everywhere].

For $f \in \mathcal{L}^0$, we also define

$$\|f\|_\infty := \inf\{ \lambda \in R_+ : \mu(|f| \geq \lambda) = 0 \},$$

with the additional convention that $\inf \varnothing := \infty$. We define

$$\begin{aligned} \mathcal{L}^\infty &:= \{ f \in \mathcal{L}^0 : \|f\|_\infty < \infty \} \\ L^\infty &:= \mathcal{L}^\infty / \sim \end{aligned}$$

Proposition 4.18. *If $\mu(\Omega) < \infty$, then $L^q \subset L^p$ for $q \geq p$.*

Proof. Let $f \in L^q$. If $A := \{ f \leq 1 \}$, we easily see that $\int_A |f|^p \, d\mu \leq \mu(\Omega) < \infty$, so integrability of $|f|^p$ is determined by its behavior on A^c. However, on A^c, we have $|f|^p \leq |f|^q$, so $\int_{A^c} |f|^p \, d\mu \leq \int_{A^c} |f|^q \, d\mu < \infty$. □

Theorem 4.19 (Hölder's inequality)**.** *For $1 \leq p, q \leq \infty$ such that $1/p + 1/q = 1$, we have*

$$\|fg\|_1 := \int |fg| \, d\mu \leq \|f\|_p \|g\|_q.$$

Note that Hölder's inequality gives a precise bound for Proposition 4.18.

Corollary 4.20. *If $\mu(\Omega) < \infty$ and $1 \leq p < q \leq \infty$, then*

$$\|f\|_p \leq \mu(\Omega)^{\frac{1}{p} - \frac{1}{q}} \|f\|_q.$$

Proof. Let r be such that $\frac{1}{p} = \frac{1}{q} + \frac{1}{r}$. Then $\frac{1}{q/p} + \frac{1}{r/p} = 1$, so applying Hölder's inequality (Theorem 4.19) to f^p and the constant function 1 gives

$$\|f^p\|_1 \leq \|f^p\|_{q/p} \|1\|_{r/p} = \|f\|_q^p \cdot \mu(\Omega)^{p/r}.$$

Taking the pth root of both sides proves the result. □

Theorem 4.21 (Minkowski inequality)**.** *For $1 \leq p \leq \infty$,*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Theorem 4.22. *For every measure space $(\Omega, \mathcal{F}, \mu)$ and $p \in [1, \infty]$, L^p is a Banach space and L^2 is a Hilbert space with $\langle f, g \rangle := \int fg \, d\mu$.*

4.2 Probability measures and modes of convergence

Proposition 4.23. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(E, \mathcal{E})$ a measurable space, and $f : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ a measurable function. Then the* ***pushforward measure***

$$\mu_f(B) := \mu(f^{-1}(B))$$

is a measure on $(E, \mathcal{E})$ such that

$$\int_\Omega g \circ f \, d\mu \, . = \int_E g \, d\mu_f$$

holds for any measurable $g : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, provided the integral exists.

If μ is a probability measure, then so is μ_f, and it is called the ***distribution*** *of f, sometimes denoted $\mu_f = \mu \circ f^{-1}$.*

Definition 4.24. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. We define the following.

$$\begin{aligned} \mathrm{E}[X] &:= \int X \, dP && \text{for } X \in L^1 \\ \mathrm{Var}(X) &:= \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 && \text{for } X \in L^2 \\ \mathrm{Cov}(X, Y) &:= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] && \text{for } X \in L^2 \subset L^1 \end{aligned}$$

Example 4.25. Let $(\Omega, \mathcal{F}, P)$ be a probability space with $X \in L^1$. The distribution $P \circ X^{-1}$ is uniquely given by the cdf $F_X(x) := P(X \leq x)$ for $x \in \mathbb{R}$.

X has the same distribution as the identity function on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P \circ X^{-1})$. In particular,

$$\mathrm{E}[X] = \int_\mathbb{R} x \, d(P \circ X^{-1}) \, .$$

We can also use the Stieltjes integral to write the expectation as the sum of two Riemann integrals.

$$\begin{aligned} \mathrm{E}[X] &= \int_\mathbb{R} x \, dF_X(x) \\ &= \int_{-\infty}^0 x \, dF_X(x) - \int_0^\infty x \, dG_X(x) && G_X(x) := 1 - F_X(x) \\ &= -\int_{-\infty}^0 F_X(x) \, dx + \int_0^\infty G_X(x) \, dx && \text{integration by parts, see homework} \\ &= \int_{-\infty}^0 (P(X > x) - 1) \, dx + \int_0^\infty P(X > x) \, dx \, . \end{aligned}$$

We define the right-quantile function $q_X : (0,1) \to \mathbb{R}$ by

$$q_X(u) := \sup\{x \in \mathbb{R} : F_X(x) \leq u\}.$$

This is in some sense the "inverse" of F_X; it may be incorrect on a set of measure at most zero. It is a random variable on $((0,1), \mathcal{B}((0,1)), \lambda)$, where λ is the Lebesgue measure. We have

$$\lambda(q_X \leq x) = \lambda(\{u \in (0,1) : F_X(q_X(u)) \leq F_X(u)\}) = \lambda(u \leq F_X(u)$$

which implies $q_X \overset{d}{=} X$. Moreover,

$$\mathrm{E}[X] = \int_0^1 q_X(u) \, du \, .$$

Definition 4.26. Let $(\omega, \mathcal{F}, P)$ be a probability space and I a nonempty set.

1) A family of events $\{A_i\}_{i\in I}$ is **independent** if

$$P\left(\bigcap_{m=1}^{M} A_{i_m}\right) = \prod_{m=1}^{M} P(A_{i_m})$$

for every finite subset $\{i_1, \ldots, i_M\} \subset I$.

2) A family of σ-algebras $\{\mathcal{F}_i\}_{i\in I}$ is independent if the family of events $\{A_i\}_{i\in I}$ is independent for any $A_i \in \mathcal{F}_i$, $i \in I$.

3) A family of random variables $\{X_i\}_{i\in I}$ is independent if the family of σ-algebras $\{\sigma(X_i)\}_{i\in I}$ is independent.

Proposition 4.27. *Let $X_1, \ldots, X_M$ be random variables on a probability space $(\Omega, \mathcal{F}, P)$. The following are equivalent.*

1) $X_1, \ldots, X_M$ are independent.

2)

$$\mathrm{E}\left[\prod_{m=1}^{M} f_m(X_m)\right] = \prod_{m=1}^{M} \mathrm{E}[f_m(X_m)]$$

for all bounded Borel functions $f_1, \ldots, f_M$.

3)

$$\mathrm{E}[\exp(iu^T X)] = \prod_{m=1}^{M} \mathrm{E}[\exp(iu_m X_m)]$$

for any $u \in \mathbb{R}^M$.

Corollary 4.28. *If $X, Y \in L^2$ are independent, then*

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y],$$

i.e., $\mathrm{Cov}(X, Y) = 0$.

Proof. Take trunctations of the ranges of X and Y by $[-N, N]$, then take $N \to \infty$ and apply the dominated convergence theorem (Theorem 4.15). □

Theorem 4.29 (Borel-Cantelli Lemma)**.** *Let $A_1, A_2, \ldots$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, P)$.*

a) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P\left(\bigcap_{m\geq 1} \bigcup_{n\geq m} A_n\right) = 0.$$

b) If all the events are independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then

$$P\left(\bigcap_{m\geq 1} \bigcup_{n\geq m} A_n\right) = 1.$$

Proof.

a)

$$P\left(\bigcap_{m\geq 1} \bigcup_{n\geq m} A_n\right) = \lim_{m\to\infty} P\left(\bigcup_{n\geq m} A_n\right) \leq \lim_{m\to\infty} \sum_{n\geq m} P(A_n) = 0.$$

b)

$$\begin{aligned}
\log P\left(\bigcap_{n=m}^{M} A_n^c\right) &= \log \prod_{n=m}^{M} P(A_n^c) && \text{complements are independent}\\
&= \sum_{n=m}^{M} \log P(A_n^c)\\
&\le \sum_{n=m}^{M} (P(A_n^c) - 1) && \log(x) \le x - 1\\
&= -\sum_{n=m}^{M} P(A_n) \overset{M\to\infty}{\longrightarrow} -\infty,
\end{aligned}$$

implying $P\Big(\bigcap_{n\ge m} A_n^c\Big) = 0$, and thus $P\Big(\bigcup_{n\ge m} A_n\Big) = 1$. Finally,

$$P\left(\bigcap_{m\ge 1}\bigcup_{n\ge m} A_n\right) = \lim_{m\to\infty} P\left(\bigcup_{n\ge m} A_n\right) = 1.$$

□

Definition 4.30. Let $X, X_1, X_2, \dots$ be random variables on a probability space $(\Omega, \mathcal{F}, P)$. There exist the following concepts of convergence.

(i) $(X_n)_{n\ge 1}$ is said to converge to X almost surely if there exists a set $N \in \mathcal{F}$ with $P[N] = 0$ such that

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in \Omega \setminus N.$$

We denote this by $X_n \to X$ a.s.

(ii) For $p \in [1, \infty]$, $(X_n)_{n\ge 1}$ is said to converge to X in L^p if

$$\lim_{n\to\infty} \|X - X_n\|_p = 0.$$

We denote this by $X_n \to X$ in L^p or by $X_n \xrightarrow{L^p} X$.

(iii) $(X_n)_{n\ge 1}$ is said to converge to X in probability if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P[|X_n - X| \ge \epsilon] = 0.$$

We denote this by $X_n \to X$ in probability or by $X_n \xrightarrow{P} X$.

(iv) $(X_n)_{n\ge 1}$ is said to converge to X in distribution if

$$\lim_{n\to\infty} \mathrm{E}[f(X_n)] = \mathrm{E}[f(X)]$$

for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$. We denote this by $X_n \to X$ in distribution or by $X_n \xrightarrow{d} X$. (Note that this notion of convergence also makes sense if the random variables $X, X_1, X_2, \dots$ are all defined on different probability spaces.)

Proposition 4.31. *The following implications hold.*

a) (i) $\implies$ (iii)

b) (ii) $\implies$ (iii)

c) (iii) $\implies$ (iv)

d) It follows from (iii) that there exists a subsequence $(X_{n_k})_{k\geq 1}$ that converges to X a.s.

e) If the $(X_n)_{n\geq 1}$ are dominated by an L^p function, then (i) $\implies$ (ii).

Lemma 4.32. *If F is a cdf, then there are at most countably many $x \in \mathbb{R}$ at which F is not continuous.*

Proposition 4.33. *A sequence of random variables $(X_n)_{n\geq 1}$ with respective cdfs $(F_n)_{n\geq 1}$ converges in distribution to a random variable X with cdf F if and only if $F_n(x) \to F(x)$ whenever F is continuous at x.*

4.3 Uniform integrability

Definition 4.34. A family of random variables $(X_i)_{i\in I}$ on a common probability space $(\Omega, \mathcal{F}, P)$ is **uniformly integrable** if

$$\lim_{c\to\infty} \sup_{i\in I} \int_{\{|X_i|>c\}} |X_i| \, dP = 0.$$

Lemma 4.35.

1) If $(X_i)_{i\in I}$ is a family of random variables on a common probability space $(\Omega, \mathcal{F}, P)$ such that $|X_i| \leq |X|$ for all $i \in I$, where $X \in L^1(\Omega, \mathcal{F}, P)$, then the family is uniformly integrable.

2) If we have finitely many random variables $X_1, \ldots, X_n \in L^1(\Omega, \mathcal{F}, P)$, then they are uniformly integrable.

Proof. For the first statement,

$$\lim_{c\to\infty} \int_{\{|X_i|>c\}} |X_i| \, dP \leq \lim_{c\to\infty} \mathrm{E}[\mathbf{1}_{\{|X|>c\}} |X|] = 0,$$

where the last equality follows from the dominated convergence theorem (Theorem 4.15) because

$$\mathbf{1}_{\{|X|>c\}} |X| \to 0$$

almost surely as $c \to \infty$.

For the second statement, note that $|X_i| \leq |X_1| + \cdots + |X_n|$ for each $i \in \{1, \ldots, n\}$ and apply the previous statement. □

Proposition 4.36. *A family of random variables $(X_i)_{i\in I}$ on a common probability space $(\Omega, \mathcal{F}, P)$ is uniformly integrable if and only if both of the following statements hold.*

a) The family is bounded in L^1, that is,

$$\sup_{i\in I} \mathrm{E}[|X_i|] < \infty.$$

b) For each $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\int_A |X_i| \, dP \leq \epsilon$$

for all $i \in I$ and $A \in \mathcal{F}$ such that $P(A) \leq \delta$.

Proof. We first observe that for $A \in \mathcal{F}$, we have

$$\int_A |X_i| \, dP = \int_{A\cap\{|X_i|\leq c\}} |X_i| \, dP + \int_{A\cap\{|X_i|<c\}} |X_i| \, dP \leq c \cdot P(A) + \int_{A\cap\{|X_i|>c\}} |X_i| \, dP \, .$$

So, if the family is uniformly integrable, there exists $c \geq 0$ such that $\int_{\{|X_i|>c\}} |X_i| \, dP \leq 1$ for all $i \in I$. Then the above statement with $A = \Omega$ gives $\mathrm{E}[|X_i|] \leq c \cdot P(\Omega) + 1 = c + 1$ for all $i \in I$, proving a).

Fix $\epsilon > 0$. If the family is uniformly integrable, there exists $c \geq 0$ such that $\int_{\{|X_i|>c\}} |X_i|\, dP \leq \epsilon/2$ for all $i \in I$. Let $\delta = \epsilon/(2c)$. If $P(A) \leq \delta$, then by our work above we have

$$\int_A |X_i|\, dP \leq c \cdot P(A) + \int_{\{|X_i|>c\}} |X_i|\, dP \leq \epsilon,$$

proving b).

We now show the converse; suppose a) and b) hold, and let $K := \sup_{i\in I} \mathrm{E}[|X_i|] < \infty$. Then

$$c \cdot P(|X_i| > c) \leq \int_{\{|X_i|>c\}} |X_i|\, dP \leq K$$

for all $i \in I$. Given $\epsilon > 0$, choose δ such that b) holds, and let $c := K/\delta$. We just showed that $P(|X_i| > c) \leq K/c = \delta$ for all $i \in I$. By b), $\int_{\{|X_i|>c\}} |X_i|\, dP \leq \epsilon$ for all $i \in I$, proving uniform integrability. □

Lemma 4.37. *If $(X_i)_{i\in I}$ and $(Y_i)_{i\in I}$ are two uniformly integrable families of random variables on a common probability space, indexed by the same index set, then $(X_i + Y_i)_{i\in I}$ is also uniformly integrable.*

Proof. We use Proposition 4.36. Since a) holds for each family, we have

$$\sup_{i\in I} \mathrm{E}[|X_i + Y_i|] \leq \sup_{i\in I} \mathrm{E}[|X_i|] + \sup_{i\in I} \mathrm{E}[|Y_i|] < \infty.$$

Fix $\epsilon > 0$. Since b) holds for each family, there exists $\delta > 0$ such that $\int_A |X_i|\, dP \leq \epsilon/2$ and $\int_A |Y_i|\, dP \leq \epsilon/2$ for all $i \in I$ and any $A \in \mathcal{F}$ such that $P(A) \leq \delta$. Then

$$\int_A |X_i + Y_i|\, dP \leq \int_A |X_i|\, dP + \int_A |Y_i|\, dP \leq \epsilon.$$

Thus, $(X_i + Y_i)_{i\in I}$ is uniformly integrable. □

Theorem 4.38. *Let X and $(X_n)_{n\geq 1}$ be random variables on a common probability space $(\Omega, \mathcal{F}, P)$ such that $X_n \to X$ almost surely and such that $(X_n)_{n\geq 1}$ is uniformly integrable. Then*

1. *$X \in L^1$,*
2. *$\mathrm{E}[X_n] \to \mathrm{E}[X]$, and*
3. *$\mathrm{E}[|X_n - X|] \to 0$ (i.e., L^1-convergence).*

Proof. By Proposition 4.31, there exists a subsequence $(X_{n_k})_{k\geq 1}$ such that $X_{n_k} \to X$ almost surely. By Fatou's lemma (Theorem 4.13) and uniform integrability, we have

$$\mathrm{E}[|X|] \leq \liminf_{k\to\infty} \mathrm{E}[|X_{n_k}|] < \infty,$$

which shows that $X \in L^1$.

By the previous lemma, the family $(X_n - X)_{n\geq 1}$ is uniformly integrable. Given $\epsilon > 0$, choose $c \geq 0$ such that

$$\int_{\{|X_n - X|>c\}} |X_n - X|\, dP \leq \epsilon/3,$$

and choose n_0 such that

$$P(|X_n - X| > \epsilon/3) \leq \epsilon/(3c)$$

for all $n \geq n_0$. Then,

$$\begin{aligned}\mathrm{E}[|X_n - X|] &= \int_{\{|X_n - X|\leq\epsilon/3\}} |X_n - X|\, dP + \int_{\{\epsilon/3<|X_n - X|\leq c\}} |X_n - X|\, dP + \int_{\{|X_n - X|>c\}} |X_n - X|\, dP \\ &\leq \epsilon/3 + c \cdot P(|X_n - X| > \epsilon/3) + \epsilon/3 \\ &\leq \epsilon,\end{aligned}$$

showing that $\mathrm{E}[|X_n - X|] \to 0$.

To show the remaining claim, note that both $\mathrm{E}[(X_n - X)^+]$ and $\mathrm{E}[(X_n - X)^-]$ are bounded by $\mathrm{E}[|X_n - X|] \to 0$ for all n, so $\mathrm{E}[X_n - X] \to 0$, i.e., $\mathrm{E}[X_n] \to \mathrm{E}[X]$. □

Theorem 4.39 (de la Vallée-Poussin)**.** *A family of random variables $(X_i)_{i\in I}$ on a common probability space $(\Omega, \mathcal{F}, P)$ is uniformly integrable if and only if there exists a function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\lim_{x\to\infty} \frac{\varphi(x)}{x} = \infty$$

and such that

$$\sup_{i\in I} \mathrm{E}[\varphi(|X_i|)] < \infty.$$

Further, if the family is uniformly integrable, this φ can be chosen to be convex and nondecreasing.

Proof. Suppose there exists such a φ for a family. Fix $\epsilon > 0$ and let $K := \sup_{i\in I} \mathrm{E}[\varphi(|X_i|)]$. There exists a $c \geq 0$ such that $\varphi(x)/x \geq K/\epsilon$ for all $x \geq c$. Then,

$$\int_{\{|X_i|>c\}} |X_i|\, dP \leq \int_{\{|X_i|>c\}} \frac{\varphi(|X_i|)\cdot\epsilon}{K}\, dP \leq \epsilon.$$

Thus, the family is uniformly integrable.

We now show the converse. Suppose the family is uniformly integrable. Then there exists a strictly increasing sequence $(c_n)_{n\geq 1}$ such that $c_1 \geq 1$ and such that

$$\int_{\{|X_i|>c_n\}} |X_i|\, dP \leq \frac{1}{2^n}$$

for all $i \in I$.

We define

$$f := \sum_{n\geq 1} n\mathbf{1}_{(c_n,c_{n+1}]} = \sum_{n\geq 1} \mathbf{1}_{(c_n,\infty)},$$

and

$$\varphi(x) := \int_0^x f(y)\, dy\,.$$

This is a piecewise-linear convex function with increasing slope. Also, $\varphi(x)/x \to \infty$ as $x \to \infty$.

We have

$$\begin{aligned}
\mathrm{E}[\varphi(|X_i|)] &= \mathrm{E}\left[\int_0^\infty \mathbf{1}_{\{|X_i|>y\}} f(y)\, dy\right] \\
&= \int_0^\infty f(y) P(|X_i| > y)\, dy && \text{Tonelli (Theorem 4.16)} \\
&= \sum_{n\geq 1} \int_{c_n}^\infty P(|X_i| > y)\, dy && \text{MCT (Theorem 4.11)} \\
&\leq \sum_{n\geq 1} \frac{1}{2^n} && \text{see below} \\
&= 1.
\end{aligned}$$

To bound the integral above, note the following.

$$\begin{aligned}
\frac{1}{2^n} &\geq \int_{\{|X_i|>c_n\}} |X_i|\, dP \\
&= \int_\Omega \int_0^\infty \mathbf{1}_{\{|X_i|>y\}\cap\{|X_i|>c_n\}}\, dy\, dP \\
&= \int_0^\infty P(|X_i| > \max\{y, c_n\})\, dy && \text{Tonelli (Theorem 4.16)} \\
&= c_n \cdot P(|X_i| > c) + \int_{c_n}^\infty P(|X_i| > y)\, dy \\
&\geq \int_{c_n}^\infty P(|X_i| > y)\, dy\,.
\end{aligned}$$

□

Corollary 4.40. *A family of random variables $(X_i)_{i\in I}$ on a common probability space is uniformly integrable if*

$$\sup_{i\in I}\|X_i\|_p < \infty$$

for some $p \in (1,\infty]$.

Proof. Noting that $\|X_i\|_p = \mathrm{E}[|X_i|^p]^{1/p}$, we see that the assumption implies that $\sup_{i\in I}\mathrm{E}[|X_i|^p] < \infty$. Applying the previous theorem with $\varphi(x) := x^p$ finishes the proof. □

4.4 Jensen's inequality

Definition 4.41. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is **convex** if

$$\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$$

holds for any $x, y \in \mathbb{R}$ and $0 \leq \lambda \leq 1$.

Proposition 4.42. *Convex functions are continuous, and their one-sided derivatives*

$$\lim_{\epsilon\searrow 0}\frac{\varphi(x+\epsilon)-\varphi(x)}{\epsilon}, \quad \lim_{\epsilon\searrow 0}\frac{\varphi(x)-\varphi(x-\epsilon)}{\epsilon}$$

exist for all $x \in \mathbb{R}$.

Theorem 4.43 (Jensen's inequality)**.** *Let X be a random variable, let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function, and let $X, \varphi(X) \in L^1$. Then*

$$\mathrm{E}[\varphi(X)] \geq \varphi(\mathrm{E}[X]).$$

Proof. Let $a := \mathrm{E}[X]$ and

$$b := \lim_{\epsilon\searrow 0}\frac{\varphi(a+\epsilon)-\varphi(a)}{\epsilon}.$$

Then for all $x \in \mathbb{R}$ we have

$$b \leq \frac{\varphi(x)-\varphi(a)}{x-a}.$$

Thus,

$$\varphi(X) \geq \varphi(a) + b\cdot(X-a)$$
$$\mathrm{E}[\varphi(X)] \geq \varphi(a) + b\cdot(\mathrm{E}[X]-a) = \varphi(\mathrm{E}[X]).$$

□

Corollary 4.44. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. Then*

$$\|X\|_p \leq \|X\|_q$$

for $1 \leq p \leq q \leq \infty$.

Proof. If $q = \infty$, then

$$\mathrm{E}[|X|^p]^{1/p} \leq \mathrm{E}[\|X\|_\infty^p]^{1/p} \leq \|X\|_\infty.$$

If $q < \infty$, then noting that $x \mapsto x^{q/p}$ is a convex function, we have the following from Jensen's inequality.

$$\|X\|_q = \mathrm{E}[|X|^q]^{1/q} = \mathrm{E}[|X|^{p\cdot q/p}]^{1/q} \geq \mathrm{E}[|X|^p]^{1/p}.$$

□

4.5 Weak law of large numbers

Lemma 4.45. *If $X_1, \ldots, X_n$ are random variables, then*

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

Proof.

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^n X_i\right) &= \operatorname{E}\left[\left(\sum_{i=1}^n X_i - \operatorname{E} X_i\right)^2\right] \\ &= \operatorname{E}\left[\sum_{i=1}^n (X_i - \operatorname{E} X_i)^2 + \sum_{i\neq j}(X_i - \operatorname{E} X_i)(X_j - \operatorname{E} X_j)\right] \\ &= \sum_{i=1}^n \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

□

Theorem 4.46 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be uncorrelated (zero pairwise covariance) random variables in L^2 such that $\operatorname{E}[X_n] = m$ for all n and such that $\sup_n \operatorname{E}[X_n^2] < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^n X_i \to m$$

in L^2 (and thus, in probability as well).

Proof. Let $k := \sup_n \operatorname{E}[X_n^2] < \infty$. Then

$$\begin{aligned}\operatorname{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - m\right)^2\right] &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}(X_i) && \text{the } X_i \text{ are uncorrelated} \\ &= \frac{1}{n^2}\sum_{i=1}^n \left(\operatorname{E}[X_i^2] - m^2\right) \\ &\leq \frac{k-m^2}{n} \\ &\to 0\end{aligned}$$

as $n \to \infty$, showing L^2 convergence. Proposition 4.31 shows why convergence in L^2 implies convergence in probability. □

4.6 Types of distributions

Definition 4.47. Let Ω be a nonempty set. Given $\omega \in \Omega$, the **Dirac measure** is the map $\delta_\omega : 2^\Omega \to \{0, 1\}$ defined by

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

Definition 4.48. Let μ be a measure on $\mathbb{R}^d$.

- We call μ **discrete** if $\mu = \sum_{n\geq 1} p_n \delta_{x_n}$ for $x_1, x_2, \in \mathbb{R}^d$ and $p_1, p_2, \ldots \geq 0$.
- We call μ **continuous** if $\mu((a_1, b_1] \times \cdots \times (a_d, b_d])$ is continuous in $b \in \mathbb{R}^d$.
- We call μ **absolutely continuous** if there exists a Borel function[2] $f : \mathbb{R}^d \to \mathbb{R}_+$ such that
$$\mu((a_1, b_1] \times \cdots \times (a_d, b_d]) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(y_1, \ldots, y_d)\, dy_1 \cdots dy_d\,.$$
This f is called the **density** of μ. Moreover,
$$\mu(A) = \int_A f(x)\, dx$$
for any $A \in \mathcal{B}(\mathbb{R})^{\otimes d}$ and
$$\int_{\mathbb{R}^d} g(x)\, d\mu(x) = \int_{\mathbb{R}^d} g(x) f(x)\, dx$$
for any measurable function $g : \mathbb{R}^d \to \mathbb{R}$.

Note that if $f_1, \ldots, f_d$ are densities on $\mathbb{R}$, then $f(x) := f_1(x_1) f_2(x_2) \cdots f_d(x_d)$ is a density on $\mathbb{R}^d$, and the corresponding measure is $\mu = \mu_1 \otimes \cdots \otimes \mu_d$.

Definition 4.49. A **random vector** on a probability space $(\Omega, \mathcal{F}, P)$ is a vector $X := (X_1, \ldots, X_d)$ where $X_1, \ldots, X_d$ are random variables. It is a measurable mapping $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R})^{\otimes d})$.

- We call X **discrete** if the pushforward measure $P \circ X^{-1}$ is discrete.
- We call X **continuous** if the pushforward measure $P \circ X^{-1}$ is continuous.
- We call X **absolutely continuous** if the pushforward measure $P \circ X^{-1}$ is absolutely continuous.

Example 4.50. We list some examples of discrete distributions on $\mathbb{R}$.

1. **Discrete uniform.** $\mu(n) := 1/N$, for $n = 1, \ldots, N$.
2. **Bernoulli.** $\mu(0) := 1 - p$, $\mu(1) := p$, with $p \in [0, 1]$.
3. **Binomial.** $\mu(n) = \binom{N}{n} p^n (1-p)^{N-n}$, for $n = 0, \ldots, N$.
4. **Poisson.** $\mu(n) := e^{-\lambda} \frac{\lambda^n}{n!}$, for $n = 0, 1, \ldots$, where $\lambda > 0$.
5. **Geometric.** With $0 < p \leq 1$,
   - $\mu(n) := p(1-p)^{n-1}$, for $n = 1, 2, \ldots$,
   - $\mu(n) := p(1-p)^n$, for $n = 0, 1, \ldots$.

Example 4.51. We list some examples of densities on $\mathbb{R}$.

1. **Gaussian.** $f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$, for all $x \in \mathbb{R}$, with $\mu \in \mathbb{R}$ and $\sigma > 0$.
2. **Uniform.** $f(x) := 1/(b - a)$ for $a \leq x \leq b$.
3. **Exponential.** $f(x) := \lambda e^{-\lambda x}$ for $x \geq 0$, with $\lambda > 0$. An important property of this distribution is memorylessness, which we will explore later.
4. **Bilateral exponential.** $f(x) := \frac{1}{2} \lambda e^{-\lambda |x|}$ for $x \in \mathbb{R}$, with $\lambda > 0$. [One can also consider using a two different values of λ for the positive reals and the negative reals.]
5. **Cauchy.** $f(x) := \theta / (\pi(x^2 + \theta^2))$ for $x \in \mathbb{R}$, with $\theta > 0$. A Cauchy random variable has no moments due to its heavy tails.

If $X, Y \sim \mathcal{N}(0, 1)$ are independent, then $X + Y \overset{d}{=} \sqrt{2} X$. If $X, Y \sim \text{Cauchy}(0)$ are independent, then $X + Y \overset{d}{=} 2X$.

[2] A Borel function is measurable with respect to the Borel σ-algebras on both its domain and its range.

4.7 The characteristic function

Definition 4.52. Let X be a d-dimensional random variable with cdf $F(x_1, \dots, x_d) := P(X_1 \le x_1, \dots, X_d \le x_d)$. The **characteristic function** $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ is defined by

$$u \mapsto \mathrm{E}[\exp(iu^t X)],$$

where $u^t X := u_1 X_1 + \cdots + u_d X_d$.

If F has density f, then

$$\varphi_X(u) = \int_{\mathbb{R}^d} \exp(iu^t x) f(x)\, dx\,.$$

Definition 4.53. If $Z = V + iW$ for $V, W \in L^1$, then we define $\mathrm{E}[Z] = \mathrm{E}[V] + i\,\mathrm{E}[W]$. In particular, $\mathrm{E}[e^{iu^T X}] = \mathrm{E}[\cos(u^T X)] + i\,\mathrm{E}[\sin(u^T X)]$.

Additionally, $\mathrm{E}[|Z|] = \mathrm{E}[\sqrt{V^2 + W^2}] \le \mathrm{E}[|V| + |W|] < \infty$.

Lemma 4.54. *If $Z = V + iW$ for $V, W \in L^1$, then $|\mathrm{E}[Z]| \le \mathrm{E}[|Z|]$.*

Proof.

$$\begin{aligned}
|\mathrm{E}[Z]| &= \sup_{q\in\mathbb{Q}} \mathrm{Re}(e^{iq}\,\mathrm{E}[Z]) \\
&= \sup_{q\in\mathbb{Q}} \mathrm{E}[\mathrm{Re}(e^{iq} Z)] \\
&\le \mathrm{E}\left[\sup_{q\in\mathbb{Q}} \mathrm{Re}(e^{iqZ})\right] \\
&= \mathrm{E}[|Z|].
\end{aligned}$$

□

Lemma 4.55. *Let X be a d-dimensional random variable.*

1) $|\varphi_X(u)| \le 1$.

2) $\varphi_X(-u) = \overline{\varphi_X(u)} = \varphi_{-X}(u)$.

3) If φ_X is real-valued, then $X \overset{d}{=} -X$.

4) $u \mapsto \varphi_X(u)$ is uniformly continuous in $\mathbb{R}^d$.

Proof. For 1), note that

$$|\varphi_X(u)| \le \mathrm{E}[|\exp(iu^t X)|] \le 1.$$

For 2), note that

$$\varphi_{-X}(u) = \varphi_X(-u) = \mathrm{E}[\overline{\exp(iu^t X)}] = \overline{\varphi_X(u)}.$$

For 3), if φ_X is real valued, then 2) shows that $\varphi_X(u) = \varphi_{-X}(u)$.

For 4),

$$|\varphi_X(hu) - \varphi_X(u)| \le \mathrm{E}[|\exp(i(h+u)^t X) - \exp(iu^t X)|] \le \mathrm{E}[|\exp(ih^t X) - 1|] \to 0$$

as $h \to 0$.

□

Lemma 4.56. *Let Y be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ with $Y \ge 0$ a.s. and $\mathrm{E}[Y] = 1$. Then $Q : \mathcal{F} \to \mathbb{R}_+$ defined by*

$$Q(A) := \mathrm{E}[\mathbf{1}_A Y]$$

is a probability measure on $(\Omega, \mathcal{F})$. Moreover, if X is a random variable on $(\Omega, \mathcal{F})$, then

$$\mathrm{E}_Q[X] = \mathrm{E}[XY].$$

Proof. We know Q maps into $\mathbb{R}_+$ because $Y \geq 0$ almost surely. Clearly we have $Q(\Omega) = \mathrm{E}[Y] = 1$ and $Q(\varnothing) = \mathrm{E}[0] = 0$. Countable additivity holds readily.

$$Q\left(\bigcup_{n\geq 1} A_n\right) = \mathrm{E}\left[\sum_{n\geq 1} \mathbf{1}_{A_n} Y\right] = \sum_{n\geq 1} \mathrm{E}[\mathbf{1}_{A_n} Y] = \sum_{n\geq 1} Q(A_n).$$

We omit the proof of the change of measure formula. It is clear when X is a simple function; for general X, approximate it by simple functions. □

Theorem 4.57. *Let X be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$.*

1) If $\mathrm{E}[|X|^n] < \infty$ for some $n \in \mathbb{N}$, then we have the following.

- *$\varphi_X^{(k)}(u)$ exists for all $k \leq n$.*
- *$\varphi_X^{(k)}(u) = \mathrm{E}[(iX)^k \exp(iuX)]$.*
- *$\mathrm{E}[X^k] = (-i)^k \varphi_X^{(k)}(0)$.*
- *$\varphi_X(u) = \sum_{k=0}^n \frac{(iu)^k}{k!} \mathrm{E}[X^k] + \frac{(iu)^n}{n!}\epsilon_n(u)$ with $\lim_{u\to 0} \epsilon_n(u) = 0$.*

2) If $\varphi_X^{(2k)}(0)$ exists then $\mathrm{E}[X^{2k}] < \infty$.

Proof of 1). The third and fourth bullets follow directly from the first and second.

Consider the case $n = 1$. We first note that

$$\lim_{h\to 0} \frac{\exp(ihX) - 1}{h} = \frac{d}{dh} \exp(ihX)\bigg|_{h=0} = iX.$$

Also, from the bound $|\exp(i\theta) - 1| \leq |\theta|$ we have

$$\left|\frac{\exp(ihX) - 1}{h}\right| \leq |X|.$$

We have $|X| \in L^1$ by assumption. Thus we may use the dominated convergence theorem (Theorem 4.15) to prove the case $n = 1$.

$$\varphi_X'(u) = \lim_{h\to 0} \frac{\varphi_X(u+h) - \varphi_X(u)}{h} = \lim_{h\to 0} \mathrm{E}\left[\exp(iuX)\frac{\exp(ihX) - 1}{h}\right] = \mathrm{E}[iX \exp(iuX)].$$

We prove the result for general n by induction. Note that $\mathrm{E}[|X|^{n+1}] < \infty$ implies $\mathrm{E}[|X|^k] < \infty$ for all $k \leq n+1$ (Proposition 4.18 or Theorem 4.19), so by induction $\varphi_X^{(k)}(u)$ exists for all $k \leq n$. Then, by the same argument,

$$\varphi_X^{(n+1)}(u) = \lim_{h\to 0} \mathrm{E}\left[(iX)^n \exp(iuX)\frac{\exp(ihX) - 1}{h}\right] = \mathrm{E}[(iX)^{n+1} \exp(iuX)].$$

□

Proof of 2). Consider the case $k = 1$.

$$
\begin{aligned}
\varphi_X''(0) &= \lim_{h\to 0} \frac{1}{2}\left(\frac{\varphi(2h)-\varphi(0)}{2h} + \frac{\varphi(0)-\varphi(-2h)}{2h}\right) \\
&= \lim_{h\to 0} \frac{1}{4h^2}(\varphi(2h)+\varphi(-2h)-2\varphi(0)) \\
&= \lim_{h\to 0} \frac{1}{4h^2}\,\mathrm{E}[\exp(i2hX)+\exp(-i2hX)-2] \\
&= \lim_{h\to 0} \frac{1}{4h^2}\,\mathrm{E}[(\exp(ihX)-\exp(-ihX))^2] \\
&= -\lim_{h\to 0} \mathrm{E}\left[\left(\frac{\sin(hX)}{hX}\right)^2 X^2\right] \\
&\le -\mathrm{E}\left[\liminf_{h\to 0}\left(\frac{\sin(hX)}{hX}\right)^2 X^2\right] && \text{Fatou's lemma (Theorem 4.13)} \\
&= -\mathrm{E}[X^2].
\end{aligned}
$$

Since $\mathrm{E}[X^2] \le -\varphi_X''(0) < \infty$, we have shown the result for $k = 1$.

To show the result for for general k, we use induction. Suppose $\varphi_X^{(2k+2)}(0)$ exists. Then $\varphi_X^{(2k)}(0)$ exists, which implies $\mathrm{E}[X^{2k}] < \infty$ by the inductive hypothesis. We know $\mathrm{E}[X^{2k}] = \mathrm{E}[(X^k)^2] \ge 0$. If $\mathrm{E}[X^{2k}] = 0$, then $X^{2k} = 0$ a.s. (because it is nonnegative), so $\mathrm{E}[X^{2k+2}] = 0$ and we are finished. Thus, we may assume $\mathrm{E}[X^{2k}] > 0$.

Let $Q : \mathcal{F} \to \mathbb{R}$ be defined by

$$
Q(A) := \mathrm{E}\left[\mathbf{1}_A \frac{X^{2k}}{\mathrm{E}[X^{2k}]}\right] = \frac{\mathrm{E}[\mathbf{1}_A X^{2k}]}{\mathrm{E}[X^{2k}]}.
$$

By Lemma 4.56, Q is a probability measure. Let $\varphi_X^Q(u) := \mathrm{E}_Q[\exp(iuX)]$ be the characteristic function with respect to the measure Q. Then,

$$
\begin{aligned}
\varphi_X^Q(u) &= \mathrm{E}_Q[\exp(iuX)] \\
&= \mathrm{E}\left[\frac{X^{2k}\exp(iuX)}{\mathrm{E}[X^{2k}]}\right] && \text{Lemma 4.56} \\
&= \frac{\mathrm{E}[X^{2k}\exp(iuX)]}{\mathrm{E}[X^{2k}]} \\
&= \frac{(-1)^k\varphi_X^{(2k)}(u)}{\mathrm{E}[X^{2k}]} && \text{by part 1)}
\end{aligned}
$$

Thus, $(\varphi_X^Q)''(0)$ exists, which implies $\mathrm{E}_Q[X^2] < \infty$ by our work in the case $k = 1$. By the definition of E_Q, we have $\mathrm{E}[X^{2k+2}]/\mathrm{E}[X^{2k}] < \infty$, and finally, $\mathrm{E}[X^{2k+2}] < \infty$. □

Example 4.58. We give some examples of characteristic functions.

1) **Dirac delta.** If $P(X = a) = 1$, then $\varphi_X(u) = \exp(iua)$.

2) **Gaussian.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\varphi_X(u) = \exp(iu\mu - \frac{1}{2}\sigma^2 u^2)$. (See Lemma 4.62.)

3) **Bernoulli.** If $X \sim \mathrm{Ber}(p)$, then $\varphi_X(u) = (1-p) + p\exp(iu)$.

4) **Binomial.** If $X \sim \mathrm{Bin}(n, p)$, then it is the sum of n i.i.d. $\mathrm{Ber}(p)$ random variables $X_1, \ldots, X_n$, so using independence (Proposition 1.9),

$$
\varphi_X(u) := \mathrm{E}[\exp(iuX)] = \mathrm{E}[\exp(iu(X_1 + \cdots + X_n))] = \mathrm{E}[\exp(iuX_1)]^n = (1 - p + p\exp(iu))^n.
$$

5) **Poisson.** If $P(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}$ for $n \geq 0$, then

$$\varphi_X(u) := \mathrm{E}[\exp(iuX)] = \sum_{n\geq 0} e^{iun} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n\geq 0}\frac{(e^{iu}\lambda)^n}{n!} = \exp(\lambda(e^{iu} - 1)).$$

6) **Exponential.** If $X \sim \mathrm{Expon}(\lambda)$, then $\varphi(u) = \frac{\lambda}{\lambda - iu}$.

The most important property of the characteristic function is that it uniquely determines the distribution of the random variable.

Theorem 4.59 (Inversion theorem)**.** *Let X be a random variable with cdf F and characteristic function $\varphi(u) = \mathrm{E}[\exp(iuX)]$.*

1) Let $a < b$ in $\mathbb{R}$, and let $F(x_0^-) := \lim_{x \nearrow x_0} F(x)$. We have

$$\frac{F(b) + F(b^-)}{2} - \frac{F(a) + F(a^-)}{2} = \lim_{c\to\infty}\Phi(c),$$

where

$$\Phi(c) := \int_{-c}^{c}\frac{\exp(-ita) - \exp(-itb)}{it}\varphi(t)\,dt\,.$$

2) If $\int_{\mathbb{R}}|\varphi(u)|\,du < \infty$, then F has a density f and f is continuous.

Proof of 1). We define the function g_c as

$$\begin{aligned} g_c(x) &:= \int_{-c}^{c}\frac{\sin(t(x-a)) - \sin(t(x-b))}{t}\,dt \\ &= \int_{-c(x-a)}^{c(x-a)}\frac{\sin u}{u}\,du - \int_{-c(x-b)}^{c(x-b)}\frac{\sin u}{u}\,du \qquad\qquad u = t(x-a), u = t(x-b) \\ &= 2\int_0^{c(x-a)}\frac{\sin u}{u}\,du - 2\int_0^{c(x-b)}\frac{\sin u}{u}\,du\,. \end{aligned}$$

Recalling the Dirichlet integral

$$\int_0^\infty \frac{\sin u}{u}\,du = \pi/2,$$

we observe the limiting behavior of g_c as c tends to infinity.

$$\lim_{c\to\infty} g_c(x) = \begin{cases} 0 & x > b \text{ or } x < a, \\ 2\pi & a < x < b, \\ \pi & x \in \{a, b\}. \end{cases}$$

We can also rewrite this as a simple function.

$$\lim_{c\to\infty} g_c(x) = 2\pi\mathbf{1}_{(a,b)}(x) + \pi(\mathbf{1}_{\{a\}}(x) + \mathbf{1}_{\{b\}}(x)).$$

This allows to arrive at the desired result.

$$\begin{aligned}
\Phi(c) &= \int_{-c}^{c} \frac{\exp(-ita)-\exp(-itb)}{it}\,\mathrm{E}[\exp(itX)]\,dt \\
&= \mathrm{E}\left[\frac{1}{2\pi}\int_{-c}^{c} \frac{\exp(-it(X-a))-\exp(-it(X-b))}{it}\,dt\right] && \text{Fubini (Theorem 4.16)} \\
&= \mathrm{E}\left[\frac{1}{2\pi}\int_{-c}^{c} \frac{\sin(t(X-a))-\sin(t(X-b))}{t}\,dt\right] && \text{terms involving cosine are odd} \\
&= \frac{1}{2\pi}\,\mathrm{E}[g_c(X)] \\
&= \frac{1}{2\pi}\,\mathrm{E}[2\pi\mathbf{1}_{(a,b)}(X)+\pi(\mathbf{1}_{\{a\}}(X)+\mathbf{1}_{\{b\}}(X))] \\
&= P(a<X<b)+\frac{1}{2}P(X=a)+\frac{1}{2}P(X=b) \\
&= F(b^-)-F(a)+\frac{1}{2}(F(a)-F(a^-)+F(b)-F(b^-)) \\
&= \frac{F(b)+F(b^-)}{2}-\frac{F(a)+F(a^-)}{2}.
\end{aligned}$$

□

Proof of 2). Let

$$f(x) := \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt\,.$$

Since we have $\int_{-\infty}^{\infty}|\varphi(t)|\,dt<\infty$ by assumption, we know f is well-defined and continuous in x. [Fourier transform of L^1 function is continuous.]

Let $a<b$ with F continuous at both a and b.

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_a^b \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt\,dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi(t)\int_a^b e^{-itx}\,dx\,dt && \text{Fubini (Theorem 4.16)} \\
&= \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\frac{\exp(-ita)-\exp(-itb)}{it}\varphi(t)\,dt \\
&= \lim_{c\to\infty}\Phi(c) \\
&= F(b)-F(a). && \text{by part 1)}
\end{aligned}$$

Now let $a<b$ with F not necessarily continuous at a and b. Because there are countably many discontinuities of F, there exist decreasing sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ that converge to a and b from above, such that F is continuous at each a_n and b_n. By absolute continuity of f, we have

$$F(b)-F(a)=\lim_{n\to\infty}(F(b_n)-F(a_n))=\lim_{n\to\infty}\int_{a_n}^{b_n} f(x)\,dx=\int_a^b f(x)\,dx,$$

so f is a density for F. □

Corollary 4.60. *The distribution of a random variable is uniquely determined by its characteristic function.*

We remark that there is also an inversion formula for random d-dimensional vectors, and thus the distribution of a d-dimensional random variable is uniquely determined by its characteristic function.

Corollary 4.61. *A random vector $X := (X_1,\dots,X_d)$ has independent components if and only if $\varphi_X(u) = \varphi_{X_1}(u_1)\cdots\varphi_{X_d}(u_d)$.*

Proof. If X has independent components, then

$$\varphi_X(u) = \mathrm{E}[e^{iu^T X}] = \mathrm{E}[e^{iu_1 X_1}] \cdots \mathrm{E}[e^{iu_d X_d}] = \varphi_{X_1}(u_1) \cdots \varphi_{X_d}(u_d).$$

Conversely, suppose $\varphi_X(u) = \varphi_{X_1}(u_1) \cdots \varphi_{X_d}(u_d)$. Let $Y_1, \dots Y_d$ be independent random variables with characteristic functions $\varphi_{X_1}, \dots, \varphi_{X_d}$ (i.e., $X_k \overset{d}{=} Y_k$ for each k), and let $Y := (Y_1, \dots, Y_d)$. Then $\varphi_X = \varphi_Y$, so by the inversion formula, $X \overset{d}{=} Y$. □

Lemma 4.62. *If $X \sim \mathcal{N}(0,1)$, then its characteristic function is $\varphi_X(u) = e^{-u^2/2}$.*

Proof. If $v \in \mathbb{R}$, then

$$\begin{aligned} \mathrm{E}[e^{vX}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{vx} \, dx \\ &= \frac{e^{v^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-v)^2/2} \, dx \\ &= e^{v^2/2}. \end{aligned}$$

Now let $v \in \mathbb{C}$. The function $e^{v^2/2}$ is analytic. The function $\mathrm{E}[e^{vX}]$ is analytic, since its derivative is $\mathrm{E}[Xe^{vX}]$ (by the dominated convergence theorem). Since the two functions agree for real v, they agree for all complex v by analytic continuation. Letting $v = iu$ proves the lemma.

Note that the same proof can be adapted to show that $\varphi_X(u) = \exp(iu\mu - \frac{1}{2}\sigma^2 u^2)$ if $X \sim \mathcal{N}(\mu, \sigma^2)$. □

Example 4.63. If $X, Y \sim \mathcal{N}(0,1)$ are independent, then what is the distribution of $X + Y$? One approach is to directly find the pdf.

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, dx = \cdots = \frac{1}{2\sqrt{\pi}} e^{-z^2/4},$$

so $X + Y \sim \mathcal{N}(0, 2)$.

However, it is much easier to consider the characteristic functions and apply the inversion theorem (Theorem 4.59).

$$\mathrm{E}[e^{iu(X+Y)}] = \mathrm{E}[e^{iuX}] \, \mathrm{E}[e^{iuY}] = e^{-u^2}.$$

The first equality is due to independence of X and Y.

Theorem 4.64 (Continuity theorem)**.** *Let $X_1, X_2, \dots$ be random variables on probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ respectively, with $\varphi_n := \varphi_{X_n}$.*

1) *If $X_n \xrightarrow{d} X$ for some random variable X, then $\varphi_n \to \varphi_X$ pointwise for every $u \in \mathbb{R}$.*

2) *If $\lim_{n\to\infty} \varphi_n(u)$ exists for all $u \in \mathbb{R}$ and the limit function $\varphi(u) := \lim_{n\to\infty} \varphi_n(u)$ is continuous at $u = 0$, then φ is the characteristic function of a random variable X, and $X_n \xrightarrow{d} X$.*

Proof. If the X_n converge in distribution to X, then by definition $\mathrm{E}[f(X_n)] \to \mathrm{E}[f(X)]$ for any continuous bounded function f. The first statement then follows immediately by writing $\mathrm{E}[e^{iu^T X}] = \mathrm{E}[\cos(u^T X)] + i\,\mathrm{E}[\sin(u^T X)]$ and noting that sin and cos are continuous and bounded.

For the second statement, see §18.1 in *Probability with Martingales* by David Williams. □

Theorem 4.65 (Central limit theorem)**.** *Let $X_1, X_2, \dots$ be i.i.d. (independently and identically distributed) random variables on a common probability space $(\Omega, \mathcal{F}, P)$ with $\mathrm{E}[X_1^2] < \infty$ and $\sigma := \sqrt{\mathrm{Var}(X_1)} > 0$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mathrm{E}[X_i]}{\sigma} \xrightarrow{d} Z$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0,1)$.

Proof. Let $Y := (X_1 - \mathrm{E}[X_1])/\sigma$. Then $\mathrm{E}[Y] = 0$ and $\mathrm{E}[Y^2] = 1$. By Theorem 4.57, we have

$$\varphi_Y(u) = 1 - \frac{u^2}{2} + u^2\epsilon(u),$$

where $\lim_{u\to 0}\epsilon(u) = 0$. Then,

$$\varphi_{Y/\sqrt{n}}(u) = \varphi_Y(u/\sqrt{n}) = 1 - \frac{u^2}{2n} + \frac{u^2}{n}\epsilon(u/\sqrt{n}).$$

Let φ_n be the characteristic function of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i - \mathrm{E}[X_i]}{\sigma}$, which is the sum of n i.i.d. random variables with the same distribution as $Y/\sqrt{n}$. Then,

$$\begin{aligned}
\varphi_n(u) &= \left(1 - \frac{u^2}{2n} + \frac{u^2}{n}\epsilon(u/\sqrt{n})\right)^n \\
\log\varphi_n(u) &= n\log\left(1 - \frac{u^2}{2n} + \frac{u^2}{n}\epsilon(u/\sqrt{n})\right) \\
\lim_{n\to\infty}\log\varphi_n(u) &= \lim_{\delta\searrow 0}\frac{1}{\delta}\log\left(1 - \delta\frac{u^2}{2}\right) && \delta := 1/n \\
\lim_{n\to\infty}\log\varphi_n(u) &= \lim_{\delta\searrow 0}\frac{-u^2/2}{1 - \delta\frac{u^2}{2}} && \text{L'Hôpital's rule} \\
\lim_{n\to\infty}\log\varphi_n(u) &= -u^2/2 \\
\lim_{n\to\infty}\varphi_n(u) &= e^{-u^2/2}.
\end{aligned}$$

Applying the continuity theorem (Theorem 4.64) finishes the proof. □

4.8 Normal distributions

Definition 4.66. A $d \times d$ matrix $C = [c_{i,j}]$ is **symmetric** if $c_{i,j} = c_{j,i}$ for all i, j. A symmetric matrix is **positive semidefinite** if $u^T C u \geq 0$ for all $u \in \mathbb{R}^d$. A symmetric matrix is **positive definite** if $u^T C u > 0$ for all $u \in \mathbb{R}^d \setminus \{0\}$.

Lemma 4.67. *By the spectral theorem, a symmetric $d \times d$ real matrix C has real eigenvalues $\lambda_1, \ldots, \lambda_d$. There exists a matrix U such that $U^T U = U U^T = I_d$ (orthogonal) and such that*

$$U^T C U = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}.$$

Lemma 4.68.

- *A symmetric matrix is positive semidefinite if and only if its eigenvalues are nonnegative.*
- *A symmetric matrix is positive definite if and only if its eigenvalues are strictly positive.*
- *A symmetric matrix is positive definite if and only if it is positive semidefinite and invertible.*

Definition 4.69. Let X be a d-dimensional random vector with square-integrable components. Let $\mathrm{E}[X] := (\mathrm{E}[X_1], \ldots, \mathrm{E}[X_d])^T \in \mathbb{R}^d$ and let $\mathrm{Cov}(X) := [\mathrm{Cov}(X_i, X_j)] \in \mathbb{R}^{d\times d}$.

Lemma 4.70. $\mathrm{Cov}(X)$ *is symmetric and positive definite.*

Proof.

$$\begin{aligned} u^T \operatorname{Cov}(X) u &= \sum_{i,j} u_i u_j \operatorname{Cov}(X_i, X_j) \\ &= \operatorname{Cov}\left(\sum_{i=1}^d u_i X_i, \sum_{j=1}^d u_j X_j \right) \\ &= \operatorname{Var}\left(\sum_{i=1}^d u_i X_i \right) \\ &\geq 0. \end{aligned}$$

□

Lemma 4.71. $\operatorname{Cov}(X)$ *is positive definite if and only if* $1, X_1, \ldots, X_d$ *are linearly independent in* L^2.

Proof. $u^T \operatorname{Cov}(X) u > 0$ for all $u \in \mathbb{R}^d \setminus \{0\}$ if and only if $\operatorname{Var}(u^T X) > 0$ for all $u \in \mathbb{R}^d \setminus \{0\}$, if and only if

$$P\left(u_0 + \sum_{i=1}^d u_i X_i = 0 \right) < 1$$

for all $(u_0, u_1, \ldots, u_d) \in \mathbb{R}^{d+1} \setminus \{0\}$. □

Definition 4.72. We call a d-dimensional random variable X **normal** or **Gaussian** if $\varphi_X(u) = \exp(iu^T \mu - \frac{1}{2} u^T C u)$ for $u \in \mathbb{R}^d$, C a symmetric positive semidefinite matrix. We denote this by $X \sim \mathcal{N}_d(\mu, C)$.

We say X is **regular normal** if C is invertible, and **degenerate normal** otherwise.

Note that normal random vectors are completely characterized by their first and second moments.

Example 4.73. Consider the random vector (Z, Z) where $Z \sim \mathcal{N}(0,1)$. Its covariance matrix is $C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, which is not invertible. The distribution lies completely in a one-dimensional subspace of $\mathbb{R}^2$ so there is no density.

Lemma 4.74. *Let* $\rho(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. *Then*

$$f(x_1, \ldots, x_d) := \rho(x_1) \cdots \rho(x_d)$$

is the density of a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. *Then* $X : \mathbb{R}^d \to \mathbb{R}^d$ *defined to be the identity map is a random vector whose components are independent standard normal. Its characteristic function decomposes as*

$$\varphi_X(u) = \operatorname{E}[e^{iu^T X}] = \operatorname{E}[e^{iu_1 X_1}] \cdots \operatorname{E}[e^{iu_d X_d}] = e^{-u^T u/2}.$$

The following proposition shows that there exists a random vector following the distribution $\mathcal{N}_d(\mu, C)$ for any choice of $\mu \in \mathbb{R}^d$ and symmetric positive semidefinite $C \in \mathbb{R}^{d \times d}$.

Proposition 4.75. *Let* $\mu \in \mathbb{R}^d$ *and* $C \in \mathbb{R}^{d \times d}$ *be symmetric and positive semidefinite.*

1) There exists a symmetric positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$ *such that* $A^2 = C$. *If* $Z \sim \mathcal{N}_d(0, I_d)$, *then* $X := \mu + AZ \sim \mathcal{N}_d(\mu, C)$, *with* $\operatorname{E}[X] = \mu$ *and* $\operatorname{Cov}(X) = C$.

2) The components of X *are independent if and only if* $\operatorname{Cov}(X_i, X_j) = 0$ *for all* $i \neq j$.

3) If C *is invertible, then* X *has density*

$$\frac{1}{(2\pi)^{d/2} \sqrt{\det C}} \exp\left(-\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu) \right).$$

4) If C is not invertible, then $A\mathbb{R}^d$ is a strict subspace of $\mathbb{R}^d$ and X cannot have a density.

5) For every $v \in \mathbb{R}^k$ and $M \in \mathbb{R}^{k\times d}$, we have $Y := v + MX \sim \mathcal{N}_k(v + M\mu, MCM^T)$.

Proof. Let U be an orthogonal matrix such that $D := U^TCU$ is a diagonal matrix whose diagonal entries are the eigenvalues of C (Lemma 4.67), which are nonnegative because C is positive semidefinite. We may take the square root of these diagonal entries to obtain another diagonal matrix denoted $\sqrt{D}$. Letting $A := U\sqrt{D}U^T$ gives a symmetric positive semidefinite matrix satisfying $A^2 = U\sqrt{D}U^TU\sqrt{D}U^T = UDU^T = C$.

It is clear that $\mathrm{E}[X] = \mu$. Moreover, because the components of Z are independent, we have

$$\mathrm{Cov}(X) = \mathrm{E}[(X - \mathrm{E}\,X)(X - \mathrm{E}\,X)^T] = \mathrm{E}[AZZ^TA^T] = A\,\mathrm{E}[ZZ^T]A^T = AI_dA^T = AA^T = C.$$

Finally,

$$\varphi_X(u) = e^{iu^T\mu}\,\mathrm{E}[e^{iu^TAZ}] = e^{iu^T\mu}\varphi_Z(A^Tu) = \exp\Big(iu^T\mu - \frac{1}{2}u^TAA^Tu\Big) = \exp\Big(iu^T\mu - \frac{1}{2}u^TCu\Big),$$

proving 1).

To prove 2), note that $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$ if and only if C is diagonal, if and only if

$$\varphi_X(u) = e^{iu^T\mu - \frac{1}{2}u^TCu} = \varphi_{X_1}(u_1)\cdots\varphi_{X_d}(u_d),$$

if and only if the components are independent (Corollary 4.61).

Note that C is invertible if and only if A is invertible. To prove 3), note that for any $B \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}P(X \in B) &= P(Z \in A^{-1}(B - \mu))\\ &= \int_{A^{-1}(B-\mu)} \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}z^Tz}\,dz\\ &= \int_B \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}(x-\mu)^TC^{-1}(x-\mu)\Big)\frac{1}{\sqrt{\det C}}\,dx,\end{aligned}$$

with the change of variables $z = A^{-1}(x - \mu)$ and $dz = dx/\sqrt{\det C}$.

To prove 4), note that A is not invertible, so $A\mathbb{R}^d$ is a strict subspace of $\mathbb{R}^d$ which therefore has zero Lebesgue measure. If X had density f, then

$$P(X \in B) = \int_{B\cap(\mu + A\mathbb{R}^d)} f(x)\,dx = 0$$

for all $B \in \mathcal{B}(\mathbb{R}^d)$, a contradiction.

Finally, 5) follows using the argument and result from 1).

$$\varphi_Y(u) = e^{iu^Tv}\,\mathrm{E}[e^{iu^TMX}] = e^{iu^Tv}\varphi_X(M^Tu) = \exp\Big(iu^Tv + iu^TM\mu - \frac{1}{2}u^TMCM^Tu\Big).$$

□

Proposition 4.76. *A d-dimensional random vector X is normal if and only if v^TX is [one-dimensional] normal for all $v \in \mathbb{R}^d$. [Note that we allow one-dimensional normal distributions to have zero variance, i.e., point masses.]*

Proof. If $X \sim \mathcal{N}_d(\mu, C)$, then for any $v \in \mathbb{R}^d$, we have

$$\varphi_{v^TX}(u) = \mathrm{E}[e^{iuv^TX}] = \exp\Big(iuv^T\mu - \frac{1}{2}u^2v^TCv\Big),$$

which implies $v^TX \sim \mathcal{N}_1(v^T\mu, v^TCv)$.

Conversely, if v^TX is normal for all $v \in \mathbb{R}^d$, then X is square-integrable. Let $\mu := \mathrm{E}[X]$ and $C := \mathrm{Cov}(X)$. Then noting that the characteristic function corresponding to $Y \sim \mathcal{N}_1(a, b)$ is $\varphi_Y(w) = \exp\big(iwa - \frac{1}{2}w^2b\big)$, we have

$$\varphi_X(u) = \mathrm{E}[e^{iu^TX}] = \varphi_Y(1) = \exp\Big(iu^T\mu - \frac{1}{2}u^TCu\Big)$$

because $a = \mathrm{E}[u^TX] = u^T\mu$ and $b = \mathrm{Var}(u^TX) = u^TCu$. □

4.9 Gaussian processes

Definition 4.77. Let I be a nonempty set. A family $(X_i)_{i\in I}$ of random variables on a probability space is called a **Gaussian process** if $(X_{i_1}, \dots, X_{i_d})$ is d-dimensional normal for any d and any d-tuple $(i_1, \dots, i_d)$ of distinct elements of I.

Definition 4.78. A function $C : I^2 \to \mathbb{R}$ is **symmetric** if $C_{i,j} = C_{j,i}$ for all $i, j \in I$. Such a function is **positive semidefinite** if for any d the matrix $[C_{i_k,i_\ell}]_{k,\ell=1}^d \in \mathbb{R}^{d\times d}$ is positive semidefinite for any d-tuple $(i_1, \dots, i_d)$ of different elements of I. Positive definiteness is defined analogously.

Lemma 4.79. *If $(X_i)_{i\in I}$ is a Gaussian process, then let $\mu_i^X := \mathrm{E}[X_i]$ and $C_{i,j}^X := \mathrm{Cov}(X_i, X_j)$. Then C^X is symmetric and positive semidefinite.*

Theorem 4.80. *Let I be nonempty, and fix functions $\mu : I \to \mathbb{R}$ and $C : I^2 \to \mathbb{R}^2$ with C symmetric and positive semidefinite. Then there exists a Gaussian process $(X_i)_{i\in I}$ with mean μ and covariance C.*

Proof. For every tuple $(i_1, \dots, i_d)$ of different elements of I, let $P^{i_1,\dots,i_d}$ be the probability measure associated with the distribution

$$\mathcal{N}\left(\begin{bmatrix} u_{i_1} \\ \vdots \\ u_{i_d} \end{bmatrix}, \begin{bmatrix} c_{i_1,i_1} & \cdots & c_{i_1,i_d} \\ \vdots & \ddots & \vdots \\ c_{i_d,i_1} & \cdots & c_{i_d,i_d} \end{bmatrix}\right).$$

This family is consistent, so applying the Kolmogorov Extension Theorem (Theorem 3.46) finishes the proof. □

Example 4.81 (White noise)**.** For any nonempty set I there exists a Gaussian process $(X_i)_{i\in I}$ such that $\mathrm{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1$ for all $i \in I$, and such that $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$.

Note that if $I = \mathbb{N}$ for example, then a realization of $(X_i)_{i\in I}$ would appear to "jump around," which presents no problem because the topology on $\mathbb{N}$ is discrete. However, if $I = \mathbb{R}_+$ for example, a realization of $(X_i)_{i\in I}$ does not necessarily have **path regularity**; it would "jump around" and not be continuous.

Example 4.82 (Brownian motion)**.** There exists a Gaussian process $(X_t)_{t\in\mathbb{R}_+}$ with $\mathrm{E}[X_t] = 0$ and $\mathrm{Cov}(X_t, X_s) = t \wedge s$. To justify this, note that we need to verify that $t \wedge s$ is a positive semi-definite function. Given a tuple $(t_1, \dots, t_n) \in \mathbb{R}^n$, we may assume without loss of generality that $0 \le t_1 < t_2 < \cdots < t_n$ because the positive semidefiniteness of the matrix generated by this tuple (in the definition of positive semidefinite function) does not change when permuting the components of the tuple.

Note that the matrix generated by this tuple is

$$\begin{bmatrix} t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & \cdots & t_2 \\ t_1 & t_2 & \cdots & t_3 \\ \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & \cdots & t_n. \end{bmatrix}.$$

One could prove the positive semidefiniteness of this matrix by working with the matrix directly, but we provide an indirect approach instead.

Suppose we had a Hilbert space $\mathcal{H}$ that contained elements $f_1, \dots, f_n$ such that $\langle f_i, f_j \rangle = t_i \wedge t_j$ for each i, j. Then we can immediately see that the matrix is positive semidefinite.

$$\sum_{i=1}^n \sum_{j=1}^n u_i u_j (t_i \wedge t_j) = \sum_{i=1}^n \sum_{j=1}^n u_i u_j \langle f_i, f_j \rangle = \left\langle \sum_{i=1}^n u_i f_i, \sum_{j=1}^n u_j f_j \right\rangle \geq 0.$$

To achieve this, let $\mathcal{H} := L^2(\mathbb{R})$ and let $f_i := \mathbf{1}_{[0,t_i]}$ for each i. Then

$$\langle f_i, f_j \rangle := \int_{-\infty}^{\infty} f_i(x) f_j(x)\, dx = t_i \wedge t_j,$$

as desired.

We now observe some properties of this process.

- $X_0 = 0$ almost surely. To see this, note that $\mathrm{E}[X_0^2] = 0 \wedge 0 = 0$.
- **Stationary increments.** For $t > s$, we have $X_t - X_s \sim \mathcal{N}(0, t-s)$; note that this depends only on the length of the increment $[s,t]$ and not its location (stationarity). To show this note that $\mathrm{E}[X_t - X_s] = 0$, that
$$\mathrm{E}[(X_t - X_s)^2] = \operatorname{Cov}(X_t, X_t) - 2\operatorname{Cov}(X_t, X_s) + \operatorname{Cov}(X_s, X_s) = t - 2s + s = t - s,$$
and that $X_t - X_s$ is normal due Proposition 4.76 and the definition of a Gaussian process.
- **Independent increments.** For $t > s > v > u$, the random variables $X_t - X_s$ and $X_v - X_u$ are independent. To show this, note that
$$\operatorname{Cov}(X_t - X_s, X_v - X_u) = \operatorname{Cov}(X_t, X_v) - \operatorname{Cov}(X_s, X_v) - \operatorname{Cov}(X_t, X_u) + \operatorname{Cov}(X_s, X_u) = v - v - u + u = 0.$$

Note that despite these nice properties of the distributions, a realization $(X_t(\omega))_{t \in \mathbb{R}_+}$ is not necessarily path regular; it can "jump around." It can be shown that there exists a Gaussian process $(B_t)_{t \in \mathbb{R}_+}$ with continuous paths such that $B_t = X_t$ almost surely for each $t \in \mathbb{R}_+$. This is a nontrivial result and we omit its verification. This process $(B_t)_{t \in \mathbb{R}_+}$ is called a **Brownian motion**.

A **Lévy process** is a process with stationary and independent increments. Brownian motion is the only type of Lévy process that has continuous paths.

5 Martingales

5.1 Conditional expectation

Definition 5.1. Let μ_1 and μ_2 be two measures on a measurable space $(\Omega, \mathcal{F})$. We say μ_2 is **absolutely continuous** with respect to μ_1 (denoted $\mu_2 \ll \mu_1$) if for any $A \in \mathcal{F}$ such that $\mu_1(A) = 0$, we also have $\mu_2(A) = 0$. We say μ_1 and μ_2 are **equivalent** (denoted $\mu_1 \sim \mu_2$) if $\mu_1 \ll \mu_2$ and $\mu_2 \ll \mu_1$.

Lemma 5.2. *If $f : \Omega \to \mathbb{R}_+ \cup \{\infty\}$ is a measurable function on the measure space $(\Omega, \mathcal{F}, \mu_1)$, then*
$$\mu_2(A) := \int_A f \, d\mu_1$$
is a measure that is absolutely continuous with respect to μ_1. If f is integrable, then μ_2 is finite.

Theorem 5.3 (Radon-Nikodym)**.** *Let μ_1 and μ_2 be measures on a measurable space $(\Omega, \mathcal{F})$ such that $\mu_2 \ll \mu_1$ and μ_1 is σ-finite. Then there exists a measurable function $f : \Omega \to \mathbb{R}_+ \cup \{\infty\}$ such that*
$$\mu_2(A) = \int_A f \, d\mu_1$$
for all $A \in \mathcal{F}$. This function f is unique up to μ_1-a.e. equality.

Proposition 5.4 (Conditional expectation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra, and let $X \in L^1(\Omega, \mathcal{F}, P)$. Then there exists a unique $Y \in L^1(\Omega, \mathcal{G}, P)$ such that*
$$\mathrm{E}[\mathbf{1}_A Y] = \mathrm{E}[\mathbf{1}_A X]$$
for all $A \in \mathcal{G}$. We often denote Y by $\mathrm{E}[X \mid \mathcal{G}]$. Note that uniqueness in $L^1(\Omega, \mathcal{G}, P)$ allows for P-almost everywhere equality.

Proof. Suppose first that X is a nonnegative random variable. Then $\mu(A) := \mathrm{E}[\mathbf{1}_A X]$ is a finite measure on $\mathcal{G}$ that is absolutely continuous with respect to P. By the Radon-Nikodym theorem (Theorem 5.3), there exists a random variable $Y \in L^1(\Omega, \mathcal{G}, P)$ such that $\mu(A) = \mathrm{E}[\mathbf{1}_A Y]$ for all $A \in \mathcal{G}$, which shows the existence of $\mathrm{E}[X \mid \mathcal{G}]$ when $X \geq 0$.

If X is instead an arbitrary random variable in $L^1(\Omega, \mathcal{F}, P)$, then by the Radon-Nikodym theorem again we have $Y_1, Y_2 \in L^1(\Omega, \mathcal{G}, P)$ such that $\mathrm{E}[\mathbf{1}_A Y_1] = \mathrm{E}[\mathbf{1}_A X^+]$ and $\mathrm{E}[\mathbf{1}_A Y_2] = \mathrm{E}[\mathbf{1}_A X^-]$ for all $A \in \mathcal{G}$. Then $\mathrm{E}[\mathbf{1}_A (Y_1 - Y_2)] = \mathrm{E}[\mathbf{1}_A X]$, which shows the existence of $\mathrm{E}[X \mid \mathcal{G}]$.

To show uniqueness, suppose $Y, Z \in L^1(\Omega, \mathcal{G}, P)$ satisfy $\mathrm{E}[\mathbf{1}_A Y] = \mathrm{E}[\mathbf{1}_A Z] = \mathrm{E}[\mathbf{1}_A X]$ for all $A \in \mathcal{G}$. Then because $\{Y > Z\}$ and $\{Y < Z\}$ are in $\mathcal{G}$, this equality implies

$$\mathrm{E}\big[\mathbf{1}_{\{Y>Z\}}(Y-Z)\big] = 0,$$
$$\mathrm{E}\big[\mathbf{1}_{\{Y<Z\}}(Z-Y)\big] = 0.$$

However, these are expectations of nonnegative random variables, so they are both zero only if $Y = Z$ almost surely. □

A random variable that equals $\mathrm{E}[X \mid \mathcal{G}]$ almost everywhere is called a **version** of $\mathrm{E}[X \mid \mathcal{G}]$.

Definition 5.5. Let $X \in L^1(\Omega, \mathcal{F}, P)$, let Z be a random variable on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then we define

$$\mathrm{E}[X \mid Z] := \mathrm{E}[X \mid \sigma(Z)],$$

where we recall that $\sigma(Z) := \{Z^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$. For $A \in \mathcal{F}$, we define

$$P(A \mid \mathcal{G}) := \mathrm{E}[\mathbf{1}_A \mid \mathcal{G}],$$
$$P(A \mid Z) := \mathrm{E}[\mathbf{1}_A \mid \sigma(Z)].$$

Proposition 5.6. *Let $X, Y \in L^1(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra.*

a) If X is $\mathcal{G}$-measurable, then $\mathrm{E}[X \mid \mathcal{G}] = X$. In particular, $\mathrm{E}[c \mid \mathcal{G}] = c$ for any constant $c \in \mathbb{R}$.

b) $\mathrm{E}[aX + Y \mid \mathcal{G}] = a\,\mathrm{E}[X \mid \mathcal{G}] + \mathrm{E}[Y \mid \mathcal{G}]$ for all $a \in \mathbb{R}$.

c) $\mathrm{E}[X \mid \mathcal{G}] \geq \mathrm{E}[Y \mid \mathcal{G}]$ a.s. if $X \geq Y$ a.s.

*d) **Tower property.** $\mathrm{E}[\mathrm{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathrm{E}[X \mid \mathcal{H}]$ for every sub-σ-algebra $\mathcal{H}$ of $\mathcal{G}$.*

e) If Y is $\mathcal{G}$-measurable and $XY \in L^1$, then $\mathrm{E}[XY \mid \mathcal{G}] = Y\,\mathrm{E}[X \mid \mathcal{G}]$.

f) If X is independent of $\mathcal{G}$ (i.e., the σ-algebras $\sigma(X)$ and $\mathcal{G}$ are independent), then $\mathrm{E}[X \mid \mathcal{G}] = \mathrm{E}[X]$.

g) If $\mathcal{G} = \{\varnothing, \Omega\}$, then $\mathrm{E}[X \mid \mathcal{G}] = \mathrm{E}[X]$.

h) If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex such that $\varphi(X) \in L^1$, then $\mathrm{E}[\varphi(X) \mid \mathcal{G}] \geq \varphi(\mathrm{E}[X \mid \mathcal{G}])$ a.s.

Proposition 5.7. *Let $(\Omega, \mathcal{F}, P)$ be a probability space carrying two σ-algebras $\mathcal{G}$ and $\mathcal{H}$. Then,*

$$\mathrm{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] = \mathrm{E}[X \mid \mathcal{G}]$$

for every $X \in L^1(\Omega, \mathcal{F}, P)$ such that $\sigma(X, \mathcal{G})$ is independent of $\mathcal{H}$.

Definition 5.8. Let X be an extended random variable (taking values in $\mathbb{R} \cup \{\pm\infty\}$) on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. If $\mathrm{E}[X^-] < \infty$, we define

$$\mathrm{E}[X \mid \mathcal{G}] := \lim_{k\to\infty} \mathrm{E}[X \wedge k \mid \mathcal{G}].$$

If $\mathrm{E}[X^+] < \infty$, we define

$$\mathrm{E}[X \mid \mathcal{G}] := \lim_{k\to-\infty} \mathrm{E}[X \vee k \mid \mathcal{G}].$$

Theorem 5.9. *Let $X_1, X_2, \dots$ be a sequence in $L^1(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra.*

a) (Conditional version of Beppo Levi's monotone convergence theorem)
If there exists $Y \in L^1$ and a random variable with values in $\mathbb{R} \cup \{\infty\}$ such that $Y \leq X_n \nearrow X$ a.s., then $\mathrm{E}[X_n \mid \mathcal{G}] \nearrow \mathrm{E}[X \mid \mathcal{G}]$ a.s.

b) (Conditional version of Fatou's lemma)
If there exists $Y \in L^1$ such that $X_n \geq Y$ a.s. for all n, then

$$\liminf_{n\to\infty} \mathrm{E}[X_n \mid \mathcal{G}] \geq \mathrm{E}\Big[\liminf_{n\to\infty} X_n \mid \mathcal{G}\Big] \text{ a.s.}$$

c) (Conditional version of Lebesgue's dominated convergence theorem)
If there exist $X, Y \in L^1$ such that $|X_n| \leq Y$ for all n and $X_n \to X$ a.s., then

$$\mathrm{E}[X_n \mid \mathcal{G}] \to \mathrm{E}[X \mid \mathcal{G}] \text{ a.s.}$$

5.2 Martingale definitions

Definition 5.10. A **stochastic process** on $(\Omega, \mathcal{F}, P)$ is a family of random variables $(X_i)_{i \in I}$ where I is a nonempty index set. If $I = \mathbb{N}$, we call $(X_n)_{n \geq 0}$ a **discrete-time stochastic process**.

Definition 5.11. A sequence $(\mathcal{F}_n)_{n\geq 0}$ of σ-algebras is a **filtration** on $(\Omega, \mathcal{F}, P)$ if $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$.

Definition 5.12. A stochastic process (X_n) is **adapted** to $(\mathcal{F}_n)$ if X_n is $\mathcal{F}_n$-measurable for each $n \geq 0$. It is **predictable** with respect to $(\mathcal{F}_n)$ if X_n is $\mathcal{F}_{n-1}$-measurable for each $n \geq 1$.

Definition 5.13. Let $(X_n)_{n\geq 0}$ be such that $X_n \in L^1(\Omega, \mathcal{F}_n, P)$; note that this implies that (X_n) is adapted to $(\mathcal{F}_n)$.

- (X_n) is a **martingale** with respect to $\mathcal{F}_n$ if $\mathrm{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ for each $n \geq 0$.
- (X_n) is a **submartingale** with respect to $\mathcal{F}_n$ if $\mathrm{E}[X_{n+1} \mid \mathcal{F}_n] \geq X_n$ for each $n \geq 0$.
- (X_n) is a **supermartingale** with respect to $\mathcal{F}_n$ if $\mathrm{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n$ for each $n \geq 0$.

Be wary of the directions of the inequalities in the definitions of submartingales and supermartingales.

Lemma 5.14. *Let $(X_n)_{n\geq 1}$ be a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. If $m \leq n$, then $\mathrm{E}[X_{n+1} \mid \mathcal{F}_m] = X_m$. In particular, $\mathrm{E}[X_{n+1}] = \mathrm{E}[X_0]$.*

Proof. By repeated use of the tower property of conditional expectation (Proposition 5.6),

$$\begin{aligned} \mathrm{E}[X_{n+1} \mid \mathcal{F}_m] &= \mathrm{E}[\mathrm{E}[X_{n+1} \mid \mathcal{F}_n] \mid \mathcal{F}_m] \\ &= \mathrm{E}[X_n \mid \mathcal{F}_m] \\ &\vdots \\ &= \mathrm{E}[X_{m+1} \mid \mathcal{F}_m] \\ &= X_m. \end{aligned}$$

Similarly,

$$\mathrm{E}[X_{n+1}] = \mathrm{E}[\mathrm{E}[X_{n+1} \mid \mathcal{F}_n]] = \mathrm{E}[X_n] = \cdots = \mathrm{E}[X_0].$$

□

Definition 5.15. A stochastic process $(X_n)_{n\geq 0}$ generates the **natural filtration**

$$\mathcal{F}_n^X := \sigma(X_0, X_1, \ldots, X_n) := \sigma(\{X_k^{-1}(B) : B \in \mathcal{B}(\mathbb{R}), 0 \leq k \leq n\}).$$

A stochastic process $(X_n)_{n\geq 0}$ is a **martingale**, **submartingale**, or a **supermartingale** if it is one with respect to its natural filtration $(\mathcal{F}_n^X)$.

Note that whether a stochastic process $(X_n)_{n\geq 0}$ is a martingale depends on the filtration $\mathcal{F}$. Increasing the filtration by "adding more information" may cause a martingale to no longer be a martingale.

Example 5.16.

1) Let $Y_0, Y_1, \ldots$ be a sequence of independent random variables in $L^1(\Omega, \mathcal{F}, P)$ such that $\mathrm{E}[Y_n] = 0$ for all $n \geq 1$. Then $X_n := \sum_{i=1}^n Y_i$ is a martingale because

$$\begin{aligned} \mathrm{E}[X_{n+1} \mid \mathcal{F}_n^X] &= \mathrm{E}[X_n + Y_{n+1} \mid \mathcal{F}_n^X] \\ &= X_n + \mathrm{E}[Y_{n+1}] && X_n \text{ is } \mathcal{F}_n^X\text{-measurable; } Y_{n+1} \text{ is indep. of } \mathcal{F}_n^X \\ &= X_n. \end{aligned}$$

An example is $Y_0 := 0$ and $P(Y_n = 1) = P(Y_n = -1) = 1/2$ for $n \geq 1$; this gives the **standard Bernoulli random walk** $(X_n)_{n\geq 1}$.

2) Let $Y_0, Y_1, \ldots$ be a sequence of independent random variables in $L^1(\Omega, \mathcal{F}, P)$ such that $\mathrm{E}[Y_n] = 0$ for all $n \geq 1$. Then $X_n := \sum_{i=1}^n Y_i$ is a martingale because X_n is integrable due to the independence of the Y_i, and because

$$\begin{aligned} \mathrm{E}[X_{n+1} \mid \mathcal{F}_n^X] &= \mathrm{E}[X_n Y_{n+1} \mid \mathcal{F}_n^X] \\ &= X_n \,\mathrm{E}[Y_{n+1}] && X_n \text{ is } \mathcal{F}_n^X\text{-measurable; } Y_{n+1} \text{ is indep. of } \mathcal{F}_n^X \\ &= X_n. \end{aligned}$$

3) Let $X \in L^1(\Omega, \mathcal{F}, P)$ and let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration. Then $X_n := \mathrm{E}[X \mid \mathcal{F}_n]$ defines a uniformly integrable martingale. [See Corollary 5.47.]

Example 5.17. Let $P(Y_n = 1) = P(Y_n = -1) = 1/2$ for $n \geq 0$ and let $X_n := \sum_{n\geq 0} 2^{-n} Y_n$. Every path will converge. However, although $X_n := \sum_{n\geq 0} \frac{Y_n}{n}$ does not converge everywhere, we can show that it converges an almost all paths.

Definition 5.18. Let $(X_n)_{n\geq 0}$ and $(V_n)_{n\geq 1}$ be two stochastic processes on $(\Omega, \mathcal{F}, P)$. Let $\Delta X_n := X_n - X_{n-1}$ for $n \geq 1$. We define the **martingale transform** of X by V as

$$(V \cdot X)_n := \begin{cases} 0 & n = 0, \\ \sum_{i=1}^n V_i \Delta X_i & n \geq 1. \end{cases}$$

Theorem 5.19. *Let $(X_n)_{n\geq 0}$ be a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Let $(V_n)_{n\geq 1}$ be $\mathcal{F}_n$-predictable and such that*

$$\mathrm{E}[|V_n \Delta X_n|] < \infty$$

for $n \geq 1$. Then $((V \cdot X)_n)_{n\geq 1}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.

Proof. Note that for each fixed $n \geq 1$, $(V \cdot X)_n$ is $\mathcal{F}_n$-measurable because it is formed by adding, subtracting, and multiplying random variables V_i and X_i for $i \leq n$, each of which is $\mathcal{F}_n$-measurable; thus $((V \cdot X)_n)_{n\geq 1}$ is adapted to $(\mathcal{F}_n)_{n\geq 1}$. Similarly, since we are given that $V_n \Delta X_n \in L^1$, we see that $(V \cdot X)_n$ is also in L^1.

Finally,

$$\begin{aligned} &\mathrm{E}[(V \cdot X)_{n+1} \mid \mathcal{F}_n] \\ &= \mathrm{E}[V_{n+1} \Delta X_{n+1} \mid \mathcal{F}_n] + (V \cdot X)_n && V_i, X_i \in \mathcal{F}_n \text{ for } i \leq n \\ &= V_{n+1} \,\mathrm{E}[\Delta X_{n+1} \mid \mathcal{F}_n] + (V \cdot X)_n && V_{n+1} \in \mathcal{F}_n \text{ because } (V_n)_{n\geq 1} \text{ is } (\mathcal{F}_n)_{n\geq 0}\text{-predictable} \\ &= (V \cdot X)_n. && (X_n)_{n\geq 0} \text{ is a martingale} \end{aligned}$$

□

5.3 Stopping times and Doob's optional stopping theorem

Definition 5.20. A **stopping time** with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$ on a probability space is an extended random variable $\tau : (\Omega, \mathcal{F}, P) \to \mathbb{N} \cup \{\infty\}$ such that

$$\{\tau = n\} \in \mathcal{F}_n$$

for each $n \geq 0$.

The intuition is that τ is the a time for stopping the process, and whether or not you stop at time n (the event $\{\tau = n\}$) depends only on the history up to and including time n (the σ-algebra $\mathcal{F}_n$).

Lemma 5.21. *In the definition above, the defining condition "$\{\tau = n\} \in \mathcal{F}_n$ for each $n \geq 0$" can be replaced with the equivalent condition "$\{\tau \leq n\} \in \mathcal{F}_n$ for each $n \geq 0$."*

Proof.

$$\{\tau = n\} = \{\tau \le n\} \setminus \{\tau \le n-1\}$$
$$\{\tau \le n\} = \bigcup_{i=0}^{n} \{\tau = i\}.$$

□

Definition 5.22. A stopping time τ is **finite** if $P(\tau = \infty) = 0$. A stopping time is **bounded** if there exists $N \in \mathbb{N}$ such that $P(\tau \le N) = 1$.

Lemma 5.23. *A stopping time τ defines a* ***stopping σ-algebra***

$$\mathcal{F}_\tau := \{A \in \mathcal{F} : A \cap \{\tau = n\} \in \mathcal{F}_n, \forall n \in \mathbb{N}\}.$$

Note that the sets in $\mathcal{F}_\tau$ do not necessarily belong to any $\mathcal{F}_n$.

Lemma 5.24. *Let τ and σ be stopping times with respect to $(\mathcal{F}_n)_{n\ge 0}$.*

a) $\tau + \sigma$ is a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

b) $\tau \vee \sigma$ is a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

c) $\tau \wedge \sigma$ is a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

d) $\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.

Corollary 5.25. *If $\tau \le \sigma$ are stopping times, then $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$.*

Proof. Note that $\tau \wedge \sigma = \tau$ and use part d) of Lemma 5.24. □

Example 5.26. Constants $\tau \equiv m \in \mathbb{N}$ are stopping times, since $\{\tau = n\}$ is either $\varnothing$ or Ω.

Example 5.27 (Hitting time)**.** Let $(X_n)_{n\ge 0}$ be a stochastic process adapted to $(\mathcal{F}_n)_{n\ge 0}$ and let $B \in \mathcal{B}(\mathbb{R})$. Then

$$\tau := \inf\{n \in \mathbb{N} : X_n \in B\}$$

is a stopping time because

$$\{\tau = n\} = \{X_n \in B\} \cap \bigcap_{i=0}^{n-1} \{X_i \notin B\} \in \mathcal{F}_n.$$

Lemma 5.28. *Let $(X_n)_{n\ge 0}$ be a stochastic process adapted to $(\mathcal{F}_n)_{n\ge 0}$ and let τ be a stopping time with respect to $(\mathcal{F}_n)$. Then $X_\tau \mathbf{1}_{\{\tau<\infty\}}$ is $\mathcal{F}_\tau$-measurable. In particular, $(X_n^\tau)_{n\ge 0}$ is adapted to $(\mathcal{F}_n)$, where $X_n^\tau := X_{n\wedge\tau}$.*

Proof. We would like to show that $\{X_\tau \mathbf{1}_{\{\tau<\infty\}} \le t\}$ is in $\mathcal{F}_\tau$ for any $t \in \mathbb{R}$. Referring to the definition of $\mathcal{F}_\tau$, we see that indeed, for any $n \in \mathbb{N}$,

$$\{X_\tau \mathbf{1}_{\{\tau<\infty\}} \le t\} \cap \{\tau = n\} = \{X_n \le t\} \cap \{\tau = n\} \in \mathcal{F}_n.$$

Let $\sigma := \tau \wedge n$. By Lemma 5.24, it is a stopping time, and moreover, $\mathbf{1}_{\{\sigma<\infty\}} \equiv 1$. By our work above, X_σ is $\mathcal{F}_\sigma$-measurable. By Corollary 5.25, we have $\mathcal{F}_\sigma \subset \mathcal{F}_n$, so X_σ is $\mathcal{F}_n$-measurable. □

Corollary 5.29 (Elementary version of the optional stopping theorem)**.** *Let $(X_n)_{n\ge 0}$ be a martingale and τ a stopping time, both with respect to $(\mathcal{F}_n)_{n\ge 0}$. Then $(X_n^\tau)_{n\ge 0}$ is a martingale. In particular,* $\mathrm{E}[X_\tau] = \mathrm{E}[X_0]$ *if τ is bounded.*

Proof. If $V_n := \mathbf{1}_{\{\tau \geq n\}}$, then

$$X_n^\tau = X_0 + (V \cdot X)_n$$

because for any realization $\tau(\omega) \in \mathbb{N}$ of τ, we have

$$X_0 + (V \cdot X)_n = X_0 + \sum_{i=1}^{\tau(\omega)\wedge n} \Delta X_i = X_{\tau(\omega)\wedge n}.$$

By Theorem 5.19, $(X_n^\tau)_{n\geq 0}$ is a martingale.

If τ is bounded, then there exists $N \in \mathbb{N}$ such that $\tau = \tau \wedge N$. Thus,

$$\mathrm{E}[X_\tau] = \mathrm{E}[X_{\tau\wedge n}] = \mathrm{E}[X_{\tau\wedge 0}] = \mathrm{E}[X_0],$$

where the second equality is due to Lemma 5.14 and the fact that $(X_n^\tau)_{n\geq 0}$ is a martingale. □

Theorem 5.30 (Doob's optional stopping theorem)**.** *Let $(X_n)_{n\geq 0}$ be a martingale and let $\sigma \leq \tau$ be bounded stopping times, all with respect to $(\mathcal{F}_n)_{n\geq 0}$. Then*

$$\mathrm{E}[X_\tau \mid \mathcal{F}_\sigma] = X_\sigma.$$

In particular,

$$\mathrm{E}[X_\tau] = \mathrm{E}[X_0].$$

Proof. Pick $N \in \mathbb{N}$ such that $\sigma \leq \tau \leq N$. We have already shown that X_σ is $\mathcal{F}_\sigma$-measurable (Lemma 5.28). We just need to show

$$\mathrm{E}[X_\tau \mathbf{1}_A] = \mathrm{E}[X_\sigma \mathbf{1}_A]$$

for all $A \in \mathcal{F}_\sigma$.

$$\begin{aligned}
\mathrm{E}[X_\tau \mathbf{1}_A] &= \sum_{n=0}^{N} \mathrm{E}[X_\tau \mathbf{1}_{A\cap\{\sigma=n\}}] \\
&= \sum_{n=0}^{N} \mathrm{E}[X_N^\tau \mathbf{1}_{A\cap\{\sigma=n\}}] \\
&= \sum_{n=0}^{N} \mathrm{E}[\mathrm{E}[X_N^\tau \mathbf{1}_{A\cap\{\sigma=n\}} \mid \mathcal{F}_n]] && \text{tower property (Proposition 5.6)} \\
&= \sum_{n=0}^{N} \mathrm{E}[\mathrm{E}[X_N^\tau \mid \mathcal{F}_n] \mathbf{1}_{A\cap\{\sigma=n\}}] && A \in \mathcal{F}_\sigma \implies A \cap \{\sigma = n\} \in \mathcal{F}_n \\
&= \sum_{n=0}^{N} \mathrm{E}[X_n^\tau \mathbf{1}_{A\cap\{\sigma=n\}}] && \text{Lemma 5.14} \\
&= \sum_{n=0}^{N} \mathrm{E}[X_\sigma^\tau \mathbf{1}_{A\cap\{\sigma=n\}}] && \sigma = n \\
&= \mathrm{E}[X_{\sigma\wedge\tau} \mathbf{1}_A] \\
&= \mathrm{E}[X_\sigma \mathbf{1}_A].
\end{aligned}$$

□

5.4 Doob's decomposition theorem

Theorem 5.31 (Doob's decomposition theorem)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Then there exists a martingale $(M_n)_{n\geq 0}$ and a nondecreasing predictable process $(A_n)_{n\geq 0}$ such that $A_0 = 0$ and*

$$X_n = M_n + A_n.$$

Moreover, this decomposition is unique.

Corollary 5.32. *Let X_n be a submartingale and V_n a nonnegative predictable process, both with respect to $(\mathcal{F}_n)_{n\geq 0}$, such that $\mathrm{E}[|V_n \Delta X_n|] < \infty$. Then $((V \cdot X)_n)_{n\geq 0}$ is a submartingale with Doob's decomposition $((V \cdot M)_n)_{n\geq 0}$ and $((V \cdot A)_n)_{n\geq 0}$.*

5.5 An example: one-dimensional random walk

Let $Y_0, Y_1, \dots$ be independent random variables in L^1. Let $Y_0 = 0$ and

$$P(Y_i = 1) = p, \quad P(Y_i = -1) = 1 - p,$$

for all $i \geq 1$. Let

$$X_n := \sum_{i=0}^{n} Y_i,$$

and define $\mathcal{F}_n := \sigma(Y_0, Y_1, \dots, Y_n)$ so that (X_n) is adapted to $(\mathcal{F}_n)$.

Fix $A, B \in \mathbb{N}$ and define

$$\tau^{A,B} := \inf\{n \in \mathbb{N} : X_n \in \{A, -B\}\}.$$

It is a stopping time because

$$\{\tau^{A,B} = n\} = (\{X_n = A\} \cup \{X_n = -B\}) \cap \bigcap_{i=1}^{n-1} \{X_i \in [-B+1, A-1]\} \in \mathcal{F}_n.$$

Moreover, $\tau^{A,B}$ is finite (see Lemma 5.33 below). Note that $X_{\tau^{A,B}} = A\mathbf{1}_{\{\tau^{A,B} = A\}} - B\mathbf{1}_{\{\tau^{A,B} = -B\}}$. We consider two cases: $p = 1/2$ and $p \neq 1/2$.

If $p = 1/2$, then (X_n) is a martingale. Since $\tau^{A,B} \wedge n$ is a stopping time (Lemma 5.24) that is bounded, Doob's optional stopping theorem (Theorem 5.30) implies $\mathrm{E}[X_{\tau^{A,B} \wedge n}] = \mathrm{E}[X_0] = 0$ for all n. Because $-B \leq X_{\tau^{A,B} \wedge n} \leq A$ for all n, we may apply the dominated convergence theorem (Theorem 4.15) to get $\mathrm{E}[X_{\tau^{A,B}}] = 0$. Solving the system

$$\begin{aligned}
0 &= \mathrm{E}[X_{\tau^{A,B}}] = A \cdot P(X_{\tau^{A,B}} = A) - B \cdot (\tau^{A,B} = -B) \\
1 &= P(X_{\tau^{A,B}} = A) + P(\tau^{A,B} = -B)
\end{aligned}$$

gives

$$P(\tau^{A,B} = A) = \frac{B}{A+B}, \quad P(\tau^{A,B} = -B) = \frac{A}{A+B}.$$

We now compute the expectation of $\tau^{A,B}$ (still in the case $p = 1/2$). First, we note that $(X_n^2 - n)_{n\geq 0}$ is a martingale. It is clearly $(\mathcal{F}_n)$-adapted, and we have

$$\begin{aligned}
\mathrm{E}[X_n^2 - n \mid \mathcal{F}_{n-1}] &= \mathrm{E}[(X_{n-1} + Y_n)^2 - n \mid \mathcal{F}_{n-1}] \\
&= \mathrm{E}[X_{n-1}^2 + 2X_{n-1}Y_n + Y_n^2 - n \mid \mathcal{F}_{n-1}] \\
&= X_{n-1}^2 - n + 2X_{n-1}\,\mathrm{E}[Y_n \mid \mathcal{F}_{n-1}] + \mathrm{E}[Y_n^2 \mid \mathcal{F}_{n-1}] && X_{n-1} \text{ is } \mathcal{F}_{n-1}\text{-measurable} \\
&= X_{n-1}^2 - n + 2X_{n-1}\,\mathrm{E}[Y_n] + \mathrm{E}[Y_n^2] && Y_n \text{ is indep. of } \mathcal{F}_{n-1} \\
&= X_{n-1}^2 - n + 0 + 1 \\
&= X_{n-1}^2 - (n-1).
\end{aligned}$$

We may apply Doob's optimal stopping theorem (Theorem 5.30) to $(X_n^2 - n)$ to get

$$\mathrm{E}[X^2_{\tau^{A,B}\wedge n}] - \mathrm{E}[\tau^{A,B} \wedge n] = \mathrm{E}[X^2_{\tau^{A,B}\wedge n} - \tau^{A,B} \wedge n] = 0.$$

The monotone convergence theorem (Theorem 4.11) gives $\mathrm{E}[\tau^{A,B} \wedge n] \to \mathrm{E}[\tau^{A,B}]$ as $n \to \infty$, and the dominated convergence theorem (Theorem 4.15) gives $\mathrm{E}[X^2_{\tau^{A,B}\wedge n}] \to \mathrm{E}[X^2_{\tau^{A,B}}]$ as $n \to \infty$, so we have

$$\begin{aligned}
\mathrm{E}[\tau^{A,B}] &= \mathrm{E}[X^2_{\tau^{A,B}}] \\
&= A^2 P(X_{\tau^{A,B}} = A) + B^2 P(\tau^{A,B} = -B) \\
&= \frac{A^2 B}{A+B} + \frac{B^2 A}{A+B} \\
&= AB.
\end{aligned}$$

We now examine the hitting times

$$\begin{aligned}
\tau_A &:= \inf\{n \geq 0 : X_n = A\}, \\
\tau_{-B} &:= \inf\{n \geq 0 : X_n = -B\},
\end{aligned}$$

still in the case $p = 1/2$. Note that $\tau^{A,B} = \tau_A \wedge \tau_{-B}$. We have

$$1 \geq P(\tau_A < \infty) \geq P(\tau_A < \tau_{-B}) = P(X_{\tau^{A,B}} = A) = \frac{B}{A+B}$$

for any choice of B. Letting B tend to infinity shows that $P(\tau_A < \infty) = 1$. A similar argument shows that τ_B is also finite.

However,

$$\mathrm{E}[\tau_A] \geq \mathrm{E}[\tau^{A,B}] = AB$$

for any choice of B. Letting B tend to infinity shows that $\mathrm{E}[\tau_A] = \infty$; similarly, $\mathrm{E}[\tau_B] = \infty$. Thus, τ_A and τ_B are examples of random variables that are almost everywhere finite but not integrable.

We now consider the case $p \neq 1/2$. We claim $(Z_n)_{n \geq 0}$ is a martingale, where

$$Z_n := \left(\frac{1-p}{p}\right)^{X_n}.$$

Indeed, it is $(\mathcal{F}_n)$-adapted, and

$$\begin{aligned}
\mathrm{E}[Z_n \mid \mathcal{F}_{n-1}] &= \mathrm{E}\left[\left(\frac{1-p}{p}\right)^{X_n} \,\middle|\, \mathcal{F}_{n-1}\right] \\
&= \left(\frac{1-p}{p}\right)^{X_{n-1}} \mathrm{E}\left[\left(\frac{1-p}{p}\right)^{Y_n} \,\middle|\, \mathcal{F}_{n-1}\right] && X_{n-1} \text{ is } \mathcal{F}_{n-1}\text{-measurable} \\
&= \left(\frac{1-p}{p}\right)^{X_{n-1}} \mathrm{E}\left[\left(\frac{1-p}{p}\right)^{Y_n}\right] && Y_n \text{ is indep. of } \mathcal{F}_{n-1} \\
&= Z_{n-1}\left(\frac{1-p}{p} \cdot p + \frac{p}{1-p} \cdot (1-p)\right) \\
&= Z_{n-1}.
\end{aligned}$$

By Doob's optional stopping theorem (Theorem 5.30), $1 = \mathrm{E}[Z_0] = \mathrm{E}[Z_{\tau^{A,B}\wedge n}]$ for all n. By the dominated convergence theorem (Theorem 4.15), we have $\mathrm{E}[Z_{\tau^{A,B}}] = 1$. Solving the system

$$\begin{aligned}
1 &= \mathrm{E}[Z_{\tau^{A,B}}] = \left(\frac{1-p}{p}\right)^{A} P(X_{\tau^{A,B}} = A) + \left(\frac{1-p}{p}\right)^{-B} P(\tau^{A,B} = -B) \\
1 &= P(X_{\tau^{A,B}} = A) + P(\tau^{A,B} = -B)
\end{aligned}$$

gives

$$P(X_{\tau^{A,B}} = A) = \frac{\left(\frac{1-p}{p}\right)^B - 1}{\left(\frac{1-p}{p}\right)^{A+B} - 1}, \quad P(X_{\tau^{A,B}} = -B) = \frac{\left(\frac{1-p}{p}\right)^A - 1}{\left(\frac{1-p}{p}\right)^{A+B} - 1}.$$

We now compute the expectation of $\tau^{A,B}$. Note that $(W_n)_{n\geq 0}$ is a martingale, where

$$W_n := X_n - n(2p-1)$$

because

$$\begin{aligned}
\mathrm{E}[W_n \mid \mathcal{F}_{n-1}] &= \mathrm{E}[X_n - n(2p-1) \mid \mathcal{F}_{n-1}] \\
&= X_{n-1} - n(2p-1) + \mathrm{E}[Y_n \mid \mathcal{F}_{n-1}] && X_{n-1} \text{ is } \mathcal{F}_{n-1}\text{-measurable} \\
&= X_{n-1} - n(2p-1) + \mathrm{E}[Y_n] && Y_n \text{ is independent of } \mathcal{F}_{n-1} \\
&= X_{n-1} - n(2p-1) + (p - (1-p)) \\
&= X_{n-1} - (n-1)(2p-1) \\
&= W_{n-1}.
\end{aligned}$$

Again by Doob's optional stopping theorem (Theorem 5.30), we have

$$\mathrm{E}[X_{\tau^{A,B}\wedge n}] - (2p-1)\,\mathrm{E}[\tau^{A,B} \wedge n] = \mathrm{E}[X_{\tau^{A,B}\wedge n} - (2p-1)(\tau^{A,B} \wedge n)] = 0.$$

As before, the dominated convergence theorem (Theorem 4.15) and the monotone convergence theorem (Theorem 4.11) give $\mathrm{E}[X_{\tau^{A,B}\wedge n}] \to \mathrm{E}[X_{\tau^{A,B}}]$ and $\mathrm{E}[\tau^{A,B} \wedge n] \to \mathrm{E}[\tau^{A,B}]$ respectively, so we have

$$\mathrm{E}[\tau^{A,B}] = \frac{1}{2p-1}\,\mathrm{E}[X_{\tau^{A,B}}] = \frac{1}{2p-1}\left(A\frac{\left(\frac{1-p}{p}\right)^B - 1}{\left(\frac{1-p}{p}\right)^{A+B} - 1} - B\frac{\left(\frac{1-p}{p}\right)^A - 1}{\left(\frac{1-p}{p}\right)^{A+B} - 1}\right).$$

Finally, we examine the hitting times τ_A and τ_{-B}. Note that $\tau_{-B} \geq B$ because it takes at least B steps to reach $-B$ from the origin. Thus, τ_{-B} tends to infinity with B. This implies

$$\begin{aligned}
P(\tau_A < \infty) &= \lim_{B\to\infty} P(\tau_A < \tau_{-B}) \\
&= \lim_{B\to\infty} P(\tau^{A,B} = A) \\
&= \lim_{B\to\infty} \frac{\left(\frac{1-p}{p}\right)^B - 1}{\left(\frac{1-p}{p}\right)^{A+B} - 1} \\
&= \begin{cases} 1 & p > 1/2, \\ \left(\frac{p}{1-p}\right)^A & p < 1/2. \end{cases}
\end{aligned}$$

Lemma 5.33. *The stopping time $\tau^{A,B}$ defined above is finite, i.e., $P(\tau^{A,B} < \infty) = 1$.*

5.6 Doob's upcrossing inequality

Proposition 5.34. *Let $(X_n)_{n\geq 0}$ be a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\varphi(X_n) \in L^1$ for each n. If at least one of the conditions below holds, then $(\varphi(X_n))_{n\geq 0}$ is a submartingale.*

a) X_n is a martingale.

b) φ is nondecreasing.

Proof. The conditional version of Jensen's inequality gives

$$\mathrm{E}[\varphi(X_{n+1}) \mid \mathcal{F}_n] \geq \varphi(\mathrm{E}[X_{n+1} \mid \mathcal{F}_n]).$$

If a) holds, then $\varphi(\mathrm{E}[X_{n+1} \mid \mathcal{F}_n]) = \varphi(X_n)$, so $(\varphi(X_n))_{n\geq 0}$ is a submartingale. If b) holds, then $\varphi(\mathrm{E}[X_{n+1} \mid \mathcal{F}_n]) \geq \varphi(X_n)$ because $(X_n)_{n\geq 0}$ is a submartingale and because φ is nondecreasing; thus $(\varphi(X_n))_{n\geq 0}$ is a submartingale. □

We now provide the setup for the theorem. Let $(X_n)_{n\geq 0}$ be a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$, and let $a < b$ be real numbers. We define

$$\begin{aligned}
\tau_1 &:= \inf\{n \geq 0 : X_n \leq a\},\\
\tau_2 &:= \inf\{n > \tau_1 : X_n \geq b\},\\
\tau_3 &:= \inf\{n > \tau_2 : X_n \leq a\},\\
&\vdots\\
\tau_{2k} &:= \inf\{n > \tau_{2k-1} : X_n \geq b\},\\
\tau_{2k+1} &:= \inf\{n > \tau_{2k} : X_n \leq a\}.
\end{aligned}$$

Lemma 5.35. *The τ_m defined above are stopping times.*

Proof. Clearly τ_1 is a stopping time (Example 5.27). If τ_m is a stopping time, then

$$\{\tau_{m+1} = n\} = \bigcup_{\ell=0}^{n-1} \left(\{\tau_m = \ell\} \cap \{X_n \in S\} \cap \bigcap_{j=\ell+1}^{n-1} \{X_j \notin S\} \right) \in \mathcal{F}_n,$$

where $S := (-\infty, a]$ if m is even, and $S := [b, \infty)$ if m is odd. □

We define the **number of upcrossings** by time n as

$$\beta_n(a,b) := \begin{cases} 0 & n < \tau_2, \\ \max\{m \geq 1 : \tau_{2m} < n\} & \text{otherwise.} \end{cases}$$

This is the number of times that the process crosses from below a to above b in the time interval $[0, n]$.

Theorem 5.36 (Doob's upcrossing inequality)**.** *If $(X_n)_{n\geq 0}$ is a submartingale, then*

$$\mathrm{E}[\beta_n(a,b)] \leq \frac{1}{b-a} \mathrm{E}[(X_n - a)^+].$$

Proof. We claim $(Y_n)_{n\geq 0}$ is a nonnegative submartingale, where $Y_n := (X_n - a)^+$. It is nonnegative by definition, and Proposition 5.34 shows that it is a submartingale because the function $x \mapsto (x-a)^+$ is convex. We have

$$\begin{aligned}
\{X_n \geq b\} &= \{X_n - a \geq b - a\} = \{Y_n \geq b - a\},\\
\{X_n \leq a\} &= \{(X_n - a)^+ \leq 0\} = \{Y_n \leq 0\}.
\end{aligned}$$

This gives an important relationship between upcrossings in (X_n) and (Y_n).

$$\beta_n^X(a,b) = \beta_n^Y(0, b-a).$$

Because the number of upcrossings of a general submartingale can be transformed into an analogous function of a nonnegative submartingale, proving the theorem reduces to showing

$$\mathrm{E}[\beta_n(0,b)] < \frac{1}{b} \mathrm{E}[X_n]$$

where (X_n) is a nonnegative submartingale. Note that the nonnegativity of $(X_n)_{n\geq 0}$ now implies $X_{\tau_{2k+1}} = 0$ for all $k \geq 0$.

We define

$$V_n := \begin{cases} 0 & n \leq \tau_1, \\ 1 & \tau_{2m-1} < n \leq \tau_{2m}, \\ 0 & \tau_{2m} < n \leq \tau_{2m+1}. \end{cases}$$

This process takes the value 1 only when the process is in the process of crossing from below a to above b. We claim $(V_n)_{n\geq 0}$ is predictable with respect to $(\mathcal{F}_n)_{n\geq 0}$. Since V_n takes only two values, it suffices to show that $\{V_n = 1\} \in \mathcal{F}_{n-1}$. Indeed,

$$\{V_n = 1\} = \bigcup_{m=1}^{\infty} \{\tau_{2m-1} < n \leq \tau_{2m}\} = \bigcup_{m=1}^{\infty} (\{\tau_{2m-1} \leq n-1\} \cap \{\tau_{2m} \leq n-1\}^c) \in \mathcal{F}_{n-1}.$$

We also claim

$$(V \cdot X)_n \geq b\beta_n(0, b)$$

for each n. Indeed, we have

$$\begin{aligned}(V \cdot X)_n &:= \begin{cases} (V \cdot X)_{n-1} & V_n = 0 \\ (V \cdot X)_{n-1} + \Delta X_n & V_n = 1 \end{cases} \\ &= \begin{cases} X_{\tau_2} + X_{\tau_4} + \cdots + X_{\tau_{2k}} & \tau_{2k} < n \leq \tau_{2k+1}, \\ X_{\tau_2} + X_{\tau_4} + \cdots + X_{\tau_{2k}} + X_n & \tau_{2k+1} < n \leq \tau_{2k+2}. \end{cases} \\ &\leq b\beta_n(0, b).\end{aligned}$$

The second equality follows because we are only adding the differences ΔX_n on the upswings from 0 to b, and cancellations simply leave the value at the top of the swing. The inequality holds because the number of terms in the sum is less than $\beta_n(0, b)$ and because $X_{2k} \geq b$ for each k.

We are now equipped to finish the proof.

$$\begin{aligned}
\mathrm{E}[b\beta_n(a, b)] &\leq \mathrm{E}[(V \cdot X)_n] && \text{see above} \\
&= \mathrm{E}\left[\sum_{i=1}^{n} V_i(X_i - X_{i-1})\right] \\
&= \sum_{i=1}^{n} \mathrm{E}[V_i(X_i - X_{i-1})] \\
&= \sum_{i=1}^{n} \mathrm{E}[V_i \, \mathrm{E}[(X_i - X_{i-1}) \mid \mathcal{F}_{i-1}]] && (V_n)_{n\geq 0} \text{ is } (\mathcal{F}_n)\text{-predictable, see above} \\
&\leq \sum_{i=1}^{n} \mathrm{E}[\mathrm{E}[X_i - X_{i-1} \mid \mathcal{F}_{i-1}]] && \mathrm{E}[X_i - X_{i-1} \mid \mathcal{F}_{i-1}] \geq 0 \text{ and } V_i \in \{0, 1\} \\
&= \mathrm{E}[X_n - X_0] \\
&\leq \mathrm{E}[X_n]. && X_0 \geq 0
\end{aligned}$$

□

5.7 Convergence theorems

Theorem 5.37. *Let (X_1) be a submartingale that is bounded in L^1 (i.e., $\sup_n \|X_n\|_1 < \infty$). Then there exists a random variable X_∞ such that $X_n \to X_\infty$ almost surely, and $\|X_\infty\|_1 \leq \sup_n \|X_n\|_1$.*

Proof. Fix $a < b$. Then

$$\mathrm{E}[(X_n - a)^+] \leq \mathrm{E}[|X_n - a|] \leq \mathrm{E}[|X_n|] + |a| \leq \sup_m \|X_m\|_1 + |a| < \infty,$$

for each n, so $\sup_n \mathrm{E}[(X_n - a)^+] < \infty$. Thus, using the fact that $\beta_n(a,b)$ is increasing in n, we have

$$\begin{aligned}
\mathrm{E}\Big[\lim_{n\to\infty} \beta_n(a,b)\Big] &= \lim_{n\to\infty} \mathrm{E}[\beta_n(a,b)] && \text{monotone convergence (Theorem 4.11)} \\
&\le \sup_n \frac{\mathrm{E}[(X_n - a)^+]}{b-a} && \text{upcrossing inequality (Theorem 5.36)} \\
&< \infty.
\end{aligned}$$

This implies

$$P\Big(\lim_{n\to\infty} \beta_n(a,b) = \infty\Big) = 0.$$

If the limit of X_n does not exist, it must oscillate between its lim inf and lim sup; however, we have just shown that this almost never happens.

$$\begin{aligned}
P\Big(\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\Big) &\le \sum_{\substack{a,b\in\mathbb{Q}\\ a<b}} P\Big(\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\Big) && \text{union bound} \\
&\le \sum_{\substack{a,b\in\mathbb{Q}\\ a<b}} P\Big(\lim_{n\to\infty} \beta_n(a,b) = \infty\Big) \\
&= 0.
\end{aligned}$$

Thus there exists some X_∞ for which $X_n \to X_\infty$ almost surely (e.g., $X_\infty := \liminf_{n\to\infty} X_n$).

The final claim follows from Fatou's lemma.

$$\begin{aligned}
\|X_\infty\|_1 &\le \liminf_{n\to\infty} \|X_n\|_1 && \text{Fatou's lemma (Theorem 4.13)} \\
&\le \sup_n \|X_n\|_1.
\end{aligned}$$

□

Corollary 5.38. *Let $(X_n)_{n\ge 0}$ be a uniformly integrable submartingale. Then there exists a random variable $X_\infty \in L^1$ such that $X_n \to X_\infty$ converges almost surely and in L^1. Moreover, $\mathrm{E}[X_\infty \mid \mathcal{F}_n] \ge X_n$ for all n.*

Proof. Theorem 5.37 gives the existence of X_∞ such that $X_n \to X_\infty$ almost surely. Theorem 4.38 shows that X_∞ is in L^1 and that $X_n \to X_\infty$ in L^1.

Fix $A \in \mathcal{F}_n$. Then $\mathbf{1}_A X_m \to \mathbf{1}_A X_\infty$ almost surely, and $(\mathbf{1}_A X_m)_{m\ge 0}$ is uniformly integrable. Then,

$$\begin{aligned}
\mathrm{E}[\mathbf{1}_A X_\infty] &= \lim_{m\to\infty} \mathrm{E}[\mathbf{1}_A X_m] && \text{Theorem 4.38} \\
&\ge \mathrm{E}[\mathbf{1}_A X_n]. && \mathrm{E}[\mathbf{1}_A X_m] \ge \mathrm{E}[\mathbf{1}_A X_n] \text{ for } m \ge n \text{ (submartingale)}
\end{aligned}$$

By the definition of conditional expectation, we have $\mathrm{E}[X_\infty \mid \mathcal{F}_n] \ge \mathrm{E}[X_n \mid \mathcal{F}_n] = X_n$. □

Corollary 5.39. *A submartingale $(X_n)_{n\ge 0}$ that is bounded from above converges almost surely.*

Proof.

$$\begin{aligned}
\mathrm{E}[|X_n|] &= \mathrm{E}[X_n^+] + \mathrm{E}[X_n^-] \\
&= 2\,\mathrm{E}[X_n^+] - \mathrm{E}[X_n^+] + \mathrm{E}[X_n^-] \\
&= \mathrm{E}[X_n^+] - \mathrm{E}[X_n] \\
&\le 2 \sup_m \mathrm{E}[X_m^+] - \mathrm{E}[X_0] && \text{submartingale} \\
&< \infty. && (X_n)_{n\ge 0} \text{ is bounded from aobve}
\end{aligned}$$

Thus $(X_n)_{n\ge 0}$ is bounded in L^1, so Theorem 5.37 implies the desired result. □

The following corollary consists of the analogues of the above results for supermartingales and martingales.

Corollary 5.40.

- *A L^1-bounded supermartingale converges almost surely.*
- *A uniformly integrable supermartingale $(X_n)_{n\geq 0}$ converges almost surely and in L^1 to a random variable $X_\infty \in L^1$. Moreover, $\mathrm{E}[X_\infty \mid \mathcal{F}_n] \leq X_n$ for all n.*
- *A supermartingale that is bounded from below converges almost surely.*
- *If $(X_n)_{n\geq 0}$ is a uniformly integrable martingale, there exists $X_\infty \in L^1$ such that $X_n \to X_\infty$ almost surely and in L^1. Moreover, $\mathrm{E}[X_\infty \mid \mathcal{F}_n] = X_n$ for all n.*

Corollary 5.41. *Let $(X_n)_{n\geq 0}$ be a submartingale that is bounded in L^p (i.e., $\sup_n \|X_n\|_p < \infty$) for some $p > 1$. Then there exists a random variable $X_\infty \in L^1$ such that $X_n \to X_\infty$ almost surely and in L^1.*

Proof. Corollary 4.40 implies that $(X_n)_{n\geq 0}$ is uniformly integrable, so we may apply Corollary 5.38. □

Theorem 5.42. *If $(X_n)_{n\geq 0}$ is an L^2-bounded martingale, then it converges in L^2.*

Theorem 5.43. *If $(X_n)_{n\geq 0}$ is an L^p-bounded martingale, then it converges in L^p.*

Example 5.44. Let $Y_1, Y_2, \ldots$ be random variables such that $P(Y_n = 1) = P(Y_n = -1) = 1/2$ for all n. Consider $(X_n)_{n\geq 0}$ where

$$X_n := \sum_{k=1}^{n} \frac{Y_k}{k}.$$

It is a martingale. Moreover,

$$\begin{aligned}
\mathrm{E}[X_n^2] &= \sum_{k=1}^{n} \frac{\mathrm{E}[Y_k^2]}{k^2} + 2 \sum_{1\leq j<k\leq n} \frac{\mathrm{E}[Y_j Y_k]}{jk} \\
&= \sum_{k=1}^{n} \frac{1}{k^2} && \text{independence of the } Y_k \\
&\leq \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,
\end{aligned}$$

so the martingale is bounded in L^2. By Corollary 5.41, we have the existence of $X_\infty \in L^1$ such that $X_n \to X_\infty$ almost surely and in L^1. Theorem 5.42 also shows that it converges in L^2 as well.

Example 5.45. Let $Y_1, Y_2, \ldots$ be random variables such that $P(Y_n = 1) = P(Y_n = -1) = 1/2$ for all n. The process $(e^{Y_1+\cdots+Y_n})_{n\geq 0}$ is a strict submartingale, since $\mathrm{E}[e^{Y_k}] = \cosh(1) > 1$. So, we can normalize by this factor to obtain a martingale $(X_n)_{n\geq 0}$ where

$$X_n := \frac{e^{Y_1}}{\mathrm{E}[e^{Y_1}]} \cdots \frac{e^{Y_n}}{\mathrm{E}[e^{Y_n}]} = e^{Y_1+\cdots+Y_n - n\log\cosh(1)}.$$

Note that this martingale is positive.

Since $\|X_n\|_1 = 1$ for all n, the martingale is bounded in L^1, so Theorem 5.37 implies the existence of X_∞ such that $X_n \to X_\infty$ almost surely. [To arrive at this conclusion, we could also note that the process is a supermartingale that is bounded from below, and then apply Corollary 5.40.]

We claim $X_\infty = 0$. Indeed, note that $Y_1 + \cdots + Y_n - n\log\cosh(1)$ is the symmetric random walk with downward drift, and that exponentiating it gives the process $(X_n)_{n\geq 0}$ which we showed converges almost surely. Due to the downward drift, the only way this can happen is if $Y_1 + \cdots + Y_n - n\log\cosh(1) \to -\infty$ and $X_n \to 0$. [This is an indirect proof that adding downward drift to the symmetric random walk makes it tend to negative infinity.] However, this shows $(X_n)_{n\geq 0}$ does not converge in L^1 because $\mathrm{E}[X_n] = 1$ while $\mathrm{E}[X_\infty] = 0$. Consequently, the martingale is not uniformly integrable either (Theorem 4.38).

Note that $\mathrm{E}[X_n] = 1$ for all n, while $X_n \to 0$ almost surely. Since almost all paths tend to zero, more and more paths are close to zero as n increases. However, at the same time, the height paths that are far from zero get higher and higher as n increases, in order to maintain the average $\mathrm{E}[X_n] = 1$.

This martingale is the discrete analogue of exponential Brownian motion.

Theorem 5.46. *If $X \in L^1(\Omega, \mathcal{F}, P)$, then the collection*

$$\{\mathrm{E}[X \mid \mathcal{G}] : \mathcal{G} \subset \mathcal{F} \text{ is a sub-}\sigma\text{-algebra}\}$$

is uniformly integrable.

Corollary 5.47. *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration on a probability space $(\Omega, \mathcal{F}, P)$. If $X \in L^1(\Omega, \mathcal{F}, P)$, then $X_n := \mathrm{E}[X \mid \mathcal{F}_n]$ defines a uniformly integrable martingale.*

Definition 5.48. Given a filtration $(\mathcal{F}_n)_{n\geq 0}$, the union $\bigcup_{n\geq 0} \mathcal{F}_n$ is an algebra but not necessarily a σ-algebra. [Consider $\Omega := \mathbb{N}$ and let $\mathcal{F}_n$ be the collection of subsets of $\{1, \ldots, n\}$ and their complements in $\mathbb{N}$. For any k, the set $A_k := \{2k\} \in \mathcal{F}_{2k}$ is in $\bigcup_{n\geq 0} \mathcal{F}_n$, but $\bigcup_{k\geq 1} A_k = \{2k \mid k \geq 1\}$ is not in any $\mathcal{F}_n$ because it and its complement are both infinite.]

Thus, we define

$$\bigvee_{n\geq 0} \mathcal{F}_n := \sigma\left(\bigcup_{n\geq 0} \mathcal{F}_n\right).$$

Theorem 5.49 (Lévy's upward theorem)**.** *Let $X \in L^1(\Omega, \mathcal{F}, P)$ and let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration. Let $X_n := \mathrm{E}[X \mid \mathcal{F}_n]$ and $X_\infty := \mathrm{E}[X \mid \mathcal{F}_\infty]$ where $\mathcal{F}_\infty := \bigvee_{n\geq 0} \mathcal{F}_n$. Then $X_n \to X_\infty$ almost surely and in L^1.*

Proof. Since $(X_n)_{n\geq 0}$ is a uniformly integrable martingale (Corollary 5.47), there exists $\widetilde{X}_\infty \in L^1(\Omega, \mathcal{F}_\infty, P)$ such that $X_n \to \widetilde{X}_\infty$ almost surely and in L^1.

Let $\mathcal{A} := \bigcup_{n\geq 0} \mathcal{F}_n$; it is an algebra. For all $A \in \mathcal{A}$,

$$\begin{aligned}
\mathrm{E}[\widetilde{X}_\infty \mathbf{1}_A] &= \lim_{n\to\infty} \mathrm{E}[X_n \mathbf{1}_A] = \mathrm{E}[X \mathbf{1}_A] && \text{Theorem 4.38} \\
&= \lim_{n\to\infty} \mathrm{E}[\mathrm{E}[X \mid \mathcal{F}_n] \mathbf{1}_A] \\
&= \mathrm{E}[X \mathbf{1}_A]. && \mathbf{1}_A \text{ is } \mathcal{F}_n\text{-measurable for all large } n
\end{aligned}$$

However, we want to show this equality for all $A \in \mathcal{F}_\infty$. Let $\mathcal{M} := \{A \in \mathcal{F} : \mathrm{E}[\widetilde{X}_\infty \mathbf{1}_A] = \mathrm{E}[X \mathbf{1}_A]\}$. It is a monotone class by the monotone convergence theorem (Theorem 4.11)). By the monotone class theorem (Theorem 3.15), $\mathcal{F}_\infty \subset \mathcal{M}$, so indeed $\mathrm{E}[\widetilde{X}_\infty \mathbf{1}_A] = \mathrm{E}[X \mathbf{1}_A]$ for all $A \in \mathcal{F}_\infty$, showing $\widetilde{X}_\infty = \mathrm{E}[X \mid \mathcal{F}_\infty] =: X_\infty$. □

Our study of martingales gives us a quick proof of the following theorem.

Theorem 5.50 (Kolmogorov's zero-one law)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables. We define the **tail σ-algebra** by*

$$\mathcal{G} := \bigcap_{n\geq 0} \mathcal{G}_n$$

where $\mathcal{G}_n := \sigma(X_{n+1}, X_{n+2}, \ldots)$. Then $P(A)$ is either 0 or 1 for all $A \in \mathcal{G}$. In particular, $\mathcal{G}$-measurable random variables are constant almost surely.

Proof. Let $\mathcal{F}_n^X := \sigma(X_1, \ldots, X_n)$ be the natural filtration, and let $\mathcal{F}_\infty^X := \bigvee_{n\geq 0} \mathcal{F}_n^X$. Note that

$$\mathcal{G} \subset \sigma(X_1, X_2, \ldots) = \bigcup_{n\geq 0} \mathcal{F}_n^X \subset \mathcal{F}_\infty^X.$$

Moreover, $\mathcal{G}_n := \sigma(X_{n+1}, X_{n+2}, \ldots)$ is independent of $\mathcal{F}_n^X := \sigma(X_1, \ldots, X_n)$ for each n.

For $A \in \mathcal{G}$,

$$\begin{aligned}
\mathbf{1}_A &= \mathrm{E}[\mathbf{1}_A \mid \mathcal{F}_\infty^X] && \mathcal{G} \subset \mathcal{F}_\infty^X \\
&= \lim_{n\to\infty} \mathrm{E}[\mathbf{1}_A \mid \mathcal{F}_n^X] && \text{Theorem 5.49} \\
&= \mathrm{E}[\mathbf{1}_A]. && \mathbf{1}_A \text{ is } \mathcal{G}_n\text{-measuarable; } \mathcal{G}_n \text{ is indep. of } \mathcal{F}_n^X
\end{aligned}$$

This implies that $P(A)$ is either zero or one. □

Theorem 5.51 (Reverse martingale convergence)**.** *Let $\mathcal{G}_0 \supset \mathcal{G}_1 \supset \cdots$ be a decreasing sequence of σ-algebras on a probability space $(\Omega, \mathcal{F}, P)$ and let $X \in L^1(\Omega, \mathcal{F}, P)$. Then*

$$\mathrm{E}[X \mid \mathcal{G}_n] \to \mathrm{E}[X \mid \mathcal{G}]$$

almost surely and in L^1, where $\mathcal{G} := \bigcap_{n\geq 0} \mathcal{G}_n$.

Proof. Let $X_n := \mathrm{E}[X \mid \mathcal{G}_n]$ for each $n \geq 0$. Then $X_n, X_{n-1}, \ldots, X_1, X_0$ is a finite martingale for each $n \geq 1$. Letting $\beta_n(a,b)$ be the number of upcrossings with respect to $[a,b]$ for each finite martingale, we have

$$\begin{aligned}
\mathrm{E}[\lim_{n\to\infty} \beta_n(a,b)] &= \lim_{n\to\infty} \mathrm{E}[\beta_n(a,b)] && \text{monotone convergence (Theorem 3.15)} \\
&\leq \lim_{n\to\infty} \frac{\mathrm{E}[(X_0 - a)^+]}{b-a} && \text{Doob's upcrossing inequality (Theorem 5.36)}
\end{aligned}$$

for any $a < b$. By the same argument in the proof of Theorem 5.37, there exists a $\mathcal{G}$-measurable random variable X_∞ such that $X_n \to X_\infty$ almost surely. [To see that X_∞ is $\mathcal{G}$-measurable, note that $X_\infty = \lim_{n\geq m} X_n$ shows that X_∞ is $\mathcal{G}_n$-measurable, but this is true for any m.] By Theorem 5.46, $(X_n)_{n\geq 0}$ is uniformly integrable, so $X_n \to X_\infty$ in L^1 as well.

For each $A \in \mathcal{G}$, we have

$$\begin{aligned}
\mathrm{E}[X_\infty \mathbf{1}_A] &= \lim_{n\to\infty} \mathrm{E}[X_n \mathbf{1}_A] && L^1\text{-convergence} \\
&= \lim_{n\to\infty} \mathrm{E}[X \mathbf{1}_A] && X_n := \mathrm{E}[X \mid \mathcal{G}_n],
\end{aligned}$$

thus $X_\infty = \mathrm{E}[X \mid \mathcal{G}]$, completing the proof. □

5.8 Strong law of large numbers

Lemma 5.52. *Let $X_1, X_2, \ldots$ be i.i.d. random variables in L^1, and let $S_n := \sum_{i=1}^n X_i$ for each $n \geq 1$. Then*

$$\mathrm{E}[X_1 \mid S_n, S_{n+1}, S_{n+2}, \ldots] = \mathrm{E}[X_1 \mid S_n] = \frac{S_n}{n}.$$

Theorem 5.53 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. in $L^1(\Omega, \mathcal{F}, P)$. Then*

$$\frac{1}{n}\sum_{i=1}^n X_i \to \mathrm{E}[X_1]$$

almost surely and in L^1.

Proof. Let $S_n := \sum_{i=1}^n S_n$, let $\mathcal{G}_n := \sigma(S_n, S_{n+1}, \ldots)$ for each $n \geq 1$, and let $\mathcal{G} := \bigcap_{n\geq 1} \mathcal{G}_n$. Then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n S_n &= \frac{S_n}{n} \\
&= \mathrm{E}[X_1 \mid \mathcal{G}_n] && \text{Lemma 5.52} \\
&\to \mathrm{E}[X_1 \mid \mathcal{G}] && \text{Theorem 5.51}
\end{aligned}$$

almost surely and in L^1.

Now note that

$$\begin{aligned}
\mathrm{E}[X_1 \mid \mathcal{G}] &= \lim_{n\to\infty} \frac{S_n}{n} \\
&= \lim_{n\to\infty} \frac{X_1 + \cdots + X_{m-1}}{n} + \frac{X_m + \cdots + X_n}{n} \\
&= \lim_{n\to\infty} \frac{X_m + \cdots + X_n}{n} && \text{first term vanishes as } n \to \infty
\end{aligned}$$

is $\sigma(X_m, X_{m+1}, \ldots)$ for any $m \geq 1$, so it is $\bigcap_{m\geq 1} \sigma(X_m, X_{m+1}, \ldots)$-measurable. By Kolmogorov's zero-one law (Theorem 5.50), $\mathrm{E}[X_1 \mid \mathcal{G}]$ is constant [almost surely], so it must be $\mathrm{E}[X_1]$. □

5.9 Maximal inequalities

Theorem 5.54 (Doob's maximal inequality for probabilities)**.** *If $(X_n)_{n\geq 0}$ is a submartingale and $\lambda > 0$, then*

$$\lambda \cdot P\left(\max_{k\leq n} X_k \geq \lambda\right) \leq \mathrm{E}[X_n \mathbf{1}_{\{\max_{k\leq n} X_k \geq \lambda\}}] \leq \mathrm{E}[X_n^+].$$

Proof. The second inequality is clear, so we only show the first inequality. Let $\tau_k := n \wedge \inf\{k \in \mathbb{N} : X_k \geq \lambda\}$. Then,

$$\begin{aligned}
\mathrm{E}[X_n] &\geq \mathrm{E}[X_{\tau_\lambda}] && (X_n) \text{ is a submartingale} \\
&= \mathrm{E}[X_{\tau_\lambda} \mathbf{1}_{\{\max_{k\leq n} X_k \geq \lambda\}}] + \mathrm{E}[X_{\tau_\lambda} \mathbf{1}_{\{\max_{k\leq n} X_k < \lambda\}}] \\
&\geq \lambda \cdot P\left(\max_{k\leq n} X_k \geq \lambda\right) + \mathrm{E}[X_{\tau_\lambda} \mathbf{1}_{\{\max_{k\leq n} X_k < \lambda\}}] && \text{def. of } \tau_\lambda \\
&= \lambda \cdot P\left(\max_{k\leq n} X_k \geq \lambda\right) + \mathrm{E}[X_n \mathbf{1}_{\{\max_{k\leq n} X_k < \lambda\}}]. && \tau_\lambda = n \text{ in this event}
\end{aligned}$$

Subtracting the last term from both sides gives the first inequality. □

Corollary 5.55. *If $(X_n)_{n\geq 0}$ is a martingale, $\lambda > 0$, and $p \geq 1$, then*

$$P\left(\max_{k\leq n} |X_k| \geq \lambda\right) \leq \frac{\mathrm{E}[|X_n|^p]}{\lambda^p}.$$

Proof. By Proposition 5.34, $(|X_n|^p)_{n\geq 0}$ is a submartingale, so we may apply Theorem 5.54. □

Lemma 5.56. *For any $a, b > 0$, the following hold.*

$$a \log b \leq a \log a + \frac{b}{e},$$
$$a \log^+ b \leq a \log^+ a + \frac{b}{e}$$

where $\log^+(x) := (\log x)^+$ denotes the positive part of $\log x$.

Theorem 5.57 (Doob's L^p maximal inequality)**.** *Let $(X_n)_{n\geq 0}$ be a nonnegative submartingale. We clearly have the lower bound*

$$\left\| \max_{k\leq n} X_k \right\|_p \geq \|X_n\|_p$$

for any $p \geq 1$. However, we also have the following upper bounds.

1) If $p > 1$, then

$$\left\| \max_{k\leq n} X_k \right\|_p \leq \frac{p}{p-1} \|X_n\|_p.$$

2) For $p = 1$, we have

$$\left\| \max_{k \le n} X_k \right\|_1 \le \frac{e}{e-1}(1 + \|X_n \log^+ X_n\|_1),$$

where $\log^+ x := (\log x)^+$ denotes the positive part of $\log x$.

Proof. Let $Y := \max_{k \le n} X_k$ and let $Y_m := Y \wedge m$. Theorem 5.54 implies

$$\lambda P(Y \ge \lambda) \le \mathrm{E}[X_n \mathbf{1}_{\{Y \ge \lambda\}}].$$

Thus,

$$\lambda P(Y_m \ge \lambda) \le \mathrm{E}[X_n \mathbf{1}_{\{Y_m \ge \lambda\}}]$$

because if $m < \lambda$ then the probability on the left-hand side is zero and the inequality holds, and otherwise if $m \ge \lambda$ then $Y_m \ge \lambda \iff Y \ge \lambda$.

1) Let $p > 1$, and assume $\mathrm{E}[X_n^p] < \infty$ (otherwise, the inequality holds immediately). Note the identity

$$x^p = p \int_0^x \lambda^{p-1}\, d\lambda = p \int_0^\infty \mathbf{1}_{\{x \ge \lambda\}} \lambda^{p-1}\, d\lambda\,.$$

Applying this to $x := Y_m$ and taking the expectation gives

$$\begin{aligned}
\mathrm{E}[Y_m^p] &= p \int_0^\infty P(Y_m \ge \lambda) \lambda^{p-1}\, d\lambda && \text{Fubini (Theorem 4.16)} \\
&\le p \int_0^\infty \mathrm{E}[X_n \mathbf{1}_{\{Y_m \ge \lambda\}}] \lambda^{p-2}\, d\lambda && \text{see beginning of proof} \\
&= p\,\mathrm{E}\left[X_n \int_0^{Y_m} \lambda^{p-2}\, d\lambda\right] && \text{Fubini (Theorem 4.16)} \\
&= \frac{p}{p-1}\,\mathrm{E}[X_n Y_m^{p-1}] \\
&\le \frac{p}{p-1}\|X_n\|_p \|Y_m^{p-1}\|_q && \text{where } q \text{ is s.t. } \frac{1}{p} + \frac{1}{q} = 1\text{; Hölder's inequality (Theorem 4.19)} \\
&= \frac{p}{p-1}\|X_n\|_p\,\mathrm{E}[Y_m^p]^{1/q}. && (p-1)q = p
\end{aligned}$$

Dividing both sides by $\mathrm{E}[Y_m^p]^{1/q}$ gives $\|Y_m\|_p \le \frac{p}{p-1}\|X_n\|_p$. Since this holds for each m, taking $m \to \infty$ and using the monotone convergence theorem (Theorem 4.11) gives $\|Y\|_p \le \frac{p}{p-1}\|X_n\|_p$.

2) Assume $\mathrm{E}[X_n \log^+ X_n] < \infty$ (otherwise, the inequality holds immediately).

$$\begin{aligned}
\mathrm{E}[Y_m] - 1 &= \mathrm{E}\left[\int_0^{Y_m} d\lambda\right] - 1 \\
&= \int_0^\infty P(Y_m \ge \lambda)\, d\lambda - 1 && \text{Fubini (Theorem 4.16)} \\
&\le \int_1^\infty P(Y_m \ge \lambda)\, d\lambda && \int_0^1 P(Y_m \ge \lambda)\, d\lambda \le \int_0^1 d\lambda = 1 \\
&\le \int_1^\infty \frac{1}{\lambda}\,\mathrm{E}[X_n \mathbf{1}_{\{Y_m \ge \lambda\}}]\, d\lambda && \text{see beginning of proof} \\
&= \mathrm{E}\left[X_n \int_1^{Y_m} \frac{1}{\lambda}\, d\lambda\right] && \text{Fubini (Theorem 4.16)} \\
&= \mathrm{E}[X_n \log^+ Y_m] \\
&\le \mathrm{E}[X_n \log^+ X_n] + \frac{\mathrm{E}[Y_m]}{e}. && \text{Lemma 5.56}
\end{aligned}$$

Rearranging gives $\mathrm{E}[Y_m] \leq \frac{e}{e-1}(1+\mathrm{E}[X_n \log^+ X_n])$. Letting m tend to infinity and applying the monotone convergence theorem (Theorem 4.11) gives

$$\mathrm{E}[Y] \leq \frac{e}{e-1}(1 + \mathrm{E}[X_n \log^+ X_n]).$$

□

6 Markov chains

6.1 Kernels

Definition 6.1. Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{E})$ be measurable spaces. A mapping $K : \Omega, \mathcal{E} \to \mathbb{R}_+ \cup \{\infty\}$ is a **stochastic kernel** from $(\Omega, \mathcal{F})$ to $(E, \mathcal{E})$ if

a) $\omega \mapsto K(\omega, A)$ is a $\mathcal{F}$-measurable random variable for each $A \in \mathcal{E}$, and

b) $A \mapsto K(\omega, A)$ is a measure on $(E, \mathcal{E})$ for each $\omega \in \Omega$.

Example 6.2. If $\Omega := \{1, \dots, M\}$ and $E := \{1, \dots, N\}$ each with the discrete σ-algebra, then to define a stochastic kernel it suffices to determine $K(m, \{n\})$ for each $m \in \Omega$, $n \in E$ (due to the second condition in the definition of stochastic kernel). Thus the stochastic kernel can be represented as a matrix.

Example 6.3. Let $k : \Omega \times E \to \mathbb{R}_+ \cup \{\infty\}$ be a $\mathcal{F} \otimes \mathcal{E}$-measurable function, and let ν be a measure on $(E, \mathcal{E})$. Then

$$K(\omega, A) := \int_A k(\omega, e)\, \nu(de)$$

is a stochastic kernel.

Definition 6.4. A stochastic kernel is **finite** if $A \mapsto k(\omega, A)$ is a finite measure (i.e., $k(\omega, E) < \infty$) for each $\omega \in \Omega$. A stochastic kernel is a **transition probability kernel** if $A \mapsto k(\omega, A)$ is a probability measure (i.e., $k(\omega, E) = 1$) for each $\omega \in \Omega$.

Definition 6.5. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. A **regular conditional probability** of P with respect to $\mathcal{G}$ is a transition probability kernel $Q : (\Omega, \mathcal{G}) \to (\Omega, \mathcal{F})$ such that

$$Q(\cdot, A) = P(A \mid \mathcal{G})$$

P-almost surely for all $A \in \mathcal{F}$.

In general, the conditional probabilities $P(A \mid \mathcal{G})$ do not necessarily satisfy the conditions of being a kernel.

Lemma 6.6. *If we assume P has a regular conditional probability Q with respect to $\mathcal{G}$, then for every $X \in L^1(\Omega, \mathcal{F}, P)$,*

$$QX := \int X(\omega')Q(\cdot, d\omega')$$

equals $\mathrm{E}[X \mid \mathcal{G}]$ P-almost surely.

Definition 6.7. Let $X : (\omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. A **regular conditional distribution** of X given $\mathcal{G}$ is any transition probability kernel $Q : (\Omega, \mathcal{G} \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$Q(\cdot, A) = P(X \in A \mid \mathcal{G})$$

P-almost surely for all $A \in \mathcal{B}(\mathbb{R})$.

Theorem 6.8. *Let $X : (\omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then there exists a regular conditional distribution of X given $\mathcal{G}$.*

Proof. Let $q \in \mathbb{Q}$ and define

$$C_q := P(X \le q \mid \mathcal{G}),$$
$$\Omega_{qr} := \{C_q \le C_r\}.$$

[In the definition of Ω_{qr}, pick two versions of C_q and C_r.] Since $\{X \le q\} \subset \{X \le r\}$ if $q \le r$, we have $\Omega_{qr} = \Omega$ P-almost surely. Let

$$\Omega_0 := \bigcap_{\substack{q,r\in\mathbb{Q}\\ q<r}} \Omega_{qr}.$$

Then $\Omega_0 = \Omega$ P-almost surely as well.

For all $\omega \in \Omega_0$, the function $q \mapsto C_q(\omega)$ is a nondecreasing function $\mathbb{Q} \to [0,1]$. If we define for $t \in \mathbb{R}$ and $\omega \in \Omega_0$ the function

$$\overline{C}_t(\omega) := \lim_{\substack{q\in\mathbb{Q}\\ q\searrow t}} C_q(\omega),$$

then $t \mapsto \overline{C}_t(\omega)$ is a cdf for some distribution $\overline{Q}(\omega,\cdot)$.

Then,

$$Q(\omega, A) := \mathbf{1}_{\Omega_0}(\omega)\overline{Q}(\omega, A) + \mathbf{1}_{\Omega_0^c}(\omega)\delta_0(A)$$

is a regular conditional distribution of X given $\mathcal{G}$. Note that the choice of the Dirac measure is arbitrary; it can be replaced by any probability measure. □

6.2 Definition and basic properties

Let I be a countable state space. For simplicity we will implicitly assume I is either $\mathbb{N}$ or of the form $\{0, 1, \ldots, n\}$ for some n.

Definition 6.9. A vector $(\lambda_i)_{i\in I}$ is a **distribution** if $\lambda_i \ge 0$ for all $i \in I$ and $\sum_{i\in I} \lambda_i = 1$. A matrix $(p_{i,j})_{i,j\in I}$ is **stochastic** if each of its rows is a distribution.

Definition 6.10. A stochastic process $(X_n)_{n\ge0}$ on a probability space $(\Omega, \mathcal{F}, P)$, with $X_n : \Omega \to I$ for each n, is a **Markov chain** with initial distribution λ and transition matrix p if

a) $P(X_0 = i) = \lambda_i$, and

b) $P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = p_{i_n,i_{n+1}}$ for any $n \ge 0$ and $i_0, \ldots, i_n \in I$.

We say $(X_n)_{n\ge0}$ is $\mathrm{Markov}(\lambda, p)$.

Theorem 6.11. *A discrete-time stochastic process $(X_n)_{n\ge0}$ is* $\mathrm{Markov}(\lambda, p)$ *if and only if*

$$P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}$$

for all $n \ge 0$ and $i_0, \ldots, i_n \in I$.

Proof. If $(X_n)_{n\ge0}$ is $\mathrm{Markov}(\lambda, p)$, then

$$\begin{aligned}
&P(X_0 = i_0, \ldots, X_n = i_n)\\
&= P(X_0 = i_0) \cdot P(X_1 = i_1 \mid X_0 = i_0) \cdots P(X_n = i_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1})\\
&= P(X_0 = i_0) \cdot P(X_1 = i_1 \mid X_0 = i_0) \cdots P(X_n = i_n \mid X_{n-1} = i_{n-1})\\
&= \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}.
\end{aligned}$$

The other direction is clear. □

Corollary 6.12. *For every initial distribution λ and stochastic matrix p, there exists a* $\mathrm{Markov}(\lambda, p)$ *process.*

Proof. Apply the Kolmogorov Extension Theorem (Theorem 3.46). □

Theorem 6.13 (Markov property). *Let $(X_n)_{n\geq 0}$ be a Markov chain. If $P(X_m = i) > 0$, then, conditional on the event $\{X_m = i\}$, the process $(X_{m+n})_{n\geq 0}$ is* Markov(δ_i, p) *and is independent of $(X_0, \dots, X_m)$.*

Let the initial distribution λ be interpreted as a row vector. Then λp is a row vector with entries $(\lambda p)_j := \sum_i \lambda_i p_{i,j}$. We can also multiply p with itself.

$$\begin{aligned} p^0_{ij} &:= \delta_{i,j} && \text{identity matrix} \\ p^1_{i,j} &:= p_{i,j} \\ p^2_{i,j} &:= \sum_k p_{i,k} p_{k,j} \\ p^3_{i,j} &:= \sum_k p^2_{i,k} p_{k,j} \\ &\vdots \end{aligned}$$

This proves the following result.

Lemma 6.14. *If $(X_n)_{n\geq 0}$ is* Markov(λ, p)*, then*

$$\begin{aligned} P(X_n = j) &= (\lambda p^n)_j, \\ P(X_{n+m} = j \mid X_m = i) &= p^n_{i,j}, \end{aligned}$$

for all $m \geq 0$.

6.3 Class structure

Definition 6.15. We let P_i denote the probability measure $P(\cdot \mid X_0 = i)$, and let E_i denote the expectation with respect to this measure. We say i **leads to** j and write $i \to j$ if $P_i(X_n = j \text{ for some } n \geq 0) > 0$. We say i **communicates with** j and write $i \leftrightarrow j$ if $i \to j$ and $j \to i$. By definition we have $i \leftrightarrow i$.

Theorem 6.16. *For $i \neq j$, the following are equivalent.*

(i) $i \to j$.

(ii) $p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j} > 0$ *for some* $n \geq 1$ *and* $i_1, \dots, i_{n-1} \in I$.

(iii) $p^n_{i,j} > 0$ *for some* $n \geq 1$.

Definition 6.17. Since communication is an equivalence relation on I, it induces a partition of I into **communicating classes**. A communicating class $C \subset I$ is **closed** if $i \in C$ and $i \to j$ together imply $j \in C$. A state $i \in I$ is **absorbing** if $\{i\}$ is a closed class. We call a Markov chain or transition matrix **irreducible** if all of I is a communicating class.

6.4 Hitting times and absorption probabilities

Definition 6.18. Let $A \subset I$. We define the **hitting time** of A by

$$H^A := \inf\{n \geq 0 : X_n \in A\}.$$

We also define

$$\begin{aligned} h^A_i &:= P_i(H^A < \infty), \\ k^A_i &:= \mathrm{E}_i[H^A]. \end{aligned}$$

Theorem 6.19.

a) The vector of hitting probabilities $(h_i^A)_{i\in I}$ is the minimal nonnegative solution to the system of equations

$$h_i^A = \begin{cases} 1 & i \in A, \\ \sum_{j\in I} p_{i,j} h_j^A & i \notin A. \end{cases}$$

b) The vector of expected hitting times $(k_i^A)_{i\in I}$ is the minimal nonnegative solution to the system of equations

$$k_i^A = \begin{cases} 0 & i \in A, \\ 1 + \sum_{j\notin A} p_{i,j} k_j^A & i \notin A. \end{cases}$$

Proof. We first prove $(h_i^A)_{i\in I}$ is indeed a solution to the system. By definition, $h_i^A = 1$ for $i \in A$. If $i \notin A$, then conditioning on X_1 and applying the Markov property (Theorem 6.13) gives the appropriate expression.

$$h_i^A = \sum_{j\in I} P_i(X_1 = j) P_i(H^A < \infty \mid X_1 = j) = \sum_{j\in I} p_{i,j} h_j^A.$$

We now show that $(h_i^A)_{i\in I}$ is smaller than any other nonnegative solution $(x_i)_{i\in I}$ to the system. For $i \in A$ we have $h_i^A = x_i = 1$. For $i \notin A$, we have

$$\begin{aligned} x_i &= \sum_{j_1\in I} p_{i,j_1} x_{j_1} \\ &= \sum_{j_1\in A} p_{i,j_1} + \sum_{j_1\notin A} p_{i,j_1} x_{j_1} \\ &= P_i(X_1 \in A) + \sum_{j_1\notin A} \left[p_{i,j_1} \left(\sum_{j_2\in A} p_{j_1,j_2} + \sum_{j_2\notin A} p_{j_1,j_2} x_{j_2} \right) \right] \\ &= P_i(X_1 \in A) + P_i(X_1 \notin A, X_2 \in A) + \sum_{j_1,j_2\notin A} p_{i,j_1} p_{j_1,j_2} x_{j_2} \\ &\vdots \\ &= P_i(X_1 \in A) + \cdots + P_i(X_1 \notin A, \ldots, X_{n-1} \notin A, X_n \in A) + \sum_{j_1,\ldots,j_n\notin A} p_{i,j_1} \cdots p_{j_{n-1},j_n} x_{j_n} \\ &\vdots \end{aligned}$$

Because $x_j \geq 0$ for all $j \in I$, the above implies

$$x_i \geq P_i(X_1 \in A) + \cdots + P_i(X_1 \notin A, \ldots, X_{n-1} \notin A, X_n \in A) = P_i(H^A \leq n)$$

for all $n \geq 1$. Taking the limit as $n \to \infty$ gives $x_i \geq h_i^A$.

The proof for $(k_i^A)_{i\in I}$ is analogous. To see that $(k_i^A)_{i\in I}$ is a solution to the system, note that by definition $k_i^A = 0$ for $i \in A$, and for $i \notin A$ we condition on X_1 and apply the Markov property (Theorem 6.13) as before.

$$k_i^A = \sum_{j\in I} P_i(X_1 = j) \operatorname{E}_i[H^A \mid X_1 = j] = \sum_{j\in I} p_{i,j}(1 + k_j^A) = 1 + \sum_{j\in I} p_{i,j} k_i^A = 1 + \sum_{j\notin A} p_{i,j} k_i^A.$$

We now show that $(k_i^A)_{i\in I}$ is smaller than any nonnegative solution $(x_i)_{i\in I}$ to the system. If $i \in A$ we have $k_i^A = x_i = 0$. For $i \notin A$,

$$\begin{aligned}
x_i &= 1 + \sum_{j_1 \notin A} p_{i,j_1} x_{j_1} \\
&= P_i(H^A \geq 1) + \sum_{j_1 \notin A} \left[p_{i,j_1} \left(1 + \sum_{j_2 \notin A} p_{j_1,j_2} x_{j_2} \right) \right] \\
&= P_i(H^A \geq 1) + P_i(H^A \geq 2) + \sum_{j_1, j_2 \notin A} p_{i,j_1} p_{j_1,j_2} x_{j_2} \\
&\vdots \\
&= P_i(H^A \geq 1) + \cdots + P(H^A \geq n) + \sum_{j_1, \ldots, j_n \notin A} p_{i,j_1} \cdots p_{j_{n-1},j_n} x_{j_n} \\
&\vdots
\end{aligned}$$

Because $x_j \geq 0$ for all $j \in I$, the above implies

$$x_i \geq \sum_{k=1}^{n} P_i(H^A \geq k)$$

for all $n \geq 1$. Taking the limit as $n \to \infty$ gives

$$x_i \geq \sum_{k=1}^{\infty} P_i(H^A \geq k) = \mathrm{E}_i[H^A] = k_i^A.$$

□

6.5 Strong Markov property

Theorem 6.20 (Strong Markov property)**.** *Let $(X_n)_{n \geq 0}$ be* Markov(λ, p) *and let τ be a stopping time with respect to the natural filtration $(\mathcal{F}_n^X)_{n \geq 0}$. If i is such that $P(\tau < \infty, X_\tau = i) > 0$, then, conditional on the event $\{\tau < \infty\} \cap \{X_\tau = i\}$, the process $(X_{\tau+n})_{n \geq 0}$ is* Markov(δ_i, p) *and is independent of $\mathcal{F}_\tau^X$.*

6.6 Recurrence and transience

Definition 6.21. A state $i \in I$ is **recurrent** if

$$P_i(X_n = i \text{ for infinitely many } n) = 1,$$

and **transient** if

$$P_i(X_n = i \text{ for infinitely many } n) = 0.$$

Theorem 6.25 shows that all states are either recurrent or transient.

Definition 6.22. We define

$$\begin{aligned}
T_i^0 &:= 0 \\
T_i \equiv T_i^1 &:= \inf\{n \geq 1 : X_n = i\} \\
T_i^2 &:= \inf\{n \geq T_i^1 + 1 : X_n = i\} \\
&\vdots \\
T_i^k &:= \inf\{n \geq T_i^{k-1} + 1 : X_n = i\} \\
&\vdots
\end{aligned}$$

In short, T_i^k is the time of the kth visit to state i strictly after time 0. We also define

$$\begin{aligned}
S_i^k &:= \begin{cases} T_i^k - T_i^{k-1} & T_i^{k-1} < \infty, \\ 0 & \text{otherwise.} \end{cases} && \text{time between visits to state } i \\
V_i &:= \sum_{n \geq 0} \mathbf{1}_{\{X_n = i\}} && \text{number of visits to state } i \\
f_i &:= P_i(T_i < \infty) && \text{probability of visiting state } i \text{ if starting there.}
\end{aligned}$$

Lemma 6.23. *For $k \geq 2$, conditional on $T_i^{k-1} < \infty$, the random variable S_i^k is independent of $\mathcal{F}^X_{T^{k-1}}$, and*

$$P(S_i^k = n \mid T_i^{k-1} < \infty) = P_i(T_i = n).$$

Proof. In the event $\{T_i^{k-1} < \infty\}$, we also have $X_{T_i^{k-1}} = i$. Therefore, by the strong Markov property (Theorem 6.20), conditional on $\{T_i^{k-1} < \infty\}$, the process $(X_{T_i^{k-1}+n})_{n \geq 0}$ is Markov(δ_i, p) and is independent of $\mathcal{F}^X_{T_i^{k-1}}$. Noting that $S_i^k = \inf\{n \geq 1 : X_{T_i^{k-1}+n} = i\}$ finishes the proof. □

Lemma 6.24.

a) $\mathrm{E}_i[V_i] = \sum_{n \geq 0} p_{i,i}^n$.

b) $P_i(V_i > k) = f_i^k$ *for each* $k \geq 0$.

Proof. To prove a), note that

$$\mathrm{E}_i[V_i] = \mathrm{E}_i\left[\sum_{n \geq 0} \mathbf{1}_{\{X_n = i\}}\right] = \sum_{n \geq 0} P_i(X_n = i) = \sum_{n \geq 0} p_{i,i}^n.$$

To prove b), we use induction. Note that $X_0 = i$ implies $\{V_i > k\} = \{T_i^k < \infty\}$ for all $k \geq 0$. [Note that V_i counts the visit at time 0, while T_i^k does not.]

Clearly the statement we are asked to show holds for $k = 0, 1$.

$$\begin{aligned}
P_i(V_i > 0) &= 1 = f_i^0, \\
P_i(V_i > 1) &= P_i(T_i < \infty) = f_i^1.
\end{aligned}$$

Suppose the statement holds for some fixed $k \geq 2$. We show it also holds for $k + 1$.

$$\begin{aligned}
P_i(V_i > k+1) &= P_i(T_i^{k+1} < \infty) \\
&= P_i(T_i^k < \infty, S_i^{k+1} < \infty) \\
&= P_i(S_i^{k+1} < \infty \mid T_i^k < \infty) \cdot P(T_i^k < \infty) \\
&= f_i f_i^k && \text{Lemma 6.23 and inductive hypothesis} \\
&= f_i^{k+1}.
\end{aligned}$$

□

Theorem 6.25.

a) If $P_i(T_i < \infty) = 1$, then i is recurrent and $\sum_{n \geq 0} p_{i,i}^n = \infty$.

b) If $P_i(T_i < \infty) < 1$, then i is transient and $\sum_{n \geq 0} p_{i,i}^n < \infty$.

In particular, every state is either recurrent or transient.

Proof. If $P_i(T_i < \infty) = 1$, then Lemma 6.24 implies

$$P_i(V_i = \infty) = \lim_{k\to\infty} P_i(V_i > k) = \lim_{k\to\infty} f_i^k = 1,$$

which in turn implies that i is recurrent and

$$\sum_{n\geq 0} p_{i,i}^n = \mathrm{E}_i[V_i] = \infty.$$

If $P_i(T_i < \infty) < 1$, then

$$\sum_{n\geq 0} p_{i,i}^n = \mathrm{E}_i[V_i] = \sum_{k\geq 0} P_i(V_i > k) = \sum_{k\geq 0} f_i^k = \frac{1}{1-f_i} < \infty.$$

This implies $P(V_i < \infty) = 1$, i.e., i is transient. □

Theorem 6.26. *The states of a communicating class are either all recurrent or all transient.*

Proof. Let i and j be the states of a communicating class, and suppose i is transient. There exists $m, n \geq 0$ such that $p_{i,j}^m > 0$ and $p_{j,i}^n > 0$. For all $k \geq 0$ we have $p_{i,i}^{m+k+n} \geq p_{i,j}^m p_{j,j}^k p_{j,i}^n$, so the transience of i implies

$$\sum_{k\geq 0} p_{j,j}^k \leq \frac{1}{p_{i,j}^m p_{j,i}^n} \sum_{k\geq 0} p_{i,i}^{m+k+n} < \infty,$$

i.e., j is transient. Thus, the states of a communicating class that contains a transient state will all be transient. Otherwise, all states of the class are recurrent. □

Consequently, we can call a communicating class recurrent or transient based on what kind of states it contains.

Theorem 6.27. *A recurrent communicating class is closed.*

Proof. We prove the contrapositive. Suppose C is a communicating class that is not closed. Then there exists $i \in C$ and $j \notin C$ such that $i \to j$, i.e., $P_i(X_m = j) > 0$ for some $m \geq 0$. Then,

$$\begin{aligned} P_i(X_m = j \text{ and } X_n = i \text{ for infinitely many } n) &= 0 && j \notin C \\ \implies P_i(X_n = i \text{ for infinitely many } n) &< 1 && P_i(X_m = j) > 0, \end{aligned}$$

implying i is not recurrent. □

Theorem 6.28. *Every finite closed communicating class is recurrent.*

Proof. Let C be a finite closed communicating class, and let $(X_n)_{n\geq 0}$ be a Markov chain starting in C. Then there exists a state $i \in C$ such that

$$0 < P(X_n = i \text{ for infinitely many } n) = P(X_n = i \text{ for some } n) P_i(X_n = i \text{ for infinitely many } n),$$

where the inequality is by the definition of a closed communicating class, and the equality is due to the Markov property. This implies $P(X_n = i \text{ for some } n)$, so i is recurrent. □

Theorem 6.29. *Assume p is an irreducible and recurrent transition matrix. Then for every initial distribution λ, we have $P(T_j < \infty) = 1$ for all $j \in I$.*

Proof. Since $P(T_j < \infty) = \sum_{i\in I} \lambda_i P_i(T_j < \infty)$, it suffices to show that $P_i(T_j < \infty) = 1$ for each $i \in I$. Choose m such that $p_{j,i}^m > 0$.

$$\begin{aligned} 1 &= P_j(X_n = j \text{ for infinitely many } n) \\ &= P_j(X_n = j \text{ for some } n \geq m+1) \\ &= \sum_{k\in I} P_j(X_n = j \text{ for some } n \geq m+1 \mid X_m = k) P_j(X_m = k) \\ &= \sum_{k\in I} P_k(T_j < \infty) p_{j,k}^m. \end{aligned}$$

Since $\sum_{k\in I} p_{j,k}^m = 1$, we must have $P_k(T_j < \infty) = 1$ for every k such that $p_{j,k}^m > 0$; in particular, $P_i(T_j < \infty) = 1$. □

6.7 Recurrence and transience of random walks

Example 6.30 (One-dimensional random walk)**.** To be consistent with Section 5.5, we will use p to denote the parameter of the random walk. It should not be confused with the transition matrix p.

Let $I := \mathbb{Z}$, $0 < p < 1$, $p_{i,i+1} := q$, and $p_{i,i-1} = 1 - p$ for each i. The transition matrix is clearly irreducible.

We now determine whether the states are transient or recurrent. Note the equation

$$P_0(T_0 < \infty) = \frac{1}{2} P_1(T_0 < \infty) + \frac{1}{2} P_{-1}(T_0 < \infty).$$

Case 1. $p = 1/2$. By our results for hitting times in Section 5.5, we have $P_1(T_0 < \infty) = P_{-1}(T_0 < \infty) = 1$. By the equation above, we have $P_0(T_0 < \infty)$, proving that the state 0 here is recurrent, and thus all states are recurrent.

Case 2. $p \neq 1/2$. If $p < 1/2$, then $P_{-1}(T_0 < \infty) < 1$, while if $p > 1/2$, then $P_1(T_0 < \infty) < 1$ (again, see Section 5.5). In either case, we have $P_0(T_0 < \infty) < 1$, proving that 0 is transient, and thus all states are transient.

6.8 Invariant distributions

Definition 6.31. A measure λ [not necessarily a probability measure] on I is said to be invariant with respect to a transition matrix p if $\lambda p = \lambda$.

Theorem 6.32. *If $(X_n)_{n\geq 0}$ is* Markov(λ, p) *and λ is invariant with respect to p, then for every $m \geq 1$, $(X_{m+n})_{n\geq 0}$ is again* Markov(λ, p).

Proof. Since $P(X_m = i) = (\lambda p^m)_i = \lambda_i$, we have

$$P(X_m = i_0, \ldots, X_{m+n} = i_n) = \lambda_{i_0} p_{i_0,i_1} \cdots p_{n-1,n}$$

for any $n \geq 0$ and $i_0, \ldots, i_n \in I$. □

Theorem 6.33. *Let I be finite. Assume there exists $i \in I$ and a vector $(\pi_j)_{j\in I}$ such that $p^n_{i,j} \to \pi_j$ as $n \to \infty$ for each $j \in I$. Then $(\pi_j)_{j\in I}$ is an invariant distribution.*

Proof. Because I is finite, we can interchange the limit and the finite sum.

$$\sum_{j\in I} \pi_j = \sum_{j\in I} \lim_{n\to\infty} p^n_{i,j} = \lim_{n\to\infty} \sum_{j\in I} p^n_{i,j} = 1,$$

so $(\pi_j)_{j\in I}$ is a distribution. To see it is invariant, note that

$$\pi_j = \lim_{n\to\infty} p^n_{i,j} = \lim_{n\to\infty} \sum_{k\in I} p^n_{i,k} p_{k,j} = \sum_{k\in I} \lim_{n\to\infty} p^n_{i,k} p_{k,j} = \sum_{k\in I} \pi_k p_{k,j}.$$

□

Example 6.34 (One-dimensional random walk)**.** Consider the setup in Example 6.30. For any $i, j \in I$ we have $p^n_{i,j} \to 0$. However, the all-zero vector is not an invariant distribution (although it is an invariant measure). This counterexample shows why the condition that I be finite is necessary in Theorem 6.33.

Definition 6.35. For $i, k \in I$, we define the expected time spent in i between visits to k.

$$\gamma^k_i := \mathrm{E}_k \left[\sum_{n=0}^{T_k - 1} \mathbf{1}_{\{X_n = i\}} \right].$$

Theorem 6.36. *If p is irreducible and recurrent, then the following hold for all $k \in I$.*

a) $\gamma^k_k = 1$.

b) $\gamma^k := (\gamma_i^k)_{i\in I}$ *satisfies* $\gamma^k p = \gamma^k$.

c) $0 < \gamma_i^k < \infty$ *for all* $i \in I$.

Proof. Since $\mathbf{1}_{\{X_n=k\}} \equiv 0$ for $0 < n < T_k - 1$ and $\mathrm{E}_k[\mathbf{1}_{\{X_n=k\}}] = 1$, part a) is clear.

We now show part b). Note that $\{n \leq T_k\} = \{T_k \leq n-1\}^c \in \mathcal{F}_{n-1}^X$, we have by the Markov property (Theorem 6.13)

$$P_k(X_{n-1} = i, X_n = j, n \leq T_k) = P_k(X_{n-1} = i, n \leq T_k)p_{i,j}.$$

By recurrence,

$$P(T_k < \infty, X_0 = X_{T_k} = k) = P(T_k < \infty) = 1,$$

so we have

$$\begin{aligned}
\gamma_j^k &:= \mathrm{E}_k\left[\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=j\}}\right] \\
&= \mathrm{E}_k\left[\sum_{n=1}^{T_k} \mathbf{1}_{\{X_n=j\}}\right] \\
&= \mathrm{E}_k\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=j\}}\mathbf{1}_{\{n\leq T_k\}}\right] \\
&= \sum_{n=1}^{\infty} P_k(X_n = j, n \leq T_k) \\
&= \sum_{i\in I}\sum_{n=1}^{\infty} P_k(X_{n-1} = i, X_n = j, n \leq T_k) \\
&= \sum_{i\in I} p_{i,j} \sum_{n=1}^{\infty} P_k(X_{n-1} = i, n \leq T_k) \\
&= \sum_{i\in I} p_{i,j}\, \mathrm{E}_k\left[\sum_{m=0}^{\infty} \mathbf{1}_{\{X_m=i, m\leq T_k-1\}}\right] \\
&=: \sum_{i\in I} \gamma_i^k p_{i,j},
\end{aligned}$$

proving b).

For each $i \in I$, there exist $m, n \geq 0$ such that $p_{i,k}^m$ and $p_{k,i}^n$ are strictly positive. Using parts a) and b), we have the following two inequalities which together imply c).

$$\begin{aligned}
\gamma_i^k &\geq \gamma_k^k p_{k,i}^n > 0, \\
\gamma_i^k p_{i,k}^m &\leq \gamma_k^k = 1.
\end{aligned}$$

□

Theorem 6.37. *Let* p *be irreducible and let* λ *be an invariant measure such that* $\lambda_k = 1$*. Then* $\lambda \geq \gamma^k$*. If* p *is also recurrent, then* $\lambda = \gamma^k$*.*

Proof. For all $j \in I$,

$$\begin{aligned}\lambda_j &= \sum_{i\in I} \lambda_i p_{i,j} \\ &= p_{k,j} + \sum_{i_1 \neq k} \lambda_{i_1} p_{i_1,j} \\ &= p_{k,j} + \sum_{i_1 \neq k} \sum_{i_2 \in I} \lambda_{i_2} p_{i_2,i_1} p_{i_1,j} \\ &= p_{k,j} + \sum_{i_1 \neq k} p_{k,i_1} p_{i_1,j} + \sum_{i_1,i_2 \neq k} \lambda_{i_2} p_{i_2,i_1} p_{i_1,j} \\ &\vdots \\ &= p_{k,j} + \sum_{i_1 \neq k} p_{k,i_1} p_{i_1,j} + \cdots + \sum_{i_1,\ldots,i_{n-1} \neq k} (p_{k,i_{n-1}} p_{i_{n-1},i_{n-2}} \cdots p_{i_1,j}) + \sum_{i_1,\ldots,i_n \neq k} (\lambda_{i_n} p_{i_n,i_{n-1}} \cdots p_{i_1,j}),\end{aligned}$$

for any $n \geq 1$. Ignoring the last term shows that for $j \neq k$,

$$\lambda_j \geq P_k(X_1 = j, T_k \geq 1) + \cdots + P_k(X_n = j, T_k \geq n).$$

Taking the limit on both sides as $n \to \infty$ gives $\lambda_j \geq \gamma_j^k$ when $j \neq k$. In the case $j = k$, we already know $\lambda_k = 1 = \gamma_k^k$.

If p is also recurrent, then λ^k is an invariant measure (Theorem 6.36), so $\mu := \lambda - \gamma^k$ is also an invariant measure, and we have already shown that $\mu \geq 0$ and $\mu_k = 0$. For any given $i \in I$ there exists $n \geq 0$ such that $p_{i,k}^n > 0$. Then,

$$0 = \mu_k = \sum_{j \in I} \mu_j p_{j,k}^n \geq \mu_i p_{i,k}^n,$$

which implies $\mu_i = 0$. □

Definition 6.38. Let $m_i := \mathrm{E}_i[T_i]$ be the **expected return time**. We call a recurrent state i **positive recurrent** if $m_i < \infty$, and we call all other recurrent states **null recurrent**.

Theorem 6.39. *If p is irreducible, then the following are equivalent.*

(i) Every state is positive recurrent.

(ii) Some state is positive recurrent.

(iii) p has an invariant distribution π.

If these statements hold, then $\pi_i = 1/m_i$ for all $i \in I$, and π is unique and strictly positive.

Proof. (i) $\Longrightarrow$ (ii) is clear.

(ii) $\Longrightarrow$ (iii). If $i \in I$ is positive recurrent, then p is recurrent (Theorem 6.26) and γ^i is an invariant measure (Theorem 6.36). Since

$$\sum_{j \in I} \gamma_j^i = \sum_{j \in I} \mathrm{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_n = j\}} \right] = \mathrm{E}_i[T_i] = m_i < \infty,$$

we see that $\pi_j := \gamma_j^i / m_i$ is an invariant distribution.

(iii) $\Longrightarrow$ (i). Fix any $k \in I$. Because p is irreducible, there exists $n \geq 0$ such that $\pi_k = \sum_{i \in I} \pi_i p_{i,k}^n > 0$. If we let $\lambda_i := \pi_i / \pi_k$, then λ is an invariant measure with $\lambda_k = 1$. By Theorem 6.37, $\lambda \geq \gamma^k$, so

$$m_k = \sum_{i \in I} \gamma_i^k \leq \sum_{i \in I} \frac{\pi_i}{\pi_k} = \frac{1}{\pi_k} < \infty,$$

showing that k is positive recurrent.

If we know that all three statements hold, then we know p is recurrent, so Theorem 6.37 implies $\lambda = \gamma^k$. Replacing the analogous inequalities in the proof of "(iii) $\Longrightarrow$ (i)" with equalities gives $0 < \pi_k = 1/m_k$. □

Corollary 6.40. *If p is irreducible and I is finite, then p has a unique invariant distribution π that is strictly positive. Moreover, all states are positive recurrent.*

Proof. Since I is a finite closed communicating class, it is recurrent (Theorem 6.28), so $(\gamma_i^k)_{i\in I}$ is an invariant finite measure (Theorem 6.36). Normalizing by $\sum_{i\in I}\gamma_i^k$ gives an invariant distribution. □

Example 6.41 (Symmetric one-dimensional random walk)**.** We use the setup in Example 6.30 with $p = 1/2$. The transition matrix is irreducible and recurrent. The measure $\pi_i = 1$ for all $i \in I$ is an invariant measure. An invariant distribution would have to be a multiple of π (consequence of Theorem 6.37), but $\sum_{i\in I}\pi_i = \infty$, so there exists no invariant distribution, and thus every state is null recurrent.

6.9 Convergence to equilibrium

Example 6.42. Let I have two states and let

$$p := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then p^{2n} would be the identity and $p^{2n+1} = p$ for all n. Clearly $\pi := \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$ is an invariant distribution with respect to p. However, $p_{i,j}^n$ does not converge as $n \to \infty$ for any $i, j \in I$.

It turns out that the periodicity of the above example is the only issue preventing convergence of p^n.

Definition 6.43. A state $i \in I$ is **aperiodic** if $p_{i,i}^n > 0$ for all sufficiently large n.

Lemma 6.44. *Let p be an irreducible transition matrix and $i \in I$ an aperiodic state. Then for any $j, k \in I$, $p_{j,k}^n > 0$ for all sufficiently large n. In particular, all states are aperiodic.*

Proof. There exist $m, \ell \geq 0$ such that $p_{j,i}^m$ and $p_{i,k}^n$ are strictly positive. Then $p_{j,k}^{m+n+\ell} \geq p_{j,i}^m p_{i,i}^n p_{i,k}^\ell > 0$ for all sufficiently large n. Taking $j = k$ shows that j is aperiodic. □

Theorem 6.45 (Convergence to equilibrium)**.** *Let π be an invariant distribution of an irreducible aperiodic transition matrix p. Let λ be an initial distribution and let $(X_n)_{n\geq 0}$ be* Markov(λ, p). *Then*

$$P(X_n = j) \to \pi_j$$

as $n \to \infty$. In particular,

$$p_{i,j}^n \to \pi_j$$

as $n \to \infty$.

Proof. Let $(Y_n)_{n\geq 0}$ be Markov(π, p) and independent of $(X_n)_{n\geq 0}$. Fix $b \in I$ and let

$$T := \inf\{n \geq 0 : X_n = Y_n = b\}.$$

We first show $P(T < \infty) = 1$. To do this, we show that $W_n := (X_n, Y_n)$ is a Markov chain on $I \times I$ with transition matrix

$$\widehat{p}_{(i,k),(j,\ell)} := p_{i,j}p_{k,\ell}$$

and initial distribution

$$\mu(i,k) := \lambda_i \pi_k.$$

Since $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ are independent and since p is aperiodic for all states in I, we have for all $i, j, k, \ell \in I$

$$\widehat{p}_{(i,k),(j,\ell)}^n = p_{i,j}^n p_{k,\ell}^n > 0$$

for all sufficiently large n, implying that $\widehat{p}$ is irreducible. Since $\widehat{p}$ has the invariant distribution

$$\widehat{\pi}_{(i,k)} := \pi_i \pi_k,$$

Theorem 6.39 implies that $\widehat{p}$ is positive recurrent, i.e., $P(T < \infty) = 1$.

Next we use a technique called coupling. We define the process

$$Z_n := \begin{cases} X_n & n < T, \\ Y_n & n \geq T. \end{cases}$$

By the strong Markov property (Theorem 6.20), we see that $(Z_n)_{n\geq 0}$ is Markov(λ, p).

To conclude, note that

$$P(Z_n = j) = P(X_n = j, n < T) + P(Y_n = j, n \geq T),$$

so

$$\begin{aligned} |P(X_n = j) - \pi_j| &= |P(Z_n = j) - P(Y_n = j)| \\ &= |P(X_n = j, n < T) - P(Y_n = j, n < T)| \\ &\leq P(n < T) \\ &\to 0 \end{aligned}$$

as $n \to \infty$. [The first equality holds because $(X_n)_{n\geq 0}$ and $(Z_n)_{n\geq 0}$ follow the same distribution, and because π is an invariant distribution for p and is the initial distribution for $(Y_n)_{n\geq 0}$. The last inequality is simply $|P(A \cap B) - P(A \cap C)| \leq \max(P(A \cap B), P(A \cap C)) \leq P(A)$.] □

6.10 Ergodic theorem

The ergodic theorem for Markov chains can be seen as a generalization of the strong law of large numbers.

Theorem 6.46 (Ergodic theorem)**.** *Let $(X_n)_{n\geq 0}$ be* Markov(λ, p)*, with p irreducible, and let $V_i(n) := \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k = i\}}$.*

a)

$$\frac{V_i(n)}{n} \to \frac{1}{m_i}$$

almost surely as $n \to \infty$. [This statement is still valid if $m_i = \infty$, in which case the point of convergence is 0.]

b) If p is also positive recurrent and π is an invariant distribution, then

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \sum_{i \in I} \pi_i f(i)$$

almost surely as $n \to \infty$, for any bounded function $f : I \to \mathbb{R}$.

Proof. If p is transient, then $V_i := \sum_{n\geq 0} \mathbf{1}_{\{X_n = i\}}$ is finite almost surely and $m_i = \infty$, so

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \to 0 = \frac{1}{m_i}$$

proves a) in this case.

Let p be recurrent and fix a state i. Then $P(T_i < \infty) = 1$ (Theorem 6.25), and $(X_{T_i+n})_{n\geq 0}$ is Markov(δ_i, p) and is independent of $X_0, X_1, \dots, X_{T_i}$ by the strong Markov property (Theorem 6.20). The long-run proportion $\lim_{n\to\infty} V_i(n)/n$ of time spent in i is the same for $(X_{T_i+n})_{n\geq 0}$ and $(X_n)_{n\geq 0}$, so we may assume $\lambda = \delta_i$.

Let S_i^k be as defined in Definition 6.22. The nonnegative random variables $S_i^1, S_i^2, \dots$ are i.i.d. (Lemma 6.23) and $\mathrm{E}_i[S_i^k] = m_i$ for all k. We have

$$S_i^1 + \cdots + S_i^{V_i(n)-1} = T_i^{V_i(n)-1} \leq n - 1$$

because the left-hand side is the time of the last visit to i before n. Similarly,

$$S_i^1 + \cdots + S_i^{V_i(n)} = T_i^{V_i(n)} \geq n$$

because the left-hand side is the time of the first visit to i after $n-1$. [Note that $V_i(n)$ counts the "visit" to i at time 0, while T_i^n does not.] Thus,

$$\frac{S_i^1 + \cdots + S_i^{V_i(n)-1}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{S_i^1 + \cdots + S_i^{V_i(n)}}{V_i(n)}.$$

By the strong law of large numbers (Theorem 5.53),

$$\frac{S_i^1 + \cdots + S_i^n}{n} \to m_i$$

almost surely as $n \to \infty$. Because p is recurrent, $V_i(n) \to \infty$ almost surely as $n \to \infty$. Thus our upper and lower bound on $n/V_i(n)$ give

$$\frac{n}{V_i(n)} \to m_i$$

almost surely as $n \to \infty$. Rearranging proves a).

Suppose $(X_n)_{n\geq 0}$ has an invariant distribution $(\pi_i)_{i\in I}$, and let $f : I \to \mathbb{R}$ be a bounded function. Without loss of generality we may assume $|f| \leq 1$. For any $J \subset I$, we have

$$\begin{aligned}
\left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \sum_{i\in I} \pi_i f(i) \right| &= \left| \sum_{i \in I} \left(\frac{V_i(n)}{n} - \pi_i \right) f(i) \right| \\
&\leq \sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \sum_{i \notin J} \left| \frac{V_i(n)}{n} - \pi_i \right| \\
&\leq \sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \sum_{i \notin J} \left(\frac{V_i(n)}{n} + \pi_i \right) \\
&= \sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \left(1 - \sum_{i \in J} \frac{V_i(n)}{n} \right) + \sum_{i \notin J} \pi_i \qquad \sum_{i\in I} V_i(n) = n \\
&= \sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \sum_{i \in J} \left(\pi_i - \frac{V_i(n)}{n} \right) + 2 \sum_{i \notin J} \pi_i \qquad \sum_{i\in I} \pi_i = 1 \\
&\leq 2 \sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + 2 \sum_{i \notin J} \pi_i.
\end{aligned}$$

By part a), $V_i(n)/n \to \pi_i$ almost surely as $n \to \infty$ for all i. Given $\epsilon > 0$, choose J finite so that $\sum_{i \notin J} \pi_i < \epsilon/4$. Because J is finite, for almost all $\omega \in \Omega$ we can choose an integer $N(\omega)$ large enough so that for $n \geq N(\omega)$ we have

$$\sum_{i\in J} \left| \frac{V_i(n)}{n} - \pi_i \right| < \epsilon/4.$$

Then for $n \geq N(\omega)$ we have

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \sum_{i\in I} \pi_i f(i) \right| < \epsilon,$$

proving b). □

7 Poisson processes

Let $S_1, S_2, \ldots$ be i.i.d. Expon(λ). Let $T_n := S_1 + S_2 + \cdots + S_n$ for each $n \geq 1$. The **Poisson process** with jump intensity λ is defined as $(N_t)_{t \in \mathbb{R}}$, where

$$N_t := \sum_{n \geq 1} \mathbf{1}_{\{T_n \leq t\}}.$$

- $N_t \sim \text{Poisson}(\lambda t)$.
- $T_n \sim \text{Gamma}(n, \lambda)$ (density function $\lambda e^{-\lambda t}(\lambda t)^{n-1}/(n-1)!$).
- $(N_t)_{t \in \mathbb{R}}$ has stationary increments ($N_t - N_s$ depends only on $t - s$) and independent increments.
- $(N_t - \lambda t)_{t \in \mathbb{R}}$ is a martingale.

References

[1] David Williams. **Probability with Martingales.** Cambridge University Press, 1991.

[2] J.R. Norris. **Markov Chains.** Cambridge University Press, 1998.