Notes

EE 226A

Druv Pai

Fall 2021

Contents

1 Elements of Probability Theory

1.1 Probability Spaces and Events

1.1.1 Basic Definitions

Definition 1.1.1. A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbf{P})$. Here, Ω is a set, called the **sample space**; the $\omega \in \Omega$ are called **samples**. $\mathcal{F}$ is a collection of events; in particular, $\mathcal{F} \subseteq 2^{\Omega}$. $\mathbf{P}$ is a "probability measure", in particular, $\mathbf{P} \colon \mathcal{F} \to [0,1]$.

In particular, for $(\Omega, \mathcal{F}, \mathbf{P})$ to be a probability space, we need $\mathcal{F}$ to be a σ-algebra.

Definition 1.1.2. A set of sets $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra if

A1. $\Omega \in \mathcal{F}$.

A2. $\mathcal{F}$ is closed under complements. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

A3. $\mathcal{F}$ is closed under countable unions. If $(A_n)_{n \in \mathbb{N}}$ is a sequence of sets in $\mathcal{F}$ then $A = \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$.

Proposition 1.1.3. If $\mathcal{F}$ is a σ-algebra, then is closed under countable intersections.

Proof. DeMorgan's laws. □

Also, we need $\mathbf{P}$ to be a probability measure.

Definition 1.1.4. Probability measures $\mathbf{P}$ satisfy "Kolmogorov axioms":

A1. $\mathbf{P}(A) \geq 0$ for all $A \in \mathcal{F}$.

A2. $\mathbf{P}(\Omega) = 1$.

A3. (σ-additivity.) If $(A_n)_{n \in \mathbb{N}}$ is a sequence of *disjoint* sets in $\mathcal{F}$ then $\mathbf{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mathbf{P}(A_n)$.

Example 1.1.5. A sample space is **discrete** if Ω is countable. For example, for a coin flip $\Omega = \{\mathtt{H}, \mathtt{T}\}$. Rolling a dice n times has $\Omega = [6]^n$. A score in a basketball match might have $\Omega = \mathbb{N}^2$. Valid σ-algebras include $\mathcal{F} = 2^{\Omega}$, $\mathcal{F} = \{\emptyset, \Omega\}$, or more complicated σ-algebras. In the dice example above, we could possibly take $\mathcal{F} = \{\emptyset, \{\text{sum of rolls is even}\}, \{\text{sum of rolls is odd}\}, \Omega\}$.

Example 1.1.6. A sample space is **continuous** if it is not discrete. For example, $\Omega = [0,1)$, $\Omega = [0, +\infty)$, and so on.

If we consider discrete problems, they can sometimes have an underlying continuous sample space.

Example 1.1.7. If the underlying event is an infinite sequence of coin flips, Ω is the set of infinitely long binary strings, which are each associated with a number in $[0,1)$.

1.1.2 Existence and Construction of Probability Measures

Definition 1.1.8. The smallest σ-algebra containing a set $\mathcal{G}$ is denoted $\sigma(\mathcal{G})$. In particular, any σ-algebra containing $\mathcal{G}$ also contains $\sigma(\mathcal{G})$.

Example 1.1.9. We want to pick a random number in $[0,1)$. Let $\mathcal{G} = \{[a,b), 0 \leq a \leq b < 1\}$ and set $\mathbf{P}([a,b)) = b-a$. We want to know whether there exists a valid extension of $\mathbf{P}$ to a probability measure. The answer is "yes", and we use a tool called **Carathéodory's Extension Theorem** to do this extension. Given some technical conditions (see homework), $\mathbf{P}$ can be extended to a probability measure.

1.1.3 Properties of Probability Spaces

Theorem 1.1.10. If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, then it enjoys the following properties:

(a) Monotonicity: If $A, B \in \mathcal{F}$, $A \subseteq B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(b) Countable sub-additivity: If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$ and $A = \bigcup_{n\in\mathbb{N}} A_n$, then $\mathbf{P}(A) \leq \sum_{n\in\mathbb{N}} \mathbf{P}(A_n)$.

(c) Continuity from below: If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$ that increases ($A_n \subseteq A_{n+1}$), and $A = \bigcup_{n\in\mathbb{N}} A_n$, then $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

(d) Continuity from above: If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$ that decreases ($A_n \supseteq A_{n+1}$), and $A = \bigcap_{n\in\mathbb{N}} A_n$, then $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

Proof.

Proof of (a) Write $B = A \cup (B \setminus A)$. Then

$$\mathbf{P}(A) \leq \mathbf{P}(A) + \mathbf{P}(B \setminus A) = \mathbf{P}(B). \tag{1.1.1}$$

Proof of (b) Define $B_1 = A_1$ and $B_n = A_n \setminus \left(\bigcup_{i\in[n-1]} A_i\right)$. Then $\bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} A_n$. Hence

$$\mathbf{P}(A) = \sum_{n\in\mathbb{N}} \mathbf{P}(B_n) \leq \sum_{n\in\mathbb{N}} \mathbf{P}(A_n). \tag{1.1.2}$$

Proof of (c) Define $B_1 = A_1$ and $B_n = A_n \setminus \left(\bigcup_{i\in[n-1]} A_i\right)$. Then (by the fact that a monotone bounded sequence has a limit),

$$\mathbf{P}(A) = \sum_{n\in\mathbb{N}} \mathbf{P}(B_n) = \lim_{N\to\infty} \sum_{n\in[N]} \mathbf{P}(B_n) = \lim_{N\to\infty} \mathbf{P}(A_N). \tag{1.1.3}$$

Proof of (d) Let $B_n = A_n^c$. Then (B_n) is an increasing sequence of events whose union is $B = A^c$. Hence

$$\mathbf{P}(B) = \lim_{n\to\infty} \mathbf{P}(B_n)$$

and using that $B = \Omega \setminus A$ and $B_n = \Omega \setminus A_n$,

$$1 - \mathbf{P}(A) = \mathbf{P}(\Omega \setminus A) = \lim_{n\to\infty} \mathbf{P}(\Omega \setminus A_n) = \lim_{n\to\infty} (1 - \mathbf{P}(A_n)) \tag{1.1.4}$$

from which the conclusion follows.

□

1.1.4 Infinitely Often and Borel-Cantelli Lemmas

Definitions

Generally speaking, we construct complicated events from simple events.

Definition 1.1.11. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$, the event

$$\{A_n \text{ infinitely often}\} = \{A_n \text{ i.o.}\} = \bigcap_{n\in\mathbb{N}} \bigcup_{m\in\mathbb{N}\setminus[n-1]} A_m = \limsup_{n\to\infty} A_n. \tag{1.1.5}$$

If ω is in infinitely many A_n's, then $\omega \in \bigcup_{m\in\mathbb{N}\setminus[n]} A_m$ for each n, so $\omega \in \{A_n \text{ i.o.}\}$. If ω is in finitely many A_n's, then there exists N such that $\omega \notin \bigcup_{m\in\mathbb{N}\setminus[N]} A_m$, so that $\omega \notin \{A_n \text{ i.o.}\}$.

Definition 1.1.12. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$, the event

$$\{A_n \text{ eventually}\} = \{A_n \text{ ev.}\} = \bigcup_{n\in\mathbb{N}} \bigcap_{m\in\mathbb{N}\setminus[n-1]} A_m = \liminf_{n\to\infty} A_n. \tag{1.1.6}$$

This is the same as saying ω is in all but finitely many of the A_n's.

Borel-Cantelli Lemmas and Independence of Events

Lemma 1.1.13 (First Borel-Cantelli Lemma). If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in $\mathcal{F}$ and $\sum_{n\in\mathbb{N}} \mathbf{P}(A_n) < \infty$, then

$$\mathbf{P}\left(\limsup_{n\to\infty} A_n\right) = 0. \tag{1.1.7}$$

Proof. Let $B_1 = A_1$ and $B_n = \bigcup_{m\in\mathbb{N}\setminus[n]} A_m$. Then $(B_n)_{n\in\mathbb{N}}$ is a decreasing sequence that converges to $B = \bigcap_{n\in\mathbb{N}} B_n$. Note that $B = \limsup_{n\to\infty} A_n$. Then

$$\mathbf{P}\left(\limsup_{n\to\infty} A_n\right) = \mathbf{P}(B) = \lim_{n\to\infty} \mathbf{P}(B_n) = \lim_{n\to\infty} \mathbf{P}\left(\bigcup_{m\in\mathbb{N}\setminus[n]} A_m\right) \tag{1.1.8}$$

$$\leq \lim_{n\to\infty} \sum_{m\in\mathbb{N}\setminus[n]} \mathbf{P}(A_m) = 0. \tag{1.1.9}$$

□

Example 1.1.14. Let $A_n = \{\text{pandas procreate in year } 2000 + n\}$. Assuming $\mathbf{P}(A_n) \leq \frac{1}{n^2}$, then pandas will go extinct with probability 1 according to Lemma 1.1.13.

Definition 1.1.15. Events $(A_n)_{n\in\mathbb{N}}$ are **independent** if, for every finite subset $S \subset \mathbb{N}$,

$$\mathbf{P}\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} A_i{}'. \tag{1.1.10}$$

Proposition 1.1.16. If $(A_n)_{n\in\mathbb{N}}$ are independent events, then $(A_n^c)_{n\in\mathbb{N}}$ are also independent events.

Proof. Homework. □

Lemma 1.1.17 (Second Borel-Cantelli Lemma). If $(A_n)_{n\in\mathbb{N}}$ are independent events, and $\sum_{n\in\mathbb{N}} \mathbf{P}(A_n) = \infty$, then

$$\mathbf{P}\left(\limsup_{n\to\infty} A_n\right) = 1. \tag{1.1.11}$$

Proof. Using the same trick as before,

$$\begin{aligned}
\mathbf{P}\left(\limsup_{n\to\infty} A_n\right) &= \mathbf{P}\left(\bigcap_{n\in\mathbb{N}} \bigcup_{m\in\mathbb{N}\setminus[n]} A_m\right) = \lim_{n\to\infty} \mathbf{P}\left(\bigcup_{m\in\mathbb{N}\setminus[n]} A_i\right) \tag{1.1.12}\\
&= \lim_{n\to\infty}\left(1-\mathbf{P}\left(\bigcap_{m\in\mathbb{N}\setminus[n]} A_m^c\right)\right) \tag{1.1.13}\\
&= 1-\lim_{n\to\infty}\mathbf{P}\left(\bigcap_{m\in\mathbb{N}\setminus[n]} A_m^c\right) \tag{1.1.14}\\
&= 1-\lim_{n\to\infty}\prod_{m\in\mathbb{N}\setminus[n]}\mathbf{P}(A_m^c) \tag{1.1.15}\\
&= 1-\lim_{n\to\infty}\prod_{m\in\mathbb{N}\setminus[n]}(1-\mathbf{P}(A_m)) \tag{1.1.16}\\
&= 1-\lim_{n\to\infty}\lim_{k\to\infty}\prod_{m\in[k]\setminus[n-1]}(1-\mathbf{P}(A_m)) \tag{1.1.17}\\
&\le 1-\lim_{n\to\infty}\lim_{k\to\infty}\exp\left(-\sum_{m\in[k]\setminus[n-1]}\mathbf{P}(A_m)\right) \tag{1.1.18}\\
&\le 1-\lim_{n\to\infty} 0 = 1. \tag{1.1.19}
\end{aligned}$$

□

1.2 Random Variables and Expectation

1.2.1 Random Variables

Definition 1.2.1 (Random Variable). A **random variable** is a function $X\colon \Omega \to \overline{\mathbb{R}}$ such that X is $\mathcal{F}$-measurable.

Corollary 1.2.2. The events $\{\omega\colon X(\omega) \le \alpha\} \in \mathcal{F}$ for $\alpha \in \overline{\mathbb{R}}$.

We can compute probabilities:

$$\mathbf{P}(X \le \alpha) = \mathbf{P}(\{\omega\colon X(\omega) \le \alpha\}) \quad \text{for all } \alpha \in \overline{\mathbb{R}}. \tag{1.2.1}$$

From homework, it is ensured that $\{X \in B\} = \{\omega\colon X(\omega) \in B\} = X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}_{\mathbb{R}}$ the **Borel σ-algebra on** $\mathbb{R}$ (σ-algebra generated by the open sets). Hence, we are justified in computing probabilities $\mathbf{P}(X \in B)$ for any Borel set B.

Random variables are well-behaved under "natural" operations.

Proposition 1.2.3. If X, Y are random variables, then

(a) $|X|^p$ is a random variable for all $p \in \mathbb{R}$;

(b) $X + Y$ is a random variable;

(c) XY is a random variable.

Proof. We do the proof assuming X and Y are real-valued, since this simplifies it slightly.

Proof of (a) Homework.

Proof of (b) Note that

$$\{X+Y>\alpha\} = \bigcup_{q\in\mathbb{Q}} \underbrace{\{X>q\}}_{\in\mathcal{F}} \cup \underbrace{\{Y>\alpha-q\}}_{\in\mathcal{F}}. \tag{1.2.2}$$

Hence $\{X+Y>\alpha\}\in\mathcal{F}$ and its complement $\{X+Y\le\alpha\}\in\mathcal{F}$.

Proof of (c) It follows from the expression

$$XY = \frac{|X+Y|^2 - |X-Y|^2}{4}. \tag{1.2.3}$$

□

Random variables are also closed under certain operations such as taking limits. This is important to the study of stochastic processes.

Proposition 1.2.4. If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables defined on a common probability space $(\Omega,\mathcal{F},\mathbf{P})$, then

1. $\sup_{n\in\mathbb{N}} X_n$ and $\inf_{n\in\mathbb{N}} X_n$ are random variables;
2. $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are random variables;
3. if $\lim_{n\to\infty} X_n$ exists, it is also a random variable.

Proof.

Proof of (a) Note that $\sup_{n\in\mathbb{N}} X_n(\omega)\le\alpha$ if and only if $X_n(\omega)\le\alpha$ for each $n\in\mathbb{N}$. Then

$$\left\{\sup_{n\in\mathbb{N}} X_n \le \alpha\right\} = \bigcap_{n\in\mathbb{N}} \underbrace{\{X_n\le\alpha\}}_{\in\mathcal{F}}. \tag{1.2.4}$$

Hence $\sup_{n\in\mathbb{N}} X_n$ is $\mathcal{F}$-measurable and hence a random variable. Note that $\inf_{n\in\mathbb{N}} X_n = -\sup_{n\in\mathbb{N}}(-X_n)$ is thus also $\mathcal{F}$-measurable and hence a random variable.

Proof of (b) Note that

$$\limsup_{n\to\infty} X_n = \inf_{m\in\mathbb{N}} \sup_{n\ge m} X_n \tag{1.2.5}$$

and

$$\liminf_{n\to\infty} X_n = \sup_{m\in\mathbb{N}} \inf_{n\ge m} X_n. \tag{1.2.6}$$

Hence $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are random variables by (a).

Proof of (c) If $\lim_{n\to\infty} X_n$ exists then

$$\lim_{n\to\infty} X_n = \liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n \tag{1.2.7}$$

and hence $\lim_{n\to\infty} X_n$ is a random variable by (b).

□

1.2.2 Almost-Sure Equivalence

Definition 1.2.5 (Almost Sure Equivalence). If X and Y are random variables and $\mathbf{P}(X = Y) = 1$, then we say $X = Y$ **P-almost surely (a.s.)**.

Example 1.2.6. Let X and Y be random variables on

$$\left(\Omega = [0,1), \mathcal{F} = \mathcal{B}_{[0,1)}, \mathbf{P} = \text{uniform measure on } \Omega\right) \tag{1.2.8}$$

Let

$$X(\omega) = 0 \quad \text{for all } \omega \in \Omega \qquad \text{Y} \qquad (\omega) = \begin{cases} 0 & \omega \in \mathbb{Q} \\ 1 & \omega \notin \mathbb{Q} \end{cases}. \tag{1.2.9}$$

Then

$$\mathbf{P}(X \neq Y) = \mathbf{P}(\mathbb{Q} \cap [0,1)) = \sum_{q \in \mathbb{Q} \cap [0,1)} \mathbf{P}(\{q\}) = 0. \tag{1.2.10}$$

Often, we define limits modulo a.s. equivalence.

Example 1.2.7. Let X and $(X_n)_{n \in \mathbb{N}}$ be random variables on

$$\left(\Omega = [0,1), \mathcal{F} = \mathcal{B}_{[0,1)}, \mathbf{P} = \text{uniform measure on } \Omega\right). \tag{1.2.11}$$

Let

$$X_n(\omega) = \begin{cases} (-1)^n & \omega \in \mathbb{Q} \\ X(\omega) + \frac{1}{n} & \omega \notin \mathbb{Q} \end{cases}. \tag{1.2.12}$$

Then $\lim_{n \to \infty} X_n(\omega)$ exists $\mathbf{P}$-a.s. and is equal to X.

1.2.3 Distributions

Associated to every random variable is its **distribution** or **law**.

Definition 1.2.8 (Distribution). For a random variable X on $(\Omega, \mathcal{F}, \mathbf{P})$, its distribution (or law) is a probability measure on $\left(\overline{\mathbb{R}}, \mathcal{B}_{\overline{\mathbb{R}}}\right)$ (to be precise, $L_X \colon \mathcal{B}_{\overline{\mathbb{R}}} \to [0,1]$) defined by

$$L_X(B) = \mathbf{P}(X \in B) \quad \text{for all } B \in \mathcal{B}_{\overline{\mathbb{R}}} \quad . \tag{1.2.13}$$

In technical terms, it is the pushforward measure of X w.r.t. $\mathbf{P}$.

Definition 1.2.9 (Distribution Function). For a random variable X, we define its **distribution function** $F_X \colon \overline{\mathbb{R}} \to [0,1]$ as

$$F_X(x) = L_X([-\infty, x]) = \mathbf{P}(X \leq x) \quad \text{for all } x \in \overline{\mathbb{R}}.$$

Evidently L_X determines F_X. It turns out that F_X determines L_X as well, since one can find L_X on intervals from F_X and use Carathéodory's Extension Theorem.

Theorem 1.2.10. The function $F \colon \overline{\mathbb{R}} \to [0,1]$ is a distribution function for a random variable X if

(a) F is non-decreasing;

(b) F is right-continuous, i.e., $\lim_{y \to x^+} F(y) = F(x)$;

If it moreover holds that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, then F is the distribution function of a $\mathbf{P}$-a.s. real-valued random variable.

Proof Sketch. Assume X is a random variable and let F be its distribution function. To show (a),

$$F(x) = \mathbf{P}(X \le x) \le \mathbf{P}\big(X \le x'\big) = F\big(x'\big) \quad \text{for all } x, x' \in \overline{\mathbb{R}} \text{ with } x \le x'. \tag{1.2.14}$$

To show (b), use continuity from above to show that for any positive sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ whose limit is 0,

$$\begin{aligned} \{X \le x\} &= \bigcap_{n\in\mathbb{N}} \{X \le x + \varepsilon_n\} && (1.2.15) \\ \implies F_X(x) &= \mathbf{P}(X \le x) && (1.2.16) \\ &= \lim_{n\to\infty} \mathbf{P}(X \le \varepsilon_n) && (1.2.17) \\ &= \lim_{n\to\infty} F(x + \varepsilon_n) && (1.2.18) \\ &= \lim_{y\to x^+} F(y). && (1.2.19) \end{aligned}$$

To go the other direction, fix F satisfying (a) and (b). We need to construct a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable X with F its distribution function.

Let $\mathbf{P}$ be the uniform distribution on $\big(\Omega = [0,1), \mathcal{F} = \mathcal{B}_{[0,1)}\big)$. Define

$$X(\omega) = \sup\{y \colon F(y) < \omega\}. \tag{1.2.20}$$

Then we can show $F_X = F$ through some tedium. □

Definition 1.2.11 (Discrete Random Variable). A distribution of random variable X is **discrete** if its distribution function is a step function. That is, there exists a countable set $(x_n)_{n\in\mathbb{N}}$ and $p_X \colon \mathbb{R} \to [0,1]$ such that

$$F_X(x) = \sum_{n\in\mathbb{N}} p_X(x_i) 1_{[x_i,\infty]}(x) \tag{1.2.21}$$

Definition 1.2.12 (Continuous Random Variable). A distribution of random variable X is **continuous** if its distribution function is absolutely continuous. That is, there exists a function $f_X \colon \mathbb{R} \to [0, +\infty)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, \mathrm{d}u. \tag{1.2.22}$$

In this case f_X is the **density** of X.

1.2.4 Random Vectors and Independence

A collection of random variables $(X_i)_{i\in[n]}$ on a common probability space is called a **random vector**. The distribution is the Borel probability measure:

$$\begin{aligned} L_X(B) &= \mathbf{P}((X_1, \ldots, X_n) \in B) \quad \text{for all } B \in \mathcal{B}_{\overline{\mathbb{R}}^n} && (1.2.23) \\ F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \mathbf{P}(X_1 \le x_1, \ldots, X_n \le x_n) \quad \text{for all } (x_1, \ldots, x_n) \in \overline{\mathbb{R}}^n. && (1.2.24) \end{aligned}$$

Definition 1.2.13 (Independence of Sequences of Random Variables). The random variables $(X_i)_{i\in[n]}$ are **independent** if

$$\mathbf{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i\in[n]} \mathbf{P}(X_i \in B_i) \quad \text{for all } B_i \in \mathcal{B}_{\overline{\mathbb{R}}}. \tag{1.2.25}$$

A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ are independent if all finite subsets of the sequence are independent.

Corollary 1.2.14. The random variables $(X_i)_{i\in[n]}$ are independent if

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i\in[n]} F_{X_i}(x_i). \tag{1.2.26}$$

1.2.5 Expectation

Expectation of Indicator Random Variables

Definition 1.2.15. The **indicator random variable** of a set $A \in \mathcal{F}$ is the function

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}. \tag{1.2.27}$$

We define the operator $\mathbf{E}$: indicator random variables $\to \overline{\mathbb{R}}$ by

$$\mathbf{E}(1_A) = \mathbf{P}(A). \tag{1.2.28}$$

Expectation of Simple Random Variables

Definition 1.2.16. Simple random variables are linear combinations of indicators (without loss of generality, over disjoint events).

We define the operator $\mathbf{E}$: simple random variables $\to \overline{\mathbb{R}}$ by:

$$\mathbf{E}\left(\sum_{i\in[n]} a_i 1_{A_i}\right) = \sum_{i\in[n]} a_i \, \mathbf{P}(A_i). \tag{1.2.29}$$

Expectation of Non-Negative Random Variables

Let $X \geq 0$ be a random variable. We define the operator $\mathbf{E}$: non-negative random variables $\to \overline{\mathbb{R}}$ by

$$\mathbf{E}(X) = \sup \left\{ \mathbf{E}(S) \colon 0 \leq S \leq X,\ S \text{ is a simple random variable} \right\}. \tag{1.2.30}$$

Lemma 1.2.17 (Fatou's Lemma). Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of non-negative random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. Then

$$\mathbf{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbf{E}(X_n). \tag{1.2.31}$$

Proposition 1.2.18. Let X, Y be non-negative random variables. Then

(a) (Uniqueness.) If $X = Y$ $\mathbf{P}$-a.s., then $\mathbf{E}(X) = \mathbf{E}(Y)$.

(b) (Monotonicity.) If $X \leq Y$ $\mathbf{P}$-a.s., then $\mathbf{E}(X) \leq \mathbf{E}(Y)$.

Proof.

Proof of (a). Let $E = \{X \neq Y\}$. Then $\mathbf{P}(E) = 0$. Take any simple $S \leq X$ such that $S = \sum_{i\in[n]} a_i 1_{A_i}$. Then

$$\begin{aligned} \mathbf{E}(S) &= \sum_{i\in[n]} a_i \, \mathbf{P}(A_i) = \sum_{i\in[n]} a_i \left(\mathbf{P}(A_i \cap E) + \mathbf{P}(A_i \cap E^c) \right) && (1.2.32) \\ &= \sum_{i\in[n]} a_i \, \mathbf{P}(A_i \cap E^c) && (1.2.33) \\ &= \mathbf{E}([S]_{E^c}). && (1.2.34) \end{aligned}$$

Then

$$\begin{aligned} \mathbf{E}(X) &= \sup \left\{ \mathbf{E}(S) \colon 0 \leq S \leq X, S \text{ is a simple function} \right\} && (1.2.35) \\ &= \sup \left\{ \mathbf{E}(S) \colon 0 \leq S \leq X, S \text{ is a simple function}, \operatorname{Supp}(S) \subseteq E^c \right\} && (1.2.36) \end{aligned}$$

$$
\begin{aligned}
&= \sup \{\mathbf{E}(S) \colon 0 \le S \le Y, S \text{ is a simple function}, \operatorname{Supp}(S) \subseteq E^c\} && (1.2.37)\\
&= \sup \{\mathbf{E}(S) \colon 0 \le S \le Y, S \text{ is a simple function}\} && (1.2.38)\\
&= \mathbf{E}(Y). && (1.2.39)
\end{aligned}
$$

Proof of (b). Without loss of generality, by the same argument in (1), we can assume that $X \le Y$ pointwise. Then

$$
\begin{aligned}
\mathbf{E}(X) &= \sup \{\mathbf{E}(S) \colon 0 \le S \le X, S \text{ is a simple function}\} && (1.2.40)\\
&\le \sup \{\mathbf{E}(S) \colon 0 \le S \le Y, S \text{ is a simple function}\} && (1.2.41)\\
&= \mathbf{E}(Y) && (1.2.42)
\end{aligned}
$$

where the inequality in the first line to the second line is because the first set is a subset of the second.

□

Theorem 1.2.19 (Monotone Convergence Theorem). Let $(X_n)_{n\in\mathbb{N}}$ be an essentially increasing sequence of non-negative random variables, i.e., $0 \le X_1 \le X_2 \le \ldots$. Then

$$\lim_{n\to\infty} \mathbf{E}(X_n) = \mathbf{E}\Big(\lim_{n\to\infty} X_n\Big). \tag{1.2.43}$$

Proof. By excising the union of all the "error" sets $E = \bigcup_{n\in\mathbb{N}} \{X_n \ne X_{n+1}\}$ which occurs with probability 0 due to countable additivity, we can consider the $(X_n)_{n\in\mathbb{N}}$ to be pointwise increasing (as opposed to essentially increasing). Then

$$
\begin{aligned}
\mathbf{E}\Big(\lim_{n\to\infty} X_n\Big) &\underbrace{\le}_{\text{Fatou}} \liminf_{n\to\infty} \mathbf{E}(X_n) && (1.2.44)\\
&\le \limsup_{n\to\infty} \mathbf{E}(X_n) && (1.2.45)\\
&\underbrace{\le}_{\text{Monotonicity}} \quad \mathbf{E}\Big(\lim_{n\to\infty} X_n\Big). && (1.2.46)
\end{aligned}
$$

Hence

$$\lim_{n\to\infty} \mathbf{E}(X_n) = \limsup_{n\to\infty} \mathbf{E}(X_n) = \liminf_{n\to\infty} \mathbf{E}(X_n) = \mathbf{E}\Big(\lim_{n\to\infty} X_n\Big) \tag{1.2.47}$$

as desired. □

Corollary 1.2.20. If $a, b \ge 0$ and X, Y are non-negative random variables, then

$$\mathbf{E}(aX + bY) = a\,\mathbf{E}(X) + b\,\mathbf{E}(Y). \tag{1.2.48}$$

Proof. Assume without loss of generality (via scaling) that $a = b = 1$.

The proof idea is to quantize X and Y and use the definition of E on simple functions. Take our favorite sequence of special functions $X_n = \min\left\{\frac{\lfloor 2^n X\rfloor}{2^n}, n\right\}$ which converges pointwise to X and is increasing, and furthermore by Theorem 1.2.19,

$$\lim_{n\to\infty} \mathbf{E}(X_n) = \mathbf{E}(X). \tag{1.2.49}$$

Similarly $Y_n = \min\left\{\frac{\lfloor 2^n Y\rfloor}{2^n}, n\right\}$. Then $X_n + Y_n$ converges pointwise to $X + Y$ and is monotone increasing, so by Theorem 1.2.19 we have

$$\mathbf{E}(X) + \mathbf{E}(Y) = \lim_{n\to\infty} (\mathbf{E}(X_n) + \mathbf{E}(Y_n)) = \lim_{n\to\infty} \mathbf{E}(X_n + Y_n) = \mathbf{E}(X + Y). \tag{1.2.50}$$

□

Expectation of Signed Random Variables

If X is a random variable, decompose $X = X^+ - X^-$ where $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$.

Definition 1.2.21 (Integrable Function). An (sometimes $\mathbf{P}$-)integrable function X is one that has $\mathbf{E}(|X|) = \mathbf{E}(X^+) + \mathbf{E}(X^-) < \infty$.

Now we define $\mathbf{E}$: integrable random variables $\to \overline{\mathbb{R}}$ by

$$\mathbf{E}(X) = \mathbf{E}\big(X^+\big) - \mathbf{E}\big(X^-\big). \tag{1.2.51}$$

Remark 1.2.22 (Proposition 1.15 in Textbook). $\mathbf{E}(X)$ is well-defined regardless of decomposition. That is, if $X = X_1 - X_2$ with $X_1, X_2 \geq 0$ and at least one is integrable, then

$$\mathbf{E}(X) = \mathbf{E}(X_1) + \mathbf{E}(X_2). \tag{1.2.52}$$

Theorem 1.2.23 (Properties of Expectation). Let X, Y be $\mathbf{P}$-integrable random variables.

(a) (Uniqueness.) If $X = Y$ $\mathbf{P}$-a.s., then $\mathbf{E}(X) = \mathbf{E}(Y)$.

(b) (Monotonicity.) If $X \leq Y$ $\mathbf{P}$-a.s., then $\mathbf{E}(X) \leq \mathbf{E}(Y)$.

(c) (Linearity.) If $a, b \in \mathbb{R}$, then $\mathbf{E}(aX + bY) = a\,\mathbf{E}(X) + b\,\mathbf{E}(Y)$.

(d) (Cauchy-Schwarz.) If X and Y are square-integrable, $\mathbf{E}(XY)^2 \leq \mathbf{E}\big(X^2\big)\,\mathbf{E}\big(Y^2\big)$.

(e) (Jensen's Inequality.) If $\phi\colon \mathbb{R} \to \mathbb{R}$ is convex and $\phi \circ X$ is integrable, then $\phi(\mathbf{E}(X)) \leq \mathbf{E}(\phi \circ X)$.

Proof. We will only prove (a), (b), and (c) – the proofs for (d) and (e) are found online.

Proof of (a). Let $E = \{X \neq Y\}$. Write

$$X^+ = X^+\left(1_E + 1_{E^c}\right) \underset{\text{a.s.}}{=} X^+ 1_{E^c} = Y^+ 1_{E^c} \underset{\text{a.s.}}{=} Y^+. \tag{1.2.53}$$

Doing the same thing for X^- and Y^-, we get

$$\begin{aligned} \mathbf{E}(X) &= \mathbf{E}\big(X^+\big) - \mathbf{E}\big(X^-\big) && (1.2.54)\\ &= \mathbf{E}\big(Y^+\big) - \mathbf{E}\big(Y^-\big) && (1.2.55)\\ &= \mathbf{E}(Y). && (1.2.56) \end{aligned}$$

Proof of (b). By the same argument, we can assume without loss of generality that $X \leq Y$ pointwise. Hence

$$Y = \underbrace{X^+ + (Y - X)}_{\geq 0} - \underbrace{X^-}_{\geq 0} \tag{1.2.57}$$

and hence

$$\begin{aligned} \mathbf{E}(X) &= \mathbf{E}\big(X^+\big) - \mathbf{E}\big(X^-\big) && (1.2.58)\\ &\leq \mathbf{E}\big(X^+ + (Y - X)\big) - \mathbf{E}\big(X^-\big) && (1.2.59)\\ &= \mathbf{E}(Y) && (1.2.60) \end{aligned}$$

where the last line uses the earlier remark.

Proof of (c). Assume again $a = b = 1$ and write $X = X^+ - X^-$, $Y = Y^+ - Y^-$. Then

$$\begin{aligned}\mathbf{E}(X) + \mathbf{E}(Y) &= \mathbf{E}\big(X^+\big) + \mathbf{E}\big(Y^+\big) - \big(\mathbf{E}\big(X^-\big) + \mathbf{E}\big(Y^-\big)\big) && (1.2.61)\\ &= \mathbf{E}\big(X^+ + Y^+\big) - \mathbf{E}\big(X^- + Y^-\big) && (1.2.62)\\ &= \mathbf{E}\big(X^+ + Y^+ - X^- - Y^-\big) && (1.2.63)\\ &= \mathbf{E}(X + Y) && (1.2.64)\end{aligned}$$

where the penultimate line uses the earlier remark.

□

Theorem 1.2.24 (Dominated Convergence Theorem). Let $(X_n)_{n\geq 1}$ be random variables with $X = \lim_{n\to\infty} X_n$ **P**-a.s.. If there is a **P**-integrable random variable Y such that $|X_n| \leq Y$ **P**-a.s. for all $n \in \mathbb{N}$, then

$$\mathbf{E}(X) = \lim_{n\to\infty} \mathbf{E}(X_n). \tag{1.2.65}$$

Proof. Note $Y + X_n \geq 0$ **P**-a.s.. By Lemma 1.2.17,

$$\begin{aligned}\mathbf{E}(Y + X) &\leq \liminf_{n\to\infty} \mathbf{E}(Y + X_n) && (1.2.66)\\ &= \liminf_{n\to\infty} (\mathbf{E}(Y) + \mathbf{E}(X_n)) && (1.2.67)\\ &= \mathbf{E}(Y) + \liminf_{n\to\infty} \mathbf{E}(X_n) && (1.2.68)\\ &\leq \mathbf{E}(Y) + \mathbf{E}(X). && (1.2.69)\end{aligned}$$

This implies

$$\liminf_{n\to\infty} \mathbf{E}(X_n) \geq \mathbf{E}(X). \tag{1.2.70}$$

Now we will do the same thing with $Y - X_n$, which is ≥ 0 **P**-a.s.. Then this pair of conclusions yield

$$\begin{aligned}&\implies \liminf_{n\to\infty} (-\mathbf{E}(X_n)) \geq -\mathbf{E}(X) && (1.2.71)\\ &\implies \limsup_{n\to\infty} \mathbf{E}(X_n) \leq \mathbf{E}(X) && (1.2.72)\end{aligned}$$

so we are done. □

Example 1.2.25. Let X and Y be independent random variables. If

1. $X, Y \geq 0$; or
2. both X and Y are **P**-integrable,

then

$$\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y). \tag{1.2.73}$$

Proof.

Proof of (1). Let $X_n = \min\left\{\frac{\lfloor 2^n X\rfloor}{2^n}, n\right\}$ and similarly to Y. Then they are simple functions which converge monotonically to X, and similarly define the same for Y. By monotone convergence theorem $\lim_{n\to\infty} \mathbf{E}(X_n Y_n) = \mathbf{E}(XY)$. But by independence $\mathbf{E}(X_n Y_n) = \mathbf{E}(X_n)\,\mathbf{E}(Y_n)$, and by monotone convergence theorem $\lim_{n\to\infty} \mathbf{E}(X_n) = \mathbf{E}(X)$ and $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$, so $\lim_{n\to\infty} \mathbf{E}(X_n)\,\mathbf{E}(Y_n) = \mathbf{E}(X)\,\mathbf{E}(Y)$. Hence $\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$.

Proof of (2). By (a),

$$\mathbf{E}(|XY|) = \mathbf{E}(|X|)\,\mathbf{E}(|Y|) < \infty \tag{1.2.74}$$

so that $|XY|$ is **P**-integrable. For X^+ define $((X^+))_{n\in\mathbb{N}}$ as a sequence of simple random variables that converge monotonically to X^+, and similarly for X^-, Y^+, and Y^-. Then $(X^+)_n - (X^-)_n$ converges **P**-a.s. to X and $|(X^+)_n - (X^-)_n| \leq |X|$. By Theorem 1.2.24 $\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$.

□

1.2.6 Important Examples

Theorem 1.2.26 (Markov's Inequality). If X is a random variable and $\lambda \geq 0$, then

$$\mathbf{P}(X \geq \lambda) \leq \frac{\mathbf{E}(X)}{\lambda}. \tag{1.2.75}$$

Proof. The simple calculation:

$$\begin{aligned} X &\geq \lambda 1_{\{X \geq \lambda\}} && (1.2.76)\\ &\geq 0 && (1.2.77)\\ \implies \mathbf{E}(X) &\geq \mathbf{E}\big(\lambda 1_{\{X \geq \lambda\}}\big) && (1.2.78)\\ &= \lambda \mathbf{E}\big(1_{\{X\geq\lambda\}}\big) && (1.2.79)\\ &= \lambda \mathbf{P}(X \geq \lambda). && (1.2.80) \end{aligned}$$

□

Theorem 1.2.27. If $1 \leq p \leq q < +\infty$, then

$$\mathbf{E}(|X|^p)^{1/p} \leq \mathbf{E}(|X|^q)^{1/q}. \tag{1.2.81}$$

Proof. Take $Y = |X|$ and $Y_n = \min\{Y, n\}$. Then by Jensen's inequality applied to the function $X \mapsto |X|^{q/p}$,

$$\begin{aligned} \mathbf{E}(|Y_n|^p) &= \mathbf{E}\Big((Y_n^p)^{1/p}\Big) && (1.2.82)\\ &\geq \mathbf{E}(Y^p)^{1/p} && (1.2.83)\\ \implies \mathbf{E}(Y^q)^{1/q} \overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbf{E}(Y_n^q)^{1/q} &\geq \lim_{n\to\infty} \mathbf{E}(Y_n^p)^{1/p} \overset{\text{MCT}}{=} \mathbf{E}(Y^p)^{1/p}. && (1.2.84) \end{aligned}$$

□

1.2.7 Functions of Random Variables

Theorem 1.2.28. If X is a random variable $(\Omega, \mathcal{F}, \mathbf{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and g is a measurable function $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, then $g \circ X$ is a random variable.

Corollary 1.2.29 (Law of Unconscious Statistician). If X is a random variable $(\Omega, \mathcal{F}, \mathbf{P}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and g is a measurable function $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, then

$$\mathbf{E}_{(\Omega,\mathcal{F},\mathbf{P})}(g \circ X) = \mathbf{E}_{(\mathbb{R},\mathcal{B}_{\mathbb{R}},L_X)}(g). \tag{1.2.85}$$

1.3 Conditional Probability and Conditional Expectation

1.3.1 Conditional Probability

Definition 1.3.1 (Conditional Probability). For events $A, B \in \mathcal{F}$ with $\mathbf{P}(B) > 0$, the **conditional probability** of A given B is defined as

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}. \tag{1.3.1}$$

Remark 1.3.2. The measure $\mathbf{P}(\cdot \mid B)$ is a probability measure on the measurable space $(B, \{F \cap B \colon F \in \mathcal{F}\})$.

Theorem 1.3.3 (Bayes' Rule). If $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$, then

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

Proof. We can write

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B \mid A) \tag{1.3.2}$$

$$= \mathbf{P}(B)\,\mathbf{P}(A \mid B). \tag{1.3.3}$$

□

1.3.2 Conditional Expectation

The following proposition follows directly from undergraduate notions of expectation:

$$\mathbf{E}(\mathbf{E}(X \mid Y)g(Y)) = \mathbf{E}(Xg(Y)) \quad \text{for all bounded measurable functions } g. \tag{1.3.4}$$

This is sometimes referred to as the “tower property”. This turns out to be the correct way to abstractly define the conditional expectation.

Definition 1.3.4 (Conditional Expectation). A measurable function h is a **conditional expectation** of X given Y if

$$\mathbf{E}(h(Y)g(Y)) = \mathbf{E}(Xg(Y)) \quad \text{for all bounded measurable functions } g. \tag{1.3.5}$$

Theorem 1.3.5. Let X be an integrable random variable and Y a random vector. The conditional expectation of X given Y exists, is integrable, and is $\mathbf{P}$-a.s. unique.

Notation 1.3.6. We denote it by $\mathbf{E}(X \mid Y)$. By definition it is a function of Y, but it is also a random variable and can be viewed as a function of ω.

Proof of Theorem 1.3.5. For proof of existence, see textbook.

For proof of integrability, note that $\operatorname{sign}(\cdot)$ is bounded and measurable, so

$$\mathbf{E}(|h(Y)|) = \mathbf{E}(h(Y)\operatorname{sign}(h(Y))) \tag{1.3.6}$$

$$= \mathbf{E}(X\operatorname{sign}(h(Y))) \tag{1.3.7}$$

$$\leq \mathbf{E}(|X|) \tag{1.3.8}$$

$$< \infty. \tag{1.3.9}$$

For proof of uniqueness, start by supposing h_1, h_2 are conditional expectations of X given Y. Then

$$\mathbf{E}(h_1(Y)g(Y)) = \mathbf{E}(Xg(Y)) = \mathbf{E}(h_2(Y)g(Y)) \tag{1.3.10}$$

for all bounded measurable functions g. Since h_1 and h_2 are integrable and g is bounded, their products are integrable, so we may use linearity of the integral to get

$$\mathbf{E}((h_1(Y) - h_2(Y))\,g(Y)) = 0. \tag{1.3.11}$$

Taking $g = \operatorname{sign}(\cdot) \circ (h_1 - h_2)$,

$$\mathbf{E}(|h_1(Y) - h_2(Y)|) = 0. \tag{1.3.12}$$

This implies $h_1 \circ Y = h_2 \circ Y$ $\mathbf{P}$-a.s.. □

Corollary 1.3.7. If X is integrable, then

$$\mathbf{E}(\mathbf{E}(X \mid Y, Z) \mid Y) = \mathbf{E}(X \mid Y). \tag{1.3.13}$$

Proof. Since $\mathbf{E}(X \mid Y, Z)$ is integrable, the conditional expectation $\mathbf{E}(\mathbf{E}(X \mid Y, Z) \mid Y)$ exists and satisfies the definition of conditional expectation:

$$\mathbf{E}(\mathbf{E}(\mathbf{E}(X \mid Z, Y) \mid Y)g(Y)) = \mathbf{E}(\mathbf{E}(X \mid Y, Z)g(Y)) \tag{1.3.14}$$

$$= \mathbf{E}(Xg(Y)) \tag{1.3.15}$$
$$= \mathbf{E}(\mathbf{E}(X \mid Y)g(Y)) \tag{1.3.16}$$

for all bounded measurable functions g. □

Theorem 1.3.8 (Properties of Conditional Expectation). If X_1, X_2 are integrable, then

(a) $\mathbf{E}(aX_1 + bX_2 \mid Y) = a\,\mathbf{E}(X_1 \mid Y) + b\,\mathbf{E}(X_2 \mid Y)$ $\mathbf{P}$-a.s..

(b) If $X_1 \geq X_2$, then $\mathbf{E}(X_1 \mid Y) \geq \mathbf{E}(X_2 \mid Y)$ $\mathbf{P}$-a.s..

(c) (Cauchy-Schwarz.) $\mathbf{E}(X_1X_2 \mid Y)^2 \leq \mathbf{E}\big(X_1^2 \bigm| Y\big)\,\mathbf{E}\big(X_2^2 \bigm| Y\big)$ $\mathbf{P}$-a.s..

(d) (Jensen.) If $\phi\colon \mathbb{R} \to \mathbb{R}$ is convex and $\phi \circ X$ is integrable, then $\phi(\mathbf{E}(X \mid Y)) \leq \mathbf{E}(\phi(X) \mid Y)$ $\mathbf{P}$-a.s..

Proof. We will only give a proof for (a); the rest go similarly. Let g be bounded and measurable. Then

$$\mathbf{E}(\mathbf{E}(aX_1 + bX_2 \mid Y)g(Y)) = \mathbf{E}((aX_1 + bX_2)\,g(Y)) \tag{1.3.17}$$
$$= a\,\mathbf{E}(X_1g(Y)) + b\,\mathbf{E}(X_2g(Y)) \tag{1.3.18}$$
$$= a\,\mathbf{E}(\mathbf{E}(X_1 \mid Y)g(Y)) + b\,\mathbf{E}(\mathbf{E}(X_2 \mid Y)g(Y)) \tag{1.3.19}$$
$$= \mathbf{E}((a\,\mathbf{E}(X_1 \mid Y) + b\,\mathbf{E}(X_2 \mid Y))\,g(Y)). \tag{1.3.20}$$

This is true for all such g so we are done. □

Remark 1.3.9. Versions of Lemma 1.2.17, Theorem 1.2.24, and Theorem 1.2.19 also hold for conditional expectation. For example, if $X = \lim_{n\to\infty} X_n$ where the limit is taken $\mathbf{P}$-a.s., and there exists integrable Z such that $|X_n| \leq Z$ for all n, then

$$\mathbf{E}(X \mid Y) = \lim_{n\to\infty} \mathbf{E}(X_n \mid Y) \quad \mathbf{P}\text{-a.s.} \tag{1.3.21}$$

Theorem 1.3.10 (Minimum Mean Square Error Property of Conditional Expectation). For any measurable f, and X with finite second moment,

$$\mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) \leq \mathbf{E}\Big((X - f(Y))^2\Big). \tag{1.3.22}$$

Proof. If $\mathbf{E}\Big(f(Y)^2\Big) = \infty$, then the right hand side is $+\infty$ and the inequality is obvious.

Now assume that $\mathbf{E}\Big(f(Y)^2\Big) < \infty$. Then

$$\mathbf{E}\Big((X - f(Y))^2\Big) = \mathbf{E}\Big((X - \mathbf{E}(X \mid Y) - (f(Y) - \mathbf{E}(X \mid Y)))^2\Big) \tag{1.3.23}$$
$$= \mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) + \mathbf{E}\Big((f(Y) - \mathbf{E}(X \mid Y))^2\Big) \tag{1.3.24}$$
$$- 2\,\mathbf{E}((X - \mathbf{E}(X \mid Y))\,(f(Y) - \mathbf{E}(X \mid Y))). \tag{1.3.25}$$

Integrability can be shown in the following calculation:

$$\mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) \leq 2\,\mathbf{E}\Big(X^2 + \mathbf{E}(X \mid Y)^2\Big) \tag{1.3.26}$$
$$= 2\,\mathbf{E}\big(X^2\big) + 2\,\mathbf{E}\Big(\mathbf{E}(X \mid Y)^2\Big) \tag{1.3.27}$$
$$\leq 2\,\mathbf{E}\big(X^2\big) + 2\,\mathbf{E}\big(X^2\big) \tag{1.3.28}$$
$$< \infty. \tag{1.3.29}$$

Similarly

$$\mathbf{E}\Big((f(Y) - \mathbf{E}(X \mid Y))^2\Big) < \infty. \tag{1.3.30}$$

So by Cauchy-Schwarz all terms are integrable. Continuing on the earlier calculations,

$$\begin{align}
\mathbf{E}\Big((X - f(Y))^2\Big) &= \mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) + \mathbf{E}\Big((f(Y) - \mathbf{E}(X \mid Y))^2\Big) \tag{1.3.31}\\
&\quad - 2\,\mathbf{E}((X - \mathbf{E}(X \mid Y))\,(f(Y) - \mathbf{E}(X \mid Y))) \tag{1.3.32}\\
&\geq \mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) \tag{1.3.33}\\
&\quad - 2\,\mathbf{E}\left((X - \mathbf{E}(X \mid Y))\,\underbrace{(f(Y) - \mathbf{E}(X \mid Y))}_{=g(Y)}\right) \tag{1.3.34}\\
&= \mathbf{E}\Big((X - \mathbf{E}(X \mid Y))^2\Big) - 2\,\mathbf{E}((X - \mathbf{E}(X \mid Y))\,g(Y)). \tag{1.3.35}\\
& \tag{1.3.36}
\end{align}$$

Now take $g_n = g1_{\{|g| \leq n\}}$. Then

$$|(X - \mathbf{E}(X \mid Y))\,g_n(Y)| \leq \underbrace{|(X - \mathbf{E}(X \mid Y))\,g(Y)|}_{\text{integrable}}. \tag{1.3.37}$$

Also

$$(X - \mathbf{E}(X \mid Y))\,g_n(Y) \to (X - \mathbf{E}(X \mid Y))\,g(Y) \quad \text{pointwise } \mathbf{P}\text{-a.s.} \tag{1.3.38}$$

By Theorem 1.2.24,

$$\mathbf{E}((X - \mathbf{E}(X \mid Y))\,g(Y)) = \lim_{n\to\infty} \underbrace{\mathbf{E}((X - \mathbf{E}(X \mid Y))\,g_n(Y))}_{=0} = 0. \tag{1.3.39}$$

This is because $g_n(Y)$ is bounded and measurable so the equality follows by the definition of conditional expectation. □

1.4 Stochastic Processes

Definition 1.4.1 (Stochastic Process). A **stochastic process** is a collection of random variables $(X_t)_{t\in\tau}$ on a common probability space.

Remark 1.4.2. Usually $\tau = \mathbb{Z}$ or $\tau = \mathbb{N}$ or $\tau = \mathbb{R}$ or $\tau = [0, \infty)$.

Definition 1.4.3. A **sample path** is a collection $(X_t\,(\omega))_{t\in\tau}$ for $\omega \in \Omega$.

In practice, we define a stochastic process by its distributions, in particular a collection of *consistent* finite-dimensional distributions. A theorem called *Kolmogorov's extension theorem* tells us that there is a probability space and random variables on this space with these finite-dimensional distributions. In this sense *consistency* means that if $\mathbf{P}_n$ is a Borel probability measure on $\mathbb{R}^n$, then

$$\mathbf{P}_{n+1}(B_n \times \mathbb{R}) = \mathbf{P}_n(B_n) \quad \text{for all } B_n \in \mathcal{B}_{\mathbb{R}^n}. \tag{1.4.1}$$

2 Limit Theorems and Modes of Convergence

2.1 Asymptotic Convergence of the Empirical Mean

2.1.1 Weak Law of Large Numbers

Theorem 2.1.1 (Weak Law of Large Numbers). Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. with $\mathbf{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. For any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{1}{n}\sum_{i\in[n]} X_i - \mu\right| > \epsilon\right) = 0. \tag{2.1.1}$$

Proof. Let $S_n = \sum_{i\in[n]} X_n$. Then

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) &= \mathbf{P}\left(\left(\frac{1}{n}S_n - \mu\right)^2 > \epsilon^2\right) && (2.1.2)\\
&\le \frac{\mathbf{E}\left(\left(\frac{1}{n}\sum_{i\in[n]}(X_i-\mu)\right)^2\right)}{\epsilon^2} && (2.1.3)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\sum_{j\in[n]} \underbrace{\mathbf{E}((X_i-\mu)(X_j-\mu))}_{\text{integrable by Cauchy-Schwarz}} && (2.1.4)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\sum_{\substack{j\in[n]\\ j\ne i}} \mathbf{E}((X_i-\mu)(X_j-\mu)) + \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\mathbf{E}\left((X_i-\mu)^2\right) && (2.1.5)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\sum_{\substack{j\in[n]\\ j\ne i}} \mathbf{E}(X_i-\mu)\,\mathbf{E}(X_j-\mu) + \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\mathbf{E}\left((X_i-\mu)^2\right) && (2.1.6)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\mathbf{E}\left((X_i-\mu)^2\right) && (2.1.7)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\mathrm{Var}(X_i) && (2.1.8)\\
&= \frac{1}{n^2\epsilon^2}\sum_{i\in[n]}\sigma^2 && (2.1.9)\\
&= \frac{\sigma^2}{n\epsilon^2} && (2.1.10)
\end{aligned}$$

which tends to 0 as $n \to \infty$. □

Remark 2.1.2. Actually, we only require $\mathbf{E}(|X_1|) < \infty$. The proof strategy is to truncate the random variables and take the truncation bound to ∞; the result follows from monotone convergence.

Similarly, we can relax $(X_n)_{n\in\mathbb{N}}$ being independent to being uncorrelated.

2.1.2 Strong Law of Large Numbers

Theorem 2.1.3 (Strong Law of Large Numbers). Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. with $\mathbf{E}(X_1) = \mu$ and $\mathbf{E}\left((X_1-\mu)^4\right) = \sigma_4 < \infty$. Then

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i\in[n]} X_i = \mu\right) = 1. \tag{2.1.11}$$

Proof. Again, let $S_n = \sum_{i\in[n]} X_n$. Then

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) &= \mathbf{P}\left(\left(\frac{1}{n}S_n - \mu\right)^4 > \epsilon^4\right) && (2.1.12)\\
&\le \frac{\mathbf{E}\left(\left(\frac{1}{n}S_n - \mu\right)^4\right)}{\epsilon^4} && (2.1.13)\\
&= \frac{\mathbf{E}\left(\left(\frac{1}{n}\sum_{i\in[n]}(X_i-\mu)\right)^4\right)}{\epsilon^4} && (2.1.14)\\
&= \frac{1}{n^4\epsilon^2}\sum_{i\in[n]}\sum_{j\in[n]}\sum_{k\in[n]}\sum_{\ell\in[n]} \mathbf{E}((X_i-\mu)(X_j-\mu)(X_k-\mu)(X_\ell-\mu)). && (2.1.15)
\end{aligned}$$

Since X_n's are i.i.d., the summand is only nonzero when $i=j=k=\ell$ (n such terms), or $i=j$ and $k=\ell$, or $i=k$ and $j=\ell$, or $i=\ell$ and $j=k$ ($\binom{n}{2}\binom{4}{2}$ such terms). Hence

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) &\le \frac{1}{n^4\epsilon^2}\sum_{i\in[n]}\sum_{j\in[n]}\sum_{k\in[n]}\sum_{\ell\in[n]} \mathbf{E}((X_i-\mu)(X_j-\mu)(X_k-\mu)(X_\ell-\mu)) && (2.1.16)\\
&= \frac{1}{n^4\epsilon^4}\left(n\sigma_4 + \binom{n}{2}\binom{4}{2}\sigma_4\right) && (2.1.17)\\
&= O\left(\frac{\sigma_4}{n^2\epsilon^4}\right). && (2.1.18)
\end{aligned}$$

By Borel-Cantelli,

$$\mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon \text{ i.o.}\right) = 0. \tag{2.1.19}$$

Now note that

$$\left\{\lim_{n\to\infty}\frac{1}{n}S_n \neq \mu\right\} = \bigcup_{k\in\mathbb{N}}\left\{\left|\frac{1}{n}S_n - \mu\right| > 2^{-k} \text{ i.o.}\right\}. \tag{2.1.20}$$

By continuity from below,

$$\begin{aligned}
\mathbf{P}\left(\lim_{n\to\infty}\frac{1}{n}S_n \neq \mu\right) &= \mathbf{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\left|\frac{1}{n}S_n - \mu\right| > 2^{-k} \text{ i.o.}\right\}\right) && (2.1.21)\\
&= \lim_{k\to\infty}\mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > 2^{-k} \text{ i.o.}\right) && (2.1.22)\\
&= 0. && (2.1.23)
\end{aligned}$$

□

Remark 2.1.4. Actually, we only require $\mathbf{E}(|X_1|) < \infty$. It is more complicated to prove this than our truncation method.

Similarly, we can relax $(X_n)_{n\in\mathbb{N}}$ being independent to having every collection of 4 random variables being uncorrelated.

This connects our axiomatic probability to the notion of frequentist probability. Say $(X_n)_{n\in\mathbb{N}}$ model outcomes of experiments performed under identical conditions (so are i.i.d., say with a random variable X).

The empirical CDF of X is

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i\in[n]} 1_{\{X_i\le x\}} && (2.1.24)\\ &\to F_X(x) \quad L_X\text{-a.s..} && (2.1.25)\end{aligned}$$

Example 2.1.5. Let $f\colon [0,1]\to\mathbb{R}$ be integrable and measurable. Let $(U_n)_{n\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$. Then by Theorem 2.1.3,

$$\mathbf{P}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i\in[n]} f(U_i) = \int_0^1 f(u)\,\mathrm{d}u\right) = 1. \tag{2.1.26}$$

This is called **Monte-Carlo integration**.

2.1.3 Central Limit Theorem

Theorem 2.1.6 (Central Limit Theorem). Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. random variables with covariance $\operatorname{Var}(X_n) = \sigma^2 < \infty$ and mean $\mathbf{E}(X_n) = \mu < \infty$, and define $S_n = \sum_{i\in[n]} X_i$. Then for every real x,

$$\lim_{n\to\infty}\mathbf{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x) \quad \text{for all } x\in\mathbb{R} \tag{2.1.27}$$

where $\Phi(x)$ is the CDF of a standard normal distribution, $\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2}\,\mathrm{d}u$.

In applications, the classical CLT is sometimes not directly applicable. Fortunately, there are many variations on the classical CLT that strengthen it in various ways. For example, the "identically distributed" assumption can be suitably relaxed.

Theorem 2.1.7 (Lindeberg's Central Limit Theorem). Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent but not necessarily identically distributed random variables. Assume the expected values $\mathbf{E}(X_n) = \mu_n$ and variances $\operatorname{Var}(X_n) = \sigma_n^2$ exist and are finite, and let $s_n^2 = \sum_{i\in[n]}\sigma_i^2$. If

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{i\in[n]}\mathbf{E}\Big((X_n-\mu_n)^2\cdot 1_{\{|X_i-\mu_i|>\epsilon s_n\}}\Big) = 0 \tag{2.1.28}$$

for all $\epsilon > 0$, then

$$\lim_{n\to\infty}\mathbf{P}\left(\frac{1}{s_n}\sum_{i\in[n]}(X_n-\mu_n)\le x\right) = \Phi(x) \quad \text{for all } x\in\mathbb{R} \tag{2.1.29}$$

As an example, the Berry-Esseen theorem gives quantitative rates of convergence:

Theorem 2.1.8 (Berry-Esseen Theorem). Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. with zero mean, variance σ^2, and finite third moment. Letting $F_n(x)$ denote the distribution function of the standardized sum $\frac{1}{\sigma\sqrt{n}}\sum_{i\in[n]} X_i$, we have

$$|F_n(x)-\Phi(x)| \le \frac{3\,\mathbf{E}\Big(|X_1|^3\Big)}{\sigma^3\sqrt{n}} \quad \text{for all } n\ge 1 \text{ and } x\in\mathbb{R}.$$

2.2 Convergence of Sequences of Random Variables

2.2.1 Convergence in Probability

Definition 2.2.1 (Convergence in Probability). Suppose $(X_n)_{n\in\mathbb{N}}$ and X are random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. We say $X_n \to X$ **in probability** (sometimes $\mathbf{P}$-i.p.) if

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \epsilon) = 0 \quad \text{for every } \epsilon > 0. \tag{2.2.1}$$

Example 2.2.2. In Theorem 2.1.1, $\frac{1}{n}S_n \to \mu$ $\mathbf{P}$-i.p..

2.2.2 Convergence Almost Surely

Definition 2.2.3 (Convergence Almost Surely). Suppose $(X_n)_{n\in\mathbb{N}}$ and X are random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. We say $X_n \to X$ **almost surely** (sometimes $\mathbf{P}$-a.s.) if

$$\mathbf{P}\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{2.2.2}$$

Example 2.2.4. In Theorem 2.1.3, $\frac{1}{n}S_n \to \mu$ $\mathbf{P}$-a.s..

2.2.3 Convergence in Distribution

Definition 2.2.5 (Convergence in Distribution). Suppose $(X_n)_{n\in\mathbb{N}}$ and X are random variables. We say $X_n \to X$ **in distribution** (sometimes i.d.) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{at all continuity points } x \text{ of } F_X. \tag{2.2.3}$$

Remark 2.2.6. Qualification of convergence at points of continuity is essential. Consider $X_n \sim \mathcal{U}\big([-\frac{1}{n}, \frac{1}{n}]\big)$. If $X = 0$, then intuitively $X_n \to X$ i.d., and

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0. \end{cases} \tag{2.2.4}$$

However,

$$F_X(0) = 1 \quad \text{but} \quad F_{X_n}(0) = \frac{1}{2} \quad \text{for every } n \in \mathbb{N}. \tag{2.2.5}$$

Hence $\lim_{n\to\infty} F_{X_n}(0) \neq F_X(0)$.

2.2.4 Convergence in L^p

Definition 2.2.7 (Convergence in L^p). Let $1 \leq p \leq \infty$ and let $(X_n)_{n\in\mathbb{N}}$ and X be integrable random variables on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say $X_n \to X$ in L^p (sometimes in p^{th} moment or in $L^p(\mathbf{P})$ or $L^p(\Omega, \mathcal{F}, \mathbf{P})$) if

$$\lim_{n\to\infty} \mathbf{E}(|X_n - X|^p) = 0. \tag{2.2.6}$$

We can define the same for $p = \infty$, in which case

$$\lim_{n\to\infty} \operatorname*{ess\,sup}_{\omega\in\Omega} |X_n(\omega) - X(\omega)| = 0. \tag{2.2.7}$$

2.2.5 Hierarchies of Convergence

Theorem 2.2.8. $X_n \to X$ **P**-a.s. $\implies$ $X_n \to X$ **P**-i.p. $\implies$ $X_n \to X$ i.d..

Remark 2.2.9. Even the "strongest" mode of convergence above (i.e. **P**-a.s. convergence) does not by itself imply things like $\lim_{n\to\infty} \mathbf{E}(X_n) = \mathbf{E}(X)$. We need to use Theorem 1.2.19 or Theorem 1.2.24.

Example 2.2.10 (Gliding Bump). Take $\Omega = [0,1]$ and $\mathbf{P} = \mathcal{U}([0,1])$. For $k \geq 1$ define

$$X_{k,n}(\omega) = \begin{cases} 1 & \frac{n-1}{k} \leq \omega < \frac{n}{k}, \quad n = 1, \ldots, k \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.8}$$

Then

$$\mathbf{P}(X_{k,n} \neq 0) = \frac{1}{k}. \tag{2.2.9}$$

Hence $\lim_{k\to\infty} X_{k,n} = 0$ **P**-i.p. for every n. Consider the sequence $X_{1,1}$, $X_{2,1}$, $X_{2,2}$, $X_{3,1}$, $X_{3,2}$, $X_{3,3}$, $X_{4,1}$, ... and let $(Y_n)_{n\in\mathbb{N}}$ be this sequence. Then

$$\limsup_{n\to\infty} Y_n(\omega) = 1 > 0 = \liminf_{n\to\infty} Y_n(\omega) \quad \text{for all } \omega \in \Omega. \tag{2.2.10}$$

However, $\lim_{k\to\infty} X_{k,n} = 0$ **P**-a.s. for every n.

Proposition 2.2.11. Let $(X_n)_{n\in\mathbb{N}}$, X be such that $X_n \to X$ **P**-i.p.. There exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ such that $X_{n_k} \to X$ **P**-a.s..

Proof. Because of convergence in probability, for any $k \geq 1$ there exists n_k such that

$$\mathbf{P}\left(|X_{n_k} - X| > \frac{1}{k}\right) \leq 2^{-k}. \tag{2.2.11}$$

By Borel-Cantelli,

$$\mathbf{P}\left(|X_{n_k} - X| > \frac{1}{k} \text{ i.o.}\right) = 0. \tag{2.2.12}$$

This means that $\lim_{k\to\infty} X_{n_k} = X$ **P**-a.s.. □

Proposition 2.2.12. If $1 \leq p \leq q \leq \infty$ and $X_n \to X$ in $L^q(\mathbf{P})$ then

(a) $X_n \to X$ in $L^p(\mathbf{P})$.

(b) $X_n \to X$ **P**-i.p..

(c) $\mathbf{E}(X_n) \to \mathbf{E}(X)$.

Proof.

Proof of (a). Monotonicity of moments.

Proof of (b). Markov's inequality.

Proof of (c). By the following calculation:

$$\lim_{n\to\infty} |\mathbf{E}(X_n) - \mathbf{E}(X)| = \lim_{n\to\infty} |\mathbf{E}(X_n - X)| \leq \lim_{n\to\infty} \mathbf{E}(|X_n - X|) \overset{(a)}{=} 0. \tag{2.2.13}$$

□

Example 2.2.13. Let $(X_n) \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}\big(\frac{1}{2}\big)$ and let $X \sim \mathrm{Bernoulli}\big(\frac{1}{2}\big)$. Then $X_n \to X$ in distribution, trivially. But X_n does not converge to anything in probability. To see why, note that for any random variable Z,

$$\begin{aligned} \mathbf{P}\left(|X_n - Z| > \frac{1}{4}\right) + \mathbf{P}\left(|X_m - Z| > \frac{1}{4}\right) &\geq \mathbf{P}\left(\left\{|X_n - Z| > \frac{1}{4}\right\} \cup \left\{|X_m - Z| > \frac{1}{4}\right\}\right) && (2.2.14) \\ &\geq \mathbf{P}\left(|X_n - X_m| > \frac{1}{2}\right) && (2.2.15) \\ &= \mathbf{P}(X_n \neq X_m) && (2.2.16) \\ &= \frac{1}{2} && (2.2.17) \end{aligned}$$

for all $n \neq m$. This implies that

$$\liminf_{n\to\infty} \mathbf{P}\left(|X_n - Z| > \frac{1}{4}\right) \geq \frac{1}{4}. \qquad (2.2.18)$$

2.3 More on Convergence in Distribution

Remark 2.3.1. The Central Limit Theorem Theorem 2.1.6 is really a statement about convergence in distribution, namely it says that

$$\frac{1}{\sqrt{n}} \sum_{i\in[n]} \frac{X_i - \mu}{\sigma} \to Z \sim \mathcal{N}(0)1 \quad \text{in distribution.} \qquad (2.3.1)$$

2.3.1 Blackboxed Tools

Theorem 2.3.2. Let $(X_n)_{n\in\mathbb{N}}$ and X be random variables. Then $X_n \to X$ in distribution if and only if

$$\mathbf{E}(g(X_n)) \to \mathbf{E}(g(X)) \quad \text{for all continuous bounded } g\colon \mathbb{R} \to \mathbb{R}. \qquad (2.3.2)$$

Definition 2.3.3 (Tightness). A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is "tight" if

$$\lim_{x\to\infty} \sup_{n\geq 1} \mathbf{P}(|X_n| > x) = 0. \qquad (2.3.3)$$

Theorem 2.3.4 (Consequence of Prokhorov's Theorem). If $(X_n)_{n\in\mathbb{N}}$ is a tight sequence of random variables, then there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ and a real-valued X such that $X_{n_k} \to X$ in distribution.

2.3.2 Characteristic Functions

Definition 2.3.5 (Characteristic Function). The characteristic function ϕ_X of a real-valued random variable X is defined as

$$\phi_X(t) := \mathbf{E}\big(\mathrm{e}^{\mathrm{i}tX}\big) = \mathbf{E}(\cos(tX)) + \mathrm{i}\,\mathbf{E}(\sin(tX)) \quad \text{for } t \in \mathbb{R}. \qquad (2.3.4)$$

Remark 2.3.6. Characteristic functions are the integral of a bounded function and thus always exist. Further, they are unique in the sense that two different distributions have two different characteristic functions.

Characteristic functions are useful for studying sums of independent random variables.

Proposition 2.3.7. If X and Y are independent random variables, then

$$\phi_{aX+bY}(t) = \phi_X(at)\phi_Y(bt). \qquad (2.3.5)$$

Proof. Writing out the computation,

$$\begin{aligned}\phi_{aX+bY}(t) &= \mathbf{E}\left(\mathrm{e}^{\mathrm{i}taX}\mathrm{e}^{\mathrm{i}tbY}\right) = \mathbf{E}\left(\mathrm{e}^{\mathrm{i}taX}\right)\mathbf{E}\left(\mathrm{e}^{\mathrm{i}tbY}\right) = \mathbf{E}\left(\mathrm{e}^{\mathrm{i}(at)X}\right)\mathbf{E}\left(\mathrm{e}^{\mathrm{i}(bt)Y}\right) \qquad (2.3.6)\\ &= \phi_X(at)\phi_Y(bt). \qquad (2.3.7)\end{aligned}$$

□

Theorem 2.3.8 (Lévy's Continuity Theorem). Let $(X_n)_{n\in\mathbb{N}}$ and X be real-valued random variables. Then $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}$ if and only if $X_n \to X$ in distribution.

Lemma 2.3.9. We have the following two important facts:

(a) The function $\phi_X(t)$ is uniformly continuous in t.

(b) For every $u > 0$,

$$\mathbf{P}\left(|X| \geq \frac{2}{u}\right) \leq \frac{1}{u}\int_{-u}^{u}(1-\phi_X(t))\,\mathrm{d}t. \tag{2.3.8}$$

Proof.

Proof of (a). Observe that

$$|\phi_X(t+h) - \phi_X(t)| \leq \mathbf{E}\left(\left|\mathrm{e}^{\mathrm{i}(t+h)X} - \mathrm{e}^{\mathrm{i}tX}\right|\right) = \mathbf{E}\left(\left|\mathrm{e}^{\mathrm{i}hX} - 1\right|\right) \tag{2.3.9}$$

and the right-hand side does not depend on t.

Proof of (b). Let $T \sim \mathcal{U}([-u,u])$ independently of X. Then

$$\begin{aligned}\frac{1}{2u}\int_{-u}^{u}(1-\phi_X(t))\,\mathrm{d}t &= \mathbf{E}(1-\phi_X(T)) = \mathbf{E}\left(1-\mathrm{e}^{\mathrm{i}TX}\right) &&(2.3.10)\\ &= \mathbf{E}\left(1-\mathbf{E}\left(\mathrm{e}^{\mathrm{i}TX} \mid X\right)\right) &&(2.3.11)\\ &= \mathbf{E}\left(1-\mathbf{E}(\cos(TX) \mid X) - \mathrm{i}\,\underbrace{\mathbf{E}(\sin(TX) \mid X)}_{=0}\right) &&(2.3.12)\\ &= \mathbf{E}\left(1-\frac{\sin(uX)}{uX}\right) &&(2.3.13)\\ &\geq \frac{1}{2}\mathbf{E}\left(1_{\left\{|X|\geq\frac{2}{u}\right\}}\right) &&(2.3.14)\\ &= \frac{1}{2}\mathbf{P}\left(|X| \geq \frac{2}{u}\right). &&(2.3.15)\end{aligned}$$

□

Proof of Theorem 2.3.8. Assume $X_n \to X$ in distribution. Then $\mathbf{E}(g(X_n)) \to \mathbf{E}(g(X))$ for all bounded continuous g. Take $g(x) = \mathrm{e}^{\mathrm{i}tX}$. This implies

$$\phi_{X_n}(t) = \mathbf{E}\left(\mathrm{e}^{\mathrm{i}tX_n}\right) \to \mathbf{E}\left(\mathrm{e}^{\mathrm{i}tX}\right) = \phi_X(t) \quad \text{for all } t \in \mathbb{R}. \tag{2.3.16}$$

This is what we want.

Now assume that $\phi_{X_n}(t) \to \phi_X(t)$ for each $t \in \mathbb{R}$. Fix $\epsilon > 0$. By continuity of ϕ_X there exists $u_\epsilon > 0$ such that

$$\frac{1}{2u_\epsilon}\int_{-u_\epsilon}^{u_\epsilon}(1-\phi_X(t))\,\mathrm{d}t < \epsilon. \tag{2.3.17}$$

Since $\phi_{X_n} \to \phi_X$ pointwise, Dominated Convergence Theorem implies

$$\lim_{n\to\infty} \frac{1}{2}\mathbf{P}\left(|X_n| > \frac{2}{u_\epsilon}\right) \le \lim_{n\to\infty} \frac{1}{2u_\epsilon}\int_{-u_\epsilon}^{u_\epsilon} (1-\phi_{X_n}(t))\,\mathrm{d}t = \frac{1}{2u_\epsilon}\int_{-u_\epsilon}^{u_\epsilon} (1-\phi_X(t))\,\mathrm{d}t < \epsilon. \tag{2.3.18}$$

Thus $(X_n)_{n\in\mathbb{N}}$ is tight. By Prokhorov's Theorem, there exists some random variable X_∞ and a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ such that $X_{n_k} \to X_\infty$ in distribution. This implies that $\phi_{X_{n_k}}(t) \to \phi_{X_\infty}(t)$ for every t. But we already knew that $\phi_{X_n}(t) \to \phi_X(t)$ for every t, so $\phi_{X_{n_k}}(t) \to \phi_X(t)$ for every t. Hence $\phi_X = \phi_{X_\infty}$ so $X_n = X$ in distribution. □

We want to invoke the following smoothing lemma:

Lemma 2.3.10. If $\mathbf{E}(|X|^n) < \infty$, then

$$\phi_X(t) = \sum_{k=0}^{n} \frac{(\mathrm{i}t)^k\,\mathbf{E}\left(X^k\right)}{k!} + o(|t|^n). \tag{2.3.19}$$

Proof of Central Limit Theorem. Without loss of generality, we assume $\mathbf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = 1$. Now define $U_n := \frac{1}{\sqrt{n}}\sum_{i\in[n]} X_i$. Then

$$\phi_{U_n}(t) = \prod_{i\in[n]} \phi\left(\frac{t}{\sqrt{n}}\right) = \phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)^n \tag{2.3.20}$$

$$= \left(1 - \frac{t^2}{2n} + o\left(\frac{|t|^2}{n}\right)\right)^n \tag{2.3.21}$$

Fixing t and taking n large,

$$\lim_{n\to\infty} \phi_{U_n}(t) = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n} + o\left(\frac{|t|^2}{n}\right)\right)^n \tag{2.3.22}$$

$$= \mathrm{e}^{-t^2/2} = \phi_Z(t) \tag{2.3.23}$$

for $Z \sim \mathcal{N}(0,1)$. □

3 Gaussians

3.1 Gaussian Random Variables

Definition 3.1.1 (Gaussian Random Variable). A **standard normal random variable** $W \sim \mathcal{N}(0,1)$ has characteristic function

$$\phi_W(t) := \mathrm{e}^{-t^2/2} \tag{3.1.1}$$

and probability density function

$$f_W(w) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-w^2/2}. \tag{3.1.2}$$

If $X = \sigma Z + \mu$ we write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ which has characteristic function

$$\phi_X(t) := \exp\left(-\mathrm{i}\mu t - \frac{\sigma^2 t^2}{2}\right) \tag{3.1.3}$$

and probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/2}. \tag{3.1.4}$$

3.2 Gaussian Random Vectors

Definition 3.2.1. A random vector $Z \in \mathbb{R}^k$ is a **Gaussian vector** if it can be written as an affine transformation of i.i.d. standard normal random variables:

$$Z = AW + \mu \tag{3.2.1}$$

where $W_1, \ldots, W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

If A is not full rank, Z does not admit a density on $\mathbb{R}^k$. To compute the density of Z, we thus assume A is full rank.

Fact 3.2.2 (Change of Variables for Densities). If $X = \Gamma Y + b$ and Γ is non-singular, then

$$f_X(x) = \frac{1}{|\det(\Gamma)|} f_Y\left(\Gamma^{-1}(x-b)\right). \tag{3.2.2}$$

Now, we know that

$$f_W(w) = \prod_{i \in [k]} f_{w_i}(w_i) = \frac{1}{(2\pi)^{k/2}} \mathrm{e}^{-w^* w/2} \tag{3.2.3}$$

$$\implies f_Z(z) = \frac{1}{\det(AA^*)^{1/2}} f_W\left(A^{-1}(z-\mu)\right) \tag{3.2.4}$$

$$= \frac{1}{(2\pi)^{k/2} \det(AA^*)^{1/2}} \exp\left(-\frac{(z-\mu)^* (AA^*)^{-1} (z-\mu)}{2}\right) \tag{3.2.5}$$

$$= \frac{1}{(2\pi)^{k/2} \det(\Sigma_Z)^{1/2}} \exp\left(-\frac{1}{2} (z-\mu)^* \Sigma_Z^{-1} (z-\mu)\right) \tag{3.2.6}$$

where

$$\begin{aligned}\Sigma_Z := \operatorname{Cov}(Z) = \mathbf{E}((Z - \mathbf{E}(Z))\,(Z - \mathbf{E}(Z))^*) = \mathbf{E}((AW)\,(AW)^*) \quad &(3.2.7)\\ = \mathbf{E}(AWW^*A^*) = A\,\mathbf{E}(WW^*)A^* = AA^*. \quad &(3.2.8)\end{aligned}$$

One notes that the jointly Gaussian random variables are parameterized entirely by the mean and covariance.

Example 3.2.3. Let (X, Y) be jointly Gaussian random variables. Then we have $\operatorname{Cov}(X, Y) = 0$ if and only if X and Y are independent.

Example 3.2.4. If X and Y are only *marginally* Gaussian, then $\operatorname{Cov}(X, Y) = 0$ does *not* imply X and Y are independent. For example, if $X \sim \mathcal{N}(0, 1)$ and $B \sim \text{Rademacher}\big(\frac{1}{2}\big)$, then $Y = BX \sim \mathcal{N}(0, 1)$. Then

$$\operatorname{Cov}(X, Y) = \mathbf{E}(XY) = \mathbf{E}\big(BX^2\big) = \mathbf{E}(B)\,\mathbf{E}\big(X^2\big) = 0$$

but $|Y| = |X|$.

3.3 Jointly Gaussian Random Vectors

Notation 3.3.1. We use the notations:

$$\begin{aligned}\mu_X &:= \mathbf{E}(X) &(3.3.1)\\ \Sigma_X &:= \mathbf{E}((X - \mu_X)\,(X - \mu_X)^*) &(3.3.2)\\ \Sigma_{XY} &:= \mathbf{E}((X - \mu_X)\,(Y - \mu_Y)^*). &(3.3.3)\end{aligned}$$

Let X and Y be jointly Gaussian random vectors. Then

$$\operatorname{Cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}. \tag{3.3.4}$$

Theorem 3.3.2. If X and Y are jointly Gaussian random vectors and $\Sigma_Y > 0$, then

$$X = \mu_X + \Sigma_{XY}\Sigma_Y^{-1}\,(Y - \mu_Y) + V \tag{3.3.5}$$

where $V \sim \mathcal{N}\big(0, \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\big)$, independent of Y.

Proof. The strategy is in two parts. We let Y and V be as given, and $\Sigma \geq 0$ partitioned as stated.

1. First, we show that X and Y are jointly Gaussian.
2. Then, we show that $\operatorname{Cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$ for the given matrices.

Lemma 1. If $\Sigma \geq 0$, then $Z \sim \mathcal{N}(0, \Sigma)$ exists as a random vector.

> *Proof.* Write $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$, let $Z = \Sigma^{1/2}W$ where $W \sim \mathcal{N}(0, I)$, hence Z is jointly Gaussian with $\Sigma_Z = \Sigma$. □

Lemma 2. $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} \geq 0$.

> *Proof.* This is equivalent to the statement that $\Sigma \geq 0$ by the Schur complement. But all covariance matrices are PSD. □

Now to prove 1, we realize that X is a linear combination of entries of Y, plus some other independently generated linear combination of standard normal random variables. Thus it's clear that $\begin{bmatrix} X \\ Y \end{bmatrix}$ is jointly Gaussian.

To prove 2, we assume without loss of generality that $\mu_X = \mu_Y = 0$ and compute

$$\begin{aligned}
\mathbf{E}(XY^*) &= \mathbf{E}\big(\big(\Sigma_{XY}\Sigma_Y^{-1}Y + V\big)\,Y^*\big) && (3.3.6)\\
&= \Sigma_{XY}\Sigma_Y^{-1}\,\mathbf{E}(YY^*) && (3.3.7)\\
&= \Sigma_{XY} && (3.3.8)\\
\mathbf{E}(XX^*) &= \mathbf{E}\big(\big(\Sigma_{XY}\Sigma_Y^{-1}Y + V\big)\,X^*\big) && (3.3.9)\\
&= \mathbf{E}\big(\Sigma_{XY}\Sigma_Y^{-1}YX^*\big) + \mathbf{E}\Big(V\left(\Sigma_{XY}\Sigma_Y^{-1}Y + V\right)^*\Big) && (3.3.10)\\
&= \Sigma_{XY}\Sigma_Y^{-1}\,\mathbf{E}(YX^*) + \mathbf{E}(V)\,\mathbf{E}\Big(\big(\Sigma_{XY}\Sigma_Y^{-1}Y\big)^*\Big) + \mathbf{E}(VV^*) && (3.3.11)\\
&= \Sigma_{XY}\Sigma_Y^{-1}\,\mathbf{E}(YX^*) + \mathbf{E}(VV^*) && (3.3.12)\\
&= \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} + \big(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\big) && (3.3.13)\\
&= \Sigma_X. && (3.3.14)
\end{aligned}$$

□

Corollary 3.3.3.

$$\mathbf{E}(X \mid Y) = \mu_X + \Sigma_{XY}\Sigma_Y^{-1}\left(Y - \mu_Y\right). \tag{3.3.15}$$

Note that $\mathbf{E}(X \mid Y)$ is the MMSE estimate of X given Y. It's really special that for jointly Gaussian (X, Y), this estimate is a *affine function* of Y.

3.4 Gaussian Processes

Definition 3.4.1. A random process $(X_t)_{t\in\mathbb{R}_{>0}}$ is a **Gaussian process** if for any finite collection $\{t_1, \ldots, t_k\}$, the samples $(X_{t_i})_{i\in[k]}$ are jointly Gaussian.

Example 3.4.2. If $(X_n)_{n\in\mathbb{N}}$ is a Gaussian process, then for each $n \geq 1$, there is a deterministic vector $a_n \in \mathbb{R}^n$ and a scalar b_n such that we can write

$$X_{n+1} = a_n^* \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} + b_n + V_n \tag{3.4.1}$$

where $V_n \sim \mathcal{N}\left(0, \sigma_n^2\right)$ independently of $(X_i)_{i\in[n]}$.

Here the term $a_n^* \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} + b_n$ is the **predictable part** and V_n is the **innovation**.

The most important stochastic process is known as **Brownian motion**, which is a good model for microscopic behavior.

Definition 3.4.3. A stochastic process $(X_t)_{t\in\mathbb{R}_{>)}}$ is a Brownian motion if

(1) It has independent stationary increments.

(2) It is a zero-mean Gaussian process with $X_0 = 0$.

(3) Its sample paths are **P**-a.s. continuous:

$$\mathbf{P}(\{\omega \colon t \mapsto X_t(\omega) \text{ is continuous}\}) = 1. \tag{3.4.2}$$

By *increments* we mean random variables of the form

$$X_{t_i} - X_{t_{i-1}} \quad t_{i-1} \in \mathbb{R}_{\geq 0} \leq t_i \in \mathbb{R}_{\geq 0} \quad . \tag{3.4.3}$$

By *independent increments* we mean that $\left(X_{t_{i+1}} - X_{t_i}\right)_{i\in[k]}$ for any choice of increasing times $(t_i)_{i\in[k+1]}$.

By *stationary increments* we mean that

$$X_{t+\tau} - X_t \overset{\text{i.d.}}{=} X_{t+s+\tau} - X_{t+s} \overset{\text{i.d.}}{=} X_\tau - X_0 = X_\tau \quad \text{for all } t, s, \tau \in \mathbb{R}_{\geq 0}. \tag{3.4.4}$$

Proposition 3.4.4. Brownian motion exists.

Proposition 3.4.5. Brownian motion $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is a zero-mean Gaussian process with **P**-a.s.. continuous sample paths and

$$\operatorname{Cov}(X_s, X_t) = \operatorname{Var}(X_1) \min\{s, t\} \quad \text{for all } s, t \in \mathbb{R}_{\geq 0}. \tag{3.4.5}$$

Proof. Take any $k \geq 1$ and write

$$\begin{aligned}
\operatorname{Var}(X_1) &= \operatorname{Var}\left(\sum_{i=1}^{k}\left(X_{1-\frac{i-1}{k}} - X_{1-\frac{i}{k}}\right)\right) && (3.4.6)\\
&= \sum_{i=1}^{k} \operatorname{Var}\left(X_{1-\frac{i-1}{k}} - X_{1-\frac{i}{k}}\right) && (3.4.7)\\
&= k \operatorname{Var}\left(X_{1/k}\right) && (3.4.8)\\
\implies \operatorname{Var}\left(X_{n/k}\right) &= n \operatorname{Var}\left(X_{1/k}\right) && (3.4.9)\\
&= \frac{n}{k} \operatorname{Var}(X_1). && (3.4.10)
\end{aligned}$$

So for any rational $q \geq 0$, $\operatorname{Var}(X_q) = q \operatorname{Var}(X_1)$. Let $t \notin \mathbb{Q}_{\geq 0}$, take rationals $q_n \to t$. Continuity of sample paths means that $X_{q_n} \to X_t$ almost surely, so $X_{q_n}^2 \to X_t$ almost surely. Therefore Fatou's Lemma tells us that

$$\operatorname{Var}(X_t) \leq \lim_{n\to\infty} \operatorname{Var}(X_{q_n}) = \lim_{n\to\infty} q_n \operatorname{Var}(X_1) = t \operatorname{Var}(X_1). \tag{3.4.11}$$

To go the other direction, take $q \leq t$. Then

$$\operatorname{Var}(X_t) = \operatorname{Var}(X_t - X_q) = \operatorname{Var}(X_t - X_q) + \operatorname{Var}(X_q) \geq \operatorname{Var}(X_q) = q \operatorname{Var}(X_1). \tag{3.4.12}$$

Take a sequence of rationals $q_n \to t$ and obtain

$$\operatorname{Var}(X_t) = t \operatorname{Var}(X_1). \tag{3.4.13}$$

□

4 Linear Estimation

4.1 Introduction

When we want to estimate a random variable X from the random variable Y which we observe, we need a *loss function* ℓ to quantify the "best estimate". In general, this resolves to solving the optimization problem

$$\inf_f \mathbf{E}(\ell(X, f(Y))). \tag{4.1.1}$$

This is completely intractable if we don't know the form of ℓ. Things get a little easier if we specify $\ell(x, y) = \|x - y\|_2^2$:

$$\inf_f \mathbf{E}(\ell(X, f(Y))) = \inf_f \mathbf{E}\left(\|X - f(Y)\|_2^2\right) = \mathbf{E}(X \mid Y). \tag{4.1.2}$$

However, it's really difficult to compute $\mathbf{E}(X \mid Y)$ unless we know the full joint distribution of X and Y.

In practice, we focus on trying to compute the best *linear* estimator. This is particularly meaningful for the case where X and Y are (approximately) jointly Gaussian, since $\mathbf{E}(X \mid Y)$ is a linear function of Y.

4.2 The Hilbert Space $L^2(\Omega, \mathcal{F}, \mathbf{P})$

For a fixed probability space $(\Omega, \mathcal{F}, \mathbf{P})$, we denote $L^2\left(\Omega, \mathcal{F}, \mathbf{P}\right)$ the set of random variables X such that $\|X\|_{L^2}^2 = \mathbf{E}\left(|X|^2\right) < \infty$, quotiented out by the set of random variables which are a.s. 0. We equip it with the inner product

$$\langle X, Y \rangle_{L^2} := \mathbf{E}(XY^*) \quad X, Y \in L^2(\Omega, \mathcal{F}, \mathbf{P}) \tag{4.2.1}$$

and Euclidean norm $\|X\|_{L^2} = \mathbf{E}\left(|X|^2\right)^{1/2}$. Sometimes we call it L^2 or $L^2(\mathbf{P})$.

Proposition 4.2.1. L^2 is a vector space.

Proof. If $X \in L^2$ then $\alpha X \in L^2$ for all $\alpha \in \mathbb{R}$. If $X, Y \in L^2$ then $X + Y \in L^2$ by the inequality $(a + b)^2 \leq 2a^2 + 2b^2$. □

Proposition 4.2.2. The function $\langle X, Y \rangle_{L^2} = \mathbf{E}(XY^*)$ is an inner product.

Proof. Follows by properties of expectation. □

Proposition 4.2.3. $\|X\|_{L^2} = \sqrt{\langle X, X \rangle_{L^2}} = \left(\mathbf{E}\left(|X|^2\right)\right)^{1/2}$ is a norm on L^2.

Definition 4.2.4. For a normed space $(\mathcal{X}, \|\cdot\|)$, a sequence $(x_n)_{n\in\mathbb{N}}$ is a **Cauchy sequence** if for all $\epsilon > 0$ there exists some N_ϵ such that

$$\|x_n - x_m\| < \epsilon \quad \text{for all } m, n \geq N_\epsilon. \tag{4.2.2}$$

Proposition 4.2.5. In the context of the previous definition, if $x_n \to x$ in $\mathcal{X}$, then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence.

Proof. Easy by the triangle inequality:

$$\|x_n - x_m\| \leq \underbrace{\|x_n - x\|}_{<\frac{\epsilon}{2}\ \forall n \geq N_{\epsilon/2}} + \underbrace{\|x_m - x\|}_{<\frac{\epsilon}{2}\ \forall n \geq N_{\epsilon/2}} \tag{4.2.3}$$

$$\leq \epsilon. \tag{4.2.4}$$

□

Definition 4.2.6. A normed space $(\mathcal{X}, \|\cdot\|)$ is **complete** if every Cauchy sequence is convergent.

Theorem 4.2.7. $L^2(\Omega, \mathcal{F}, \mathbf{P})$ is a Hilbert space.

Proof. We only need to show that L^2 is complete. The basic idea is to take a Cauchy sequence and show that there exists a limit which it converges to.

Let $(X_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in L^2. We can extract a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ such that

$$\mathbf{E}\left(\left|X_{n_k} - X_{n_{k+1}}\right|^2\right) < 2^{-3k}. \tag{4.2.5}$$

Indeed, we just need $n_k > N_{2^{-3k/2}}$ in the definition of a Cauchy sequence. Then Markov's inequality tells us that

$$\mathbf{P}\left(\left|X_{n_k} - X_{n_{k+1}}\right| > 2^{-k}\right) = \mathbf{P}\left(\left|X_{n_k} - X_{n_{k+1}}\right|^2 > 2^{-2k}\right) \tag{4.2.6}$$

$$\leq \frac{\mathbf{E}\left(\left|X_{n_k} - X_{n_{k+1}}\right|^2\right)}{2^{-2k}} \tag{4.2.7}$$

$$< 2^{-k}. \tag{4.2.8}$$

By Borel-Cantelli,

$$\mathbf{P}\left(\left|X_{n_k} - X_{n_{k+1}}\right| > 2^{-k} \text{ for infinitely many } k\right) = 0. \tag{4.2.9}$$

Then

$$\mathbf{P}\left(\sum_{k\in\mathbb{N}} \left|X_{n_k} - X_{n_{k+1}}\right|^2 < \infty\right) \geq \mathbf{P}\left(\left|X_{n_k} - X_{n_{k+1}}\right| < 2^{-k} \text{ for all but finitely many } k\right) \tag{4.2.10}$$

$$= 1 - \mathbf{P}\left(\left|X_{n_k} - X_{n_{k+1}}\right| > 2^{-k} \text{ for infinitely many } k\right) \tag{4.2.11}$$

$$= 1. \tag{4.2.12}$$

Hence $X = \lim_{k\to\infty} X_{n_k}$ exists almost surely.

By Fatou's lemma, for any $n \geq 1$,

$$\mathbf{E}\left(|X_n - X|^2\right) \leq \liminf_{k\to\infty} \mathbf{E}\left(|X_n - X_{n_k}|^2\right) < \epsilon \quad \text{for all } n \geq N_\epsilon. \tag{4.2.13}$$

This tells us that $X_n - X \in L^2$, and $X_n \in L^2$, so $X \in L^2$. Also since ϵ is arbitrary, $X_n \to X$ in L^2 (in the sense that $\lim_{n\to\infty} \|X_n - X\|_{L^2} = 0$). □

If $(X_n)_{n\in\mathbb{N}}$ is a sequence in L^2, we want to figure out what it means to write $\sum_{n\in\mathbb{N}} X_n$. We say $X_{n\in\mathbb{N}} X_n = Y$ in L^2 if for every rearrangement $\sigma\colon \mathbb{N} \to \mathbb{N}$,

$$\lim_{N\to\infty} \sum_{i\in[N]} X_{\sigma(i)} = Y \quad \text{in } L^2. \tag{4.2.14}$$

More strongly, we say $X_{n\in\mathbb{N}} \to X_n = Y$ **P**-a.s. and in L^2 if Eq. (4.2.14) holds **P**-a.s. and in L^2.

A sufficient condition for convergence a.s. and in L^2 is

$$\sum_{n\in\mathbb{N}} \|X_n\|_{L^2} < \infty. \tag{4.2.15}$$

Theorem 4.2.8 (Hilbert Projection Theorem). Let H be a Hilbert space, and C a closed convex set. Then for any $x \in H$, there is a unique $y \in C$ closest to x in the sense that

$$\|x - y\| = \inf_{z \in C} \|x - z\| . \tag{4.2.16}$$

If it further holds that C is a subspace, then

$$\langle y - x, z \rangle = 0 \quad \text{for all } z \in C. \tag{4.2.17}$$

Proof. By translation, assume $x = 0$. Let $\delta = \inf_{z \in C} \|z\|$ is the distance from C to $x = 0$. Now, let $(y_n)_{n \in \mathbb{N}}$ be a sequence in C such that $\|y_n\| \to \delta$. Then

$$2\|y_n\|^2 + 2\|y_m\|^2 = \|y_n + y_m\|^2 + 2\|y_n - y_m\|^2 = 4\left\|\frac{1}{2}(y_n + y_m)\right\|^2 + \|y_n - y_m\|^2 . \tag{4.2.18}$$

Since C is convex, $\frac{1}{2}(y_n + y_m) \in C$, so that

$$2\|y_n\|^2 + 2\|y_m\|^2 \geq 4\delta^2 + \|y_n - y_m\|^2 . \tag{4.2.19}$$

If N is such that $\|y_n\|^2 \leq \delta^2 + \frac{\epsilon^2}{4}$ for all $n \geq N$, then

$$\|y_n - y_m\| \leq \epsilon, \quad n, m \geq N, \tag{4.2.20}$$

so that $(y_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, and therefore has limit y. Since C is closed, $y \in C$. Uniqueness follows by Eq. (4.2.19). Indeed, if $y, y' \in C$ and satisfy $\|y\| = \|y'\| = \delta$, then Eq. (4.2.19) implies $0 = \|y - y'\|^2$, and therefore $y = y'$.

Now, assume C is a subspace. For $z \in C$ and $t \in \mathbb{R}$, define

$$g(t) = \|y + tz - x\|^2 = \|y - x\| + 2t\langle y - x, z \rangle + t^2\|z\|^2 . \tag{4.2.21}$$

Since g achieves a minimum at $t = 0$, we have $g'(0) = 0$, which completes the proof. □

We finally arrive at the **Orthogonality Principle** for Hilbert spaces, which asserts that the projection of any point $x \in H$ into a closed subspace $U \subseteq H$ is characterized by the property that the residual is orthogonal to the subspace being projected on.

Corollary 4.2.9 (Orthogonality Principle). Let H be a Hilbert space, and $U \subseteq H$ a closed subspace. For any $x \in H$, there exists a unique closest point $u \in U$ to x. Moreover, $u \in U$ is the closest point to x if and only if

$$\langle u - x, z \rangle = 0 \quad \text{for all } z \in U. \tag{4.2.22}$$

Proof. Theorem 4.2.8 ensures that a closest point $u \in U$ to x exists, is unique, and satisfies Eq. (4.2.22). To see the converse, fix $v \in U$ and assume $u \in U$ satisfies Eq. (4.2.22). Now expand

$$\begin{aligned}
\|v - x\|^2 &= \|(v = u) + (u - x)\|^2 && (4.2.23)\\
&= \|v - u\|^2 + 2\langle v - u, u - x \rangle + \|u - x\|^2 && (4.2.24)\\
&= \|v - u\|^2 + \|u - x\|^2 && (4.2.25)\\
&\geq \|u - x\|^2 . && (4.2.26)
\end{aligned}$$

□

4.3 Best Linear Estimators

Notation 4.3.1. Let X be our random variable of interest, and $Y = (Y_i)_{i\in I}$ be our collection of observations, given some index set I. Define

$$\mathcal{L}(Y) = \text{ set of all linear estimators of } X \text{ given } Y \tag{4.3.1}$$

$$= \mathrm{Cl}_{L^2}\left(\left\{a + \sum_{i\in S} a_i Y_i \colon a, a_i \in \mathbb{R}, S \subseteq I, |S| < \infty\right\}\right). \tag{4.3.2}$$

Note that finite linear combinations form a subspace of L^2. The L^2 closure is just the set of limits in L^2 of the set.

Theorem 4.3.2 (Existence and Uniqueness of Best Linear Estimators). Let $X \in L^2$, $Y = (Y_i)_{i\in I}$ be observations. There exists a unique ($\mathbf{P}$-a.s.) random variable $\mathbf{L}(X \mid Y) \in \mathcal{L}(Y)$ such that

$$\mathbf{E}\Big(|X - \mathbf{L}(X \mid Y)|^2\Big) \le \mathbf{E}\Big(|X - L(Y)|^2\Big) \quad \text{for all } L(Y) \in \mathcal{L}(Y). \tag{4.3.3}$$

Moreover,

$$L(Y) = \mathbf{L}(X \mid Y) \quad \mathbf{P}\text{-a.s.} \tag{4.3.4}$$

$$\iff \underbrace{\mathbf{E}(L(Y)) = \mathbf{E}(X)}_{\text{unbiased}} \quad \text{and} \quad \underbrace{\mathbf{E}((X - L(Y))\, Y_i^*) = 0 \quad \text{for all } i \in I}_{\text{orthogonality principle}} \tag{4.3.5}$$

$$\iff \mathbf{E}\big((X - L(Y))\, L'(Y)^*\big) \quad \text{for all } L'(Y) \in \mathcal{L}(Y). \tag{4.3.6}$$

Proof. We claim the forward direction first. $\mathcal{L}(Y)$ is a closed subspace of L^2 by definition. So existence and uniqueness follows from Hilbert projection theorem. The same result says that for any $i \in I$ and $a, a_i \in \mathbb{R}$,

$$0 = \mathbf{E}((X - \mathbf{L}(X \mid Y))\,(a + a_i Y_i)^*) \tag{4.3.7}$$

$$= \mathbf{E}(X\,(a + a_i Y_i)^*) - \mathbf{E}(\mathbf{L}(X \mid Y)\,(a + a_i Y_i)^*) \tag{4.3.8}$$

$$= \mathbf{E}(a^* X) + \mathbf{E}(a_i^* X Y_i^*) - \mathbf{E}(a^*\,\mathbf{L}(X \mid Y)) - \mathbf{E}(a_i^*\,\mathbf{L}(X \mid Y) Y_i^*) \tag{4.3.9}$$

By taking $a_i = 0$ and $a = 1$, we see that $\mathbf{E}(X) = \mathbf{E}(\mathbf{L}(X \mid Y))$. By taking $a = 0$ and $a_i = 1$, we see that $\mathbf{E}((X - \mathbf{L}(X \mid Y))\, Y_i^*) = 0$.

Now suppose $L(Y)$ is unbiased and satisfies the orthogonality principle. Take any other $\widehat{L}(Y) \in L(Y)$. By definition there exists a sequence $(L_n(Y))_{n\in\mathbb{N}}$ is a sequence in $\mathcal{L}(Y)$ such that each L_n is a finite linear combination of the Y_i's and $L_n \to \widehat{L}$ in L^2. Then

$$\mathbf{E}\Big((X - L(Y))\,\widehat{L}(Y)^*\Big)^2 = \mathbf{E}\Big((X - L(Y))\,\Big(\widehat{L} - L_n\Big)^*\Big)^2 \tag{4.3.10}$$

$$\le \mathbf{E}\Big(|X - L(Y)|^2\Big)\,\mathbf{E}\Big(\Big|L_n - \widehat{L}\Big|^2\Big) \tag{4.3.11}$$

$$\to 0 \quad \text{as } n \to \infty. \tag{4.3.12}$$

So $X - L(Y) \perp \widehat{L} \in \mathcal{L}(Y)$, so Hilbert Projection Theorem tells us that $L(Y) = \mathbf{L}(X \mid Y)$ almost surely.

To see the second and third conditions being equal, note that Y_i and 1 are affine functions of Y, so the forward direction is trivial. For the backward direction, write $L'(Y) = \lim_{n\to\infty} L_n(Y)$ where $L_n(Y)$ is a finite affine combination of Y_i. Then by Dominated Convergence Theorem,

$$\mathbf{E}\big((X - L(Y))\, L'(Y)^*\big) \tag{4.3.13}$$

$$= \lim_{n\to\infty} \mathbf{E}((X - L(Y))\, L_n(Y)^*) \tag{4.3.14}$$

$$= \lim_{n\to\infty} \left(\mathbf{E}\left(X\left(a_n + \sum_{i\in S_n} a_{in} Y_{in}\right)^*\right) - \mathbf{E}\left(L(Y)\left(a_n + \sum_{i\in S_n} a_{in} Y_{in}\right)^*\right)\right) \tag{4.3.15}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \left(a^* \, \mathbf{E}(X) + \sum_{i\in S_n} a^*_{in} \, \mathbf{E}(XY^*_{in}) - a^* \, \mathbf{E}(L(Y)) - \sum_{i\in S_n} a^*_{in} \, \mathbf{E}(L(Y)Y^*_{in}) \right) && (4.3.16)\\
&= \lim_{n\to\infty} \left(a^*_n \, \mathbf{E}((X - L(Y))) + \sum_{i\in S_n} a^*_{in} \, \mathbf{E}((X - L(Y)) \, Y^*_{in}) \right) && (4.3.17)\\
&= \lim_{n\to\infty} \left(a^*_n \cdot 0 + \sum_{i\in S_n} a^*_{in} \cdot 0 \right) && (4.3.18)\\
&= 0. && (4.3.19)
\end{aligned}
$$

□

Example 4.3.3. If X is a random vector and Y is a random vector with $\Sigma_Y > 0$, then $\mathbf{L}(X \mid Y) = \mu_X + \Sigma_{XY}\Sigma_Y^{-1}\,(Y - \mu_Y)$.

Proof. Checking orthogonality principle,

$$
\begin{aligned}
\mathbf{E}(\mathbf{L}(X \mid Y)) &= \mu_X = \mathbf{E}(X) && (4.3.20)\\
0 &= \mathbf{E}((X - \mathbf{L}(X \mid Y))\, Y^*) && (4.3.21)\\
&= \mathbf{E}\big(\big(X - \mu_X - \Sigma_{XY}\Sigma_Y^{-1}\,(Y - \mu_Y)\big)\, Y^*\big) && (4.3.22)\\
&= \mathbf{E}((X - \mu_X)\,(Y - \mu_Y)^*) - \Sigma_{XY}\Sigma_Y^{-1}\,\mathbf{E}((Y - \mu_Y)\,(Y - \mu_Y)^*) && (4.3.23)\\
&= \Sigma_{XY} - \Sigma_{XY} && (4.3.24)\\
&= 0. && (4.3.25)
\end{aligned}
$$

□

Example 4.3.4. If $Y_n = X + Z_n$ and $Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then $\mathbf{L}(X \mid Y) = X$. This is because we can write $L_n(Y) = \frac{1}{n}\sum_{i\in[n]} Y_i = X + \frac{1}{n}\sum_{i\in[n]} Z_i$. This converges to X in L^2, so $X \in \mathcal{L}(Y)$ and is obviously the best linear estimator of X given Y.

Example 4.3.5. The best linear estimator is linear in the operator sense, i.e.,

$$
\mathbf{L}(aX_1 + bX_2 \mid Y) = a\,\mathbf{L}(X_1 \mid Y) + b\,\mathbf{L}(X_2 \mid Y). \tag{4.3.26}
$$

Proof. We want to check that the right hand side satisfies the orthogonality principle. First, we want to check that $a\,\mathbf{L}(X_1 \mid Y) + b\,\mathbf{L}(X_2 \mid Y) \in \mathcal{L}(Y)$. Indeed,

$$
\begin{aligned}
&\mathbf{E}(a\,\mathbf{L}(X_1 \mid Y) + b\,\mathbf{L}(X_2 \mid Y)) && (4.3.27)\\
&= a\,\mathbf{E}(\mathbf{L}(X_1 \mid Y)) + b\,\mathbf{E}(\mathbf{L}(X_2 \mid Y)) && (4.3.28)\\
&= a\,\mathbf{E}(X_1) + b\,\mathbf{E}(X_2) && (4.3.29)\\
&= \mathbf{E}(aX_1 + bX_2). && (4.3.30)\\
&\mathbf{E}((a\,\mathbf{L}(X_1 \mid Y) + b\,\mathbf{L}(X_2 \mid Y) - (aX_1 + bX_2))\, Y_i^*) && (4.3.31)\\
&= a\,\underbrace{\mathbf{E}((\mathbf{L}(X_1 \mid Y) - X_1)\, Y_i^*)}_{=0} + b\,\underbrace{\mathbf{E}((\mathbf{L}(X_2 \mid Y) - X_2)\, Y_i^*)}_{=0} && (4.3.32)\\
&= 0. && (4.3.33)
\end{aligned}
$$

□

Example 4.3.6. We saw earlier that if Y is a random vector with $\Sigma_Y > 0$, then

$$
\mathbf{L}(X \mid Y) = \mathbf{E}(X) + \Sigma_{XY}\Sigma_Y^{-1}\,(Y - \mathbf{E}(Y))\,. \tag{4.3.34}
$$

Suppose the Y_i's are zero mean and uncorrelated. Then

$$\mathbf{L}(X \mid Y_1, \dots, Y_n) = \mathbf{E}(X) + \sum_{i \in [n]} \frac{\mathrm{Cov}(X, Y_i)}{\mathrm{Var}(Y_i)} Y_i \tag{4.3.35}$$

$$= \mathbf{E}(X) + \sum_{i \in [n-1]} \frac{\mathrm{Cov}(X, Y_i)}{\mathrm{Var}(Y_i)} Y_i + \frac{\mathrm{Cov}(X, Y_n)}{\mathrm{Var}(Y_n)} Y_n \tag{4.3.36}$$

$$= \mathbf{L}(X \mid Y_1, \dots, Y_{n-1}) + \frac{\mathrm{Cov}(X, Y_n)}{\mathrm{Var}(Y_n)} Y_n. \tag{4.3.37}$$

4.4 Linear Innovation Sequences

Definition 4.4.1. Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence in L^2. We define corresponding **linear innovations**

$$\widehat{Y}_n := Y_n - \mathbf{L}(Y_n \mid Y_1, \dots, Y_{n-1}) \tag{4.4.1}$$

and $\widehat{Y}_1 = Y_1 - \mathbf{E}(Y_1)$.

Remark 4.4.2. The map $(1, Y_1, \dots, Y_n) \mapsto \left(1, \widehat{Y}_1, \dots, \widehat{Y}_n\right)$ is a change of basis for $\mathrm{Span}(1, Y_1, \dots, Y_n)$. In particular, $\mathbf{L}(X \mid Y_1, \dots, Y_n) = \mathbf{L}\left(X \,\middle|\, \widehat{Y}_1, \dots, \widehat{Y}_n\right)$, and $\sigma(Y_1, \dots, Y_n) = \sigma\left(\widehat{Y}_1, \dots, \widehat{Y}_n\right)$.

Remark 4.4.3. Note that

$$\mathbf{E}\left(\widehat{Y}_n\right) = \mathbf{E}(Y_n - \mathbf{L}(Y_n \mid Y_1, \dots, Y_{n-1})) = 0 \tag{4.4.2}$$

$$\mathbf{E}\left(\widehat{Y}_m \widehat{Y}_n\right) = \mathbf{E}((Y_m - \mathbf{L}(Y_m \mid Y_1, \dots, Y_{m-1}))\,(Y_n - \mathbf{L}(Y_n \mid Y_1, \dots, Y_{n-1}))) \tag{4.4.3}$$

$$= 0 \tag{4.4.4}$$

since $\widehat{Y}_n = Y_n - \mathbf{L}(Y_n \mid Y_1, \dots, Y_{n-1}) \in \mathcal{L}(Y_1, \dots, Y_n)$ for all n, and the sets $(\mathcal{L}(Y_1, \dots, Y_n))_{n \in \mathbb{N}}$ are an increasing sequence of sets. Hence $\left(\widehat{Y}_n\right)_{n \in \mathbb{N}}$ is a zero-mean uncorrelated sequence of random variables with the property

$$\mathbf{L}(X \mid Y_1, \dots, Y_n) = \mathbf{L}\left(X \,\middle|\, \widehat{Y}_1, \dots, Y_n\right). \tag{4.4.5}$$

In particular, if $(Y_n)_{n \in \mathbb{N}}$ is any sequence of observations,

$$\mathbf{L}(X \mid Y_1, \dots, Y_n) = \mathbf{L}\left(X \,\middle|\, \widehat{Y}_1, \dots, \widehat{Y}_n\right) \tag{4.4.6}$$

$$= \mathbf{L}\left(X \,\middle|\, \widehat{Y}_1, \dots, \widehat{Y}_{n-1}\right) + \frac{\mathrm{Cov}\left(X, \widehat{Y}_n\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)} \widehat{Y}_n \tag{4.4.7}$$

$$= \mathbf{L}(X \mid Y_1, \dots, Y_{n-1}) + \frac{\mathrm{Cov}\left(X, \widehat{Y}_n\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)} \widehat{Y}_n. \tag{4.4.8}$$

This is a way to recursively update our estimator as new observations arrive.

4.5 Kalman Filter

Consider a "state space model" of the following form:

$$\text{state evolution:} \qquad X_{n+1} = a_n X_n + U_n \tag{4.5.1}$$

$$\text{observation} \qquad Y_n = X_n + V_n \tag{4.5.2}$$

for $n \geq 0$. Here without loss of generality, X_0, $(U_n)_{n\in\mathbb{N}_0}$, and $(V_n)_{n\in\mathbb{N}_0}$ are all uncorrelated random variables. The following are known to us: the means and variances of all the random variables, and a_n. Without loss of generality, suppose the random variables are zero-mean.

The goal is to recursively compute $\widehat{X}_{n+1|n} := \mathbf{L}(X_{n+1} \mid Y_1, \ldots, Y_n)$ in an efficient way. We also want to track the error: $\sigma^2_{n+1|n} := \mathbf{E}\left(\left|X_{n+1} - \widehat{X}_{n+1|n}\right|^2\right)$.

Here is the Kalman Predictor:

```
Initialize $\widehat{X}_{0|-1} \leftarrow 0$
Initialize $\sigma^2_{0|-1} \leftarrow \mathrm{Var}(X_0)$
for $n \in \mathbb{N}$ do
    Compute "Kalman gain":
    $k_n = a_n\sigma^2_{n|n-1}\left(\sigma^2_{n|n-1} + \mathrm{Var}(V_n)\right)^{-1}.$        (4.5.3)
    Compute prediction estimate:
    $\widehat{X}_{n+1|n} = a_n\widehat{X}_{n|n-1} + k_n\left(Y_n - \widehat{X}_{n|n-1}\right).$        (4.5.4)
    Compute prediction error estimate:
    $\sigma^2_{n+1|n} = a_n\left(a_n - k_n\right)\sigma^2_{n|n-1} + \mathrm{Var}(U_n)$        (4.5.5)
```

Theorem 4.5.1. The Kalman prediction algorithm stated above satisfies $\widehat{X}_{n+1|n} = \mathbf{L}(X_{n+1} \mid Y_0, Y_1, \ldots, Y_i)$ and $\sigma^2_{n+1|n}$ is equal to the variance of the corresponding prediction error.

Proof. Let $\left(\widehat{Y}_n\right)_{n\in\mathbb{N}_0}$ be the linear innovation sequence corresponding to $(Y_n)_{n\in\mathbb{N}_0}$. Then

$$\begin{aligned}
\widehat{X}_{n+1|n} &= \mathbf{L}(X_{n+1} \mid Y_0, \ldots, Y_n) && (4.5.6)\\
&= \mathbf{L}(X_n + U_n \mid Y_0, \ldots, Y_n) && (4.5.7)\\
&= \mathbf{L}(X_n \mid Y_0, \ldots, Y_{n-1}) + \frac{\mathrm{Cov}\left(X_n, \widehat{Y}_n\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.8)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}\left(X_n, \widehat{Y}_n\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.9)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}(X_n, Y_n - \mathbf{L}(Y_n \mid Y_0, \ldots, Y_{n-1}))}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.10)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}(X_n, X_n + V_n - \mathbf{L}(X_n + V_n \mid Y_0, \ldots, Y_{n-1}))}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.11)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}(X_n, X_n - \mathbf{L}(X_n \mid Y_0, \ldots, Y_{n-1}))}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.12)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}\left(X_n, X_n - \widehat{X}_{n|n-1}\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.13)\\
&= \widehat{X}_{n|n-1} + \frac{\mathrm{Cov}\left(X_n - \widehat{X}_{n|n-1}, X_n - \widehat{X}_{n|n-1}\right)}{\mathrm{Var}\left(\widehat{Y}_n\right)}\widehat{Y}_n && (4.5.14)
\end{aligned}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\operatorname{Var}\left(\widehat{Y}_n\right)} \widehat{Y}_n \tag{4.5.15}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\operatorname{Cov}\left(\widehat{Y}_n, \widehat{Y}_n\right)} \widehat{Y}_n \tag{4.5.16}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\operatorname{Cov}(Y_n - \mathbf{L}(Y_n \mid Y_0, \ldots, Y_{n-1}), Y_n - \mathbf{L}(Y_n \mid Y_0, \ldots, Y_{n-1}))} \widehat{Y}_n \tag{4.5.17}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\operatorname{Cov}(Y_n, Y_n - \mathbf{L}(Y_n \mid Y_0, \ldots, Y_{n-1}))} \widehat{Y}_n \tag{4.5.18}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\operatorname{Cov}(X_n + V_n, X_n + V_n - \mathbf{L}(X_n + V_n \mid Y_0, \ldots, Y_{n-1}))} \widehat{Y}_n \tag{4.5.19}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \operatorname{Var}(V_n)} \widehat{Y}_n \tag{4.5.20}$$

$$= \widehat{X}_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \operatorname{Var}(V_n)} \left(Y_n - \widehat{Y}_{n|n-1}\right). \tag{4.5.21}$$

It remains to compute the estimation error:

$$\sigma^2_{n+1|n} = \operatorname{Cov}\left(X_{n+1} - \widehat{X}_{n+1|n}, X_{n+1} - \widehat{X}_{n+1|n}\right) \tag{4.5.22}$$

$$= \operatorname{Cov}\left(X_n + U_n, X_n + U_n - \widehat{X}_{n|n-1} - k_n \left(X_n + V_n - \widehat{X}_{n|n-1}\right)\right) \tag{4.5.23}$$

$$= \operatorname{Cov}\left(X_n, \left(X_n - \widehat{X}_{n|n-1}\right)(1 - k_n)\right) + \operatorname{Var}(U_n) \tag{4.5.24}$$

$$= (1 - k_n) \operatorname{Cov}\left(X_n - \widehat{X}_{n|n-1}, X_n - \widehat{X}_{n|n-1}\right) + \operatorname{Var}(U_n) \tag{4.5.25}$$

$$= (1 - k_n)\, \sigma^2_{n|n-1} + \operatorname{Var}(U_n). \tag{4.5.26}$$

□

5 Time Series Analysis

Time series analysis is linear estimation when the observation is a random process.

5.1 Second-Order Processes

Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stochastic process on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Definition 5.1.1 (Second-Order Process). X is a (discrete-time) **second-order process** if it has finite second moments:

$$\mathbf{E}\left(|X_n|^2\right) < \infty \quad \text{for all } n \in \mathbb{Z}. \tag{5.1.1}$$

In particular, we require all X_n to be elements of $L^2(\Omega, \mathcal{F}, \mathbf{P})$.

Proposition 5.1.2. The set of second-order processes is a vector space.

Proof. Since $L^2(\Omega, \mathcal{F}, \mathbf{P})$ is a vector space, the claim follows from definitions. □

Definition 5.1.3. For second-order processes X and Y, the second-order statistics are summarized by the

$$\textbf{mean function} \qquad \mu_X(n) := \mathbf{E}(X_n) \quad n \in \mathbb{Z} \tag{5.1.2}$$

$$\textbf{covariance function} \qquad R_{XY}(m,n) := \operatorname{Cov}(X_m, Y_n) \quad m,n \in \mathbb{Z}. \tag{5.1.3}$$

Example 5.1.4. If X is a Gaussian process, then all finite-dimensional marginals are characterized by the functions μ_X and R_{XX}. If $(X_n)_{n\in\mathbb{Z}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, then $\mu_X(n) = 0$, $R_{XX}(m,n) = \sigma^2 \delta_{mn}$.

Example 5.1.5. Consider a sinusoidal process with a random phase:

$$X_n = \alpha \sin(\omega_0 n + \Theta), \quad n \in \mathbb{Z}, \tag{5.1.4}$$

where Θ is uniform over $[0, 2\pi)$. In this case, the mean can be computed as

$$\mu_X(n) = \mathbf{E}(X_n) = \mathbf{E}(\alpha \sin(\omega_0 n + \Theta)) = 0. \tag{5.1.5}$$

and the covariance can be computed as

$$\begin{aligned} R_{XX}(m,n) &= \frac{1}{2\pi} \int_0^{2\pi} \alpha^2 \sin(\omega_0 m + \theta) \sin(\omega_0 n + \theta) \,\mathrm{d}\theta \qquad (5.1.6) \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{\alpha^2}{2} \left(\cos(\omega_0 (m-n)) - \cos(\omega_0 (m+n) + 2\theta)\right) \mathrm{d}\theta \qquad (5.1.7) \\ &= \frac{\alpha^2}{2} \cos(\omega_0 (m-n)). \qquad (5.1.8) \end{aligned}$$

An important observation is that the autocorrelation function in the above example is only a function of the difference $m - n$. This is a defining feature of so-called **wide-sense stationary** processes, which we will come to shortly.

Definition 5.1.6 (Stationary Process). A process X is **stationary** if for any integers $n_1 < \cdots < n_k$ and m,

$$F_{X_{n_1}, \ldots, X_{n_k}}(x_1, \ldots, x_k) = F_{X_{n_1+m}, \ldots, X_{n_k+m}}(x_1, \ldots, x_k). \tag{5.1.9}$$

A process is stationary if the distribution of any finite set of samples is invariant to shifts in time.

Stationarity is often too strict of an assumption. However, the concept may be relaxed to require that only the second-order statistics are invariant to time shifts. This turns out to be a fairly good model for processes encountered in practice. This motivates the following definition.

Definition 5.1.7 (WSS Process). A second-order process X is **wide-sense stationary** (WSS) if $\mu_X(n) = \mu_X(0)$ for all n, and $R_{XX}(m,n)$ is a function of only the difference $m-n$. In this case, we often abbreviate $R_{XX}(m,n)$ as $R_{XX}(m-n) := R_{XX}(m-n,0)$ to denote parameterization of the covariance function by the difference $m-n$.

A wide sense stationary process does *not* have to be stationary. WSS only states that the mean and covariance are invariant with respects to shifts in time. However, for a Gaussian process, WSS is equivalent to stationary due to the fact that jointly Gaussian random variables have distribution parameterized by first and second moments.

Example 5.1.8 (Autoregressive Process). Let $|\alpha| < 1$. If X is defined by

$$X_{n+1} = \alpha X_n + W_n, \quad n \geq 0 \tag{5.1.10}$$

and $X_n = 0$ for $n \leq 0$, and $(W_n)_{n\in\mathbb{N}_0}$ is uncorrelated, unit variance random variables, then

$$\mu_X(n) = 0 \tag{5.1.11}$$

$$R_{XX}(m,n) = \begin{cases} \alpha^{|m-n|}\frac{1-\alpha^{2n}}{1-\alpha^2} & m,n \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.1.12}$$

So for n large, $R_{XY}(m,n) \approx \alpha^{|m-n|}$, so it is approximately WSS.

Definition 5.1.9 (Jointly WSS). Processes X and Y are **jointly wide sense stationary** (JWSS) if each are WSS and the covariance function $R_{XY}(m,n) = \mathrm{Cov}(X_m, Y_n)$ depends only on the difference $m-n$. In this case, we abbreviate $R_{XY}(m,n)$ as $R_{XY}(m-n)$.

Example 5.1.10. Suppose X is WSS. Fix k and denote $Y_n = \alpha X_{n-k}$ for $n \in \mathbb{Z}$. Then Y is WSS. It holds that (X,Y) are JWSS because

$$R_{XY}(m,n) = \mathrm{Cov}(X_m, Y_n) \tag{5.1.13}$$

$$= \mathrm{Cov}(X_m, \alpha X_{n-k}) \tag{5.1.14}$$

$$= \alpha R_{XX}(m-n+k). \tag{5.1.15}$$

For WSS processes the covariance enjoys the following symmetries:

$$R_{XX}(n,n+k) = R_{XX}(0,k) = R_{XX}(k,0) = R_{XX}(0,-k). \tag{5.1.16}$$

In our compact notation, $R_{XX}(k) = R_{XX}(-k)$, so that R_{XX} is a symmetric function of k.

Unlike R_{XX}, the function R_{XY} is not symmetric in its argument. However, if X and Y are JWSS, then we have the following identities:

$$R_{XY}(n+k,n) = R_{XY}(k,0) = R_{XY}(0,-k). \tag{5.1.17}$$

In particular, noting the order of subscripts, we have

$$R_{XY}(k) = R_{YX}(-k). \tag{5.1.18}$$

5.2 Discrete-Time Fourier Transform

Definition 5.2.1. Let $1 \leq p \leq \infty$. We write $x \in \ell^p(\mathbb{Z})$ if x satisfies

$$\|x\|_{\ell^p} = \begin{cases} \left(\sum_{n\in\mathbb{Z}} |x(n)|^p\right)^{1/p} & 1 \leq p < \infty \\ \sup_{n\in\mathbb{Z}} |x(n)| & p = \infty \end{cases} < \infty. \tag{5.2.1}$$

Corollary 5.2.2. Let $1 \leq p \leq \infty$. Then $\ell^p(\mathbb{Z})$ is a Banach space, i.e., it is complete with respect to convergence in its norm $\|\cdot\|_{\ell^p}$.

Proposition 5.2.3. Let $1 \leq p \leq q \leq \infty$. Then $\ell^p(\mathbb{Z}) \subseteq \ell^q(\mathbb{Z})$ on account of $\|x\|_{\ell^q} \leq \|x\|_{\ell^p}$ (which can be shown through Jensen's inequality).

Proposition 5.2.4. For $1 < p < \infty$,

$$\ell^p(\mathbb{Z}) = \mathrm{Cl}_{\ell^p}\left(\ell^1(\mathbb{Z})\right). \tag{5.2.2}$$

To clarify, the closure is in the topology induced by $\|\cdot\|_{\ell^p}$.

Proposition 5.2.5. Of note, $\ell^2(\mathbb{Z})$ is a Hilbert space when equipped with the inner product

$$\langle x, y\rangle_{\ell^2} := \sum_{n\in\mathbb{Z}} x(n)y(n)^*, \quad x, y \in \ell^2(\mathbb{Z}). \tag{5.2.3}$$

Definition 5.2.6 (DTFT). Let $x \in \ell^1(\mathbb{Z})$. Its discrete-time Fourier transform is the complex-valued function

$$\widehat{x}(\omega) := \sum_{n\in\mathbb{Z}} x(n)\mathrm{e}^{-\mathrm{i}\omega n}, \quad \omega \in [-\pi, \pi). \tag{5.2.4}$$

Proposition 5.2.7. The mapping $x \mapsto \widehat{x}$ is a linear map from $\ell^1(\mathbb{Z})$ to $L^\infty([-\pi,\pi))$. In particular, $|\widehat{x}(\omega)| \leq \|x\|_{\ell^1} < \infty$.

Proposition 5.2.8 (IDTFT). The DTFT is invertible:

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{x}(\omega)\mathrm{e}^{\mathrm{i}\omega n}\,\mathrm{d}\omega. \tag{5.2.5}$$

Proof.

$$\begin{aligned} \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{x}(\omega)\mathrm{e}^{\mathrm{i}\omega n}\,\mathrm{d}\omega &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\sum_{k\in\mathbb{Z}} \mathrm{e}^{-\mathrm{i}\omega k}\right)\mathrm{e}^{\mathrm{i}\omega n}\,\mathrm{d}\omega && (5.2.6)\\ &= \sum_{k\in\mathbb{Z}} x(k)\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\omega(n-k)}\,\mathrm{d}\omega\right) && (5.2.7)\\ &= \sum_{k\in\mathbb{Z}} x(k)\delta(n-k) && (5.2.8)\\ &= x(n), && (5.2.9) \end{aligned}$$

where the exchange of the sum and integral is justified by Fubini-Tonelli. □

Theorem 5.2.9 (Parseval's Identity). If $x, y \in \ell^1(\mathbb{Z})$, then

$$\langle x, y\rangle_{\ell^2} = \langle \widehat{x}, \widehat{y}\rangle_{L^2_{\mathbb{C}}}. \tag{5.2.10}$$

Corollary 5.2.10. The mapping $x \mapsto \widehat{x}$ is a linear isometry from $\ell^1(\mathbb{Z}) \cap \ell^2(\mathbb{Z})$ into $L^2([-\pi,\pi))$, i.e.,

$$\|x\|_{\ell^2} = \|\widehat{x}\|_{L^2}. \tag{5.2.11}$$

We can extend the Fourier transform $x \mapsto \widehat{x}$ to a linear isometry from $\ell^2(\mathbb{Z})$ into $L^2([-\pi,\pi))$. That is, let $x \in \ell^2(\mathbb{Z})$. Then there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of elements in $\ell^1(\mathbb{Z})$ such that

$$\lim_{n\to\infty} \|x_n - x\|_{\ell^2} = 0. \tag{5.2.12}$$

By Parseval's Identity, $(\widehat{x}_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^2([-\pi,\pi))$. Therefore there exists $\widehat{x} \in L^2([-\pi,\pi))$ such that

$$\lim_{n\to\infty} \|\widehat{x}_n - \widehat{x}\|_{L^2_{\mathbb{C}}} = 0. \tag{5.2.13}$$

So we define $\widehat{x} \in L^2([-\pi,\pi))$ as the Fourier transform of $x \in \ell^2(\mathbb{Z})$.

For $x \in \ell^2(\mathbb{Z})$, we still define its DTFT as

$$\widehat{x}(\omega) = \sum_{n\in\mathbb{Z}} x(n)\mathrm{e}^{-\mathrm{i}\omega n} \tag{5.2.14}$$

where the sum converges in an L^2 sense.

Theorem 5.2.11. If $x, y \in \ell^2(\mathbb{Z})$ and $\widehat{y} \in L^\infty([-\pi,\pi))$, and $z = x * y$ (i.e., $z(n) = \sum_{k\in\mathbb{Z}} x(k)y(n-k)$), then all Fourier transforms exist, and

$$\widehat{z}(\omega) = \widehat{x}(\omega)\widehat{y}(\omega). \tag{5.2.15}$$

5.3 Spectral Theory of WSS Processes

For $x \in \ell^1(\mathbb{Z})$, we can define the sequence $a \in \ell^1(\mathbb{Z})$ via the auto-correlation

$$a(n) = \sum_{k\ in\mathbb{Z}} x(k)x(k-n), \quad n \in \mathbb{Z}. \tag{5.3.1}$$

By the convolution theorem and time-reversal property of Fourier transforms, the discrete-time Fourier transform of a is equal to

$$\widehat{a}(\omega) = \widehat{x}(\omega)\widehat{x}^*(\omega) = |\widehat{x}(\omega)|^2 \geq 0. \tag{5.3.2}$$

The function $\widehat{a}$ is called the **energy spectral density** of x, since it is a non-negative function with the property that its integral over any subset of frequencies in $[-\pi,\pi)$ is equal to the energy of the sequence x restricted to those frequencies. Indeed, integrating over the entire interval recovers the energy of the sequence x:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{a}(\omega)\,\mathrm{d}\omega = a(0) = \sum_{n\in\mathbb{Z}} |x(n)|^2, \tag{5.3.3}$$

where we used Fourier inversion and definition of a.

We would like to apply these ideas to a zero-mean WSS random process X to similarly characterize spectral properties of the process and subsequently develop a theory of linear esitmation of such processes. Unfortunately, we typically have $\sum_{n\in\mathbb{Z}} |X_n| = +\infty$ almost surely. To get around this, consider the truncated process $X_n^N = X_n 1_{\{|n|\leq N\}}$. Then $\left(X_n^N\right)_{n\in\mathbb{Z}} \in \ell^1(\mathbb{Z})$ almost surely. In this case, the energy spectral density of the truncated signal can be computed as

$$\begin{aligned} A^N(\omega) &= \widehat{X}^N(\omega)\widehat{X}^{N*}(\omega) && (5.3.4)\\ &= \left(\sum_{n=-N}^{N} X_n\mathrm{e}^{-\mathrm{i}\omega n}\right)\left(\sum_{m=-N}^{N} X_m\mathrm{e}^{\mathrm{i}\omega m}\right) && (5.3.5)\\ &= \sum_{n=-N}^{N}\sum_{m=-N}^{N} X_m X_n\mathrm{e}^{-\mathrm{i}\omega(m-n)}. && (5.3.6) \end{aligned}$$

Note that power is the average energy per unit time. Normalizing,

$$\frac{1}{2N+1}\mathbf{E}\big(A^N(\omega)\big) = \frac{1}{2N+1}\sum_{m=-N}^{N}\sum_{n=-N}^{N} R_{XX}(m-n)\mathrm{e}^{-\mathrm{i}\omega(m-n)} \tag{5.3.7}$$

$$= \frac{1}{2N+1}\sum_{k=-N}^{N} R_{XX}(k)\mathrm{e}^{-\mathrm{i}\omega k}\left(1-\frac{|k|}{2N+1}\right). \tag{5.3.8}$$

If $R_{XX} \in \ell^1(\mathbb{Z})$, then the limit as $N \to \infty$ exists on the right by dominated convergence, so that

$$S_{XX}(\omega) := \lim_{N\to\infty}\frac{1}{2N+1}\mathbf{E}\big(A^N(\omega)\big) = \sum_{n\in\mathbb{N}} R_{XX}(n)\mathrm{e}^{-\mathrm{i}\omega n} = \widehat{R}_{XX}(\omega), \quad \omega \in [-\pi,\pi). \tag{5.3.9}$$

The function S_{XX} is the **power spectral density** of the process X, and earns its name from the above derivation; it is a real non-negative function. The definition of power spectral density can be extended to WSS processes X with $R_{XX} \in \ell^2(\mathbb{Z})$ using mean-square convergence of the Fourier transform. In this case, S_{XX} continues to be real and non-negative.

As a purely mathematical gadget, we derive the "**cross-power spectral density**" as

$$S_{YX}(\omega) := \sum_{n\in\mathbb{Z}} R_{YX}(n)\mathrm{e}^{-\mathrm{i}\omega n} = \sum_{n\in\mathbb{Z}} \mathbf{E}\big(Y_m X_{m-n}^*\big)\mathrm{e}^{-\mathrm{i}\omega n}. \tag{5.3.10}$$

Here the convergence is absolute or in mean-square sense as necessary.

Example 5.3.1. A zero-mean WSS process $Z = (Z_n)_{n\in\mathbb{Z}}$ is a **white noise process** if $R_{ZZ}(n) = \delta_n \operatorname{Var}(Z_0)$, so that $S_{ZZ}(\omega) = \operatorname{Var}(Z_0)$.

Example 5.3.2. Power spectral definition is a non-negative function. By Fourier inversion,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XX}(\omega)\,\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XX}(\omega)\mathrm{e}^{\mathrm{i}\omega\cdot 0}\,\mathrm{d}\omega = R_{XX}(\omega) = \operatorname{Var}(X_0). \tag{5.3.11}$$

Definition 5.3.3. We say X has **regular covariance** if

1. $R_{XX} \in \ell^2(\mathbb{Z})$ (ensuring S_{XX} exists); and
2. $0 < \operatorname{ess\,inf}_{\omega\in[-\pi,\pi)} S_{XX}(\omega) \le \operatorname{ess\,sup}_{\omega\in[-\pi,\pi)} S_{XX}(\omega) < \infty$.

Example 5.3.4. Let $X = (Z_n)_{n\in\mathbb{N}}$ be zero-mean WSS and uncorrelated with a white noise process $Z = (Z_n)_{n\in\mathbb{Z}}$. If $R_{XX} \in \ell^1(\mathbb{Z})$ (i.e., X does not have *long-range dependence*), then the process $Y_n = X_n + Z_n$ has regular covariance. In many practical situations, the observation process will not have long range dependence, and is typically contaminated by an additive white noise process. Hence, the assumption of a regular covariance will often be satisfied.

5.4 Linear Estimation from WSS Observations

Suppose $Y = (Y_n)_{n\in\mathbb{Z}}$ is a zero-mean WSS process with regular covariance, $I \subseteq \mathbb{Z}$, and $Y_I = (Y_i)_{i\in I}$ are the observations of Y we make.

Theorem 5.4.1. For any zero-mean $X \in L^2(\Omega,\mathcal{F},\mathbf{P})$, there exists $h \in \ell^2(\mathbb{Z})$ such that

$$\mathbf{L}(X \mid Y_I) = \sum_{n\in I} h(n)Y_n \quad \text{in } L^2. \tag{5.4.1}$$

Moreover, the sequence h is unique on indices in I.

Importantly, we can express $\mathbf{L}(X \mid Y_I)$ as a linear combination of observations. This is different from before, where we were able to take L^2 limits of linear combinations to get the best linear estimators.

Proof. Let a be any sequence of real numbers, with only finitely many nonzero elements. By linearity and Fourier inversions,

$$\mathbf{E}\left(\left|\sum_{n\in\mathbb{Z}} a(n)Y_n\right|^2\right) = \sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} a(m)a(n)R_{YY}(m-n) \tag{5.4.2}$$

$$= \sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} a(m)a(n)\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{YY}(\omega)\mathrm{e}^{\mathrm{i}\omega(m-n)}\,\mathrm{d}\omega \tag{5.4.3}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{YY}(\omega)\left(\sum_{m\in\mathbb{Z}} a(m)\mathrm{e}^{\mathrm{i}\omega m}\right)\left(\sum_{n\in\mathbb{Z}} a(n)\mathrm{e}^{-\mathrm{i}\omega n}\right) \tag{5.4.4}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{YY}(\omega)\left|\widehat{a}(\omega)\right|^2\mathrm{d}\omega. \tag{5.4.5}$$

Using the fact that R_{YY} is regular, there is $\lambda > 0$ such that

$$\frac{\lambda}{2\pi}\int_{-\pi}^{\pi}\left|\widehat{a}(\omega)\right|^2\mathrm{d}\omega \leq \mathbf{E}\left(\left|\sum_{n\in\mathbb{Z}} a(n)Y_n\right|^2\right) \leq \frac{\lambda^{-1}}{2\pi}\int_{-\pi}^{\pi}\left|\widehat{a}(\omega)\right|^2\mathrm{d}\omega \tag{5.4.6}$$

$$\lambda\left\|\widehat{a}\right\|_{L^2}^2 \leq \mathbf{E}\left(\left|\sum_{n\in\mathbb{Z}} a(n)Y_n\right|^2\right) \leq \lambda^{-1}\left\|\widehat{a}\right\|_{L^2}^2 \tag{5.4.7}$$

$$\lambda\left\|a\right\|_{\ell^2}^2 \leq \mathbf{E}\left(\left|\sum_{n\in\mathbb{Z}} a(n)Y_n\right|^2\right) \leq \lambda^{-1}\left\|a\right\|_{\ell^2}^2. \tag{5.4.8}$$

By definition of $\mathbf{L}(X \mid Y_I)$, there exists a sequence of finite linear combinations $(L_n(Y))_{n\in\mathbb{N}}$ such that

$$L_n(Y) = \sum_{i\in I} a_n(i)Y_i \tag{5.4.9}$$

and $L_n(Y) \to \mathbf{L}(X \mid Y_I)$ in L^2. Hence

$$\lambda\left\|a_m - a_n\right\|_{\ell^2}^2 \leq \left\|L_m(Y) - L_n(Y)\right\|_{L^2}^2 \to 0 \quad \text{as } m, n \to \infty. \tag{5.4.10}$$

Hence $(a_n)_{n\in\mathbb{Z}}$ is Cauchy as a sequence in $\ell^2(\mathbb{Z})$. Thus there exists a $h \in \ell^2(\mathbb{Z})$ such that $\lim_{n\to\infty}\left\|a_n - h\right\|_{\ell^2} = 0$ and $h(n) \neq 0$ for $n \notin I$.

To show that $\sum_{n\in I} h(n)Y_n$ converges in L^2 to $\mathbf{L}(X \mid Y_I)$, we fix a rearrangement $\sigma = \{1, \dots, |I|\}$. Define

$$Z_{\sigma,N} := \sum_{n=1}^{N} h(\sigma(n))Y_{\sigma(n)} = \sum_{n\in I} h_{\sigma,N}(n)Y_n. \tag{5.4.11}$$

Then

$$\mathbf{E}\Big(\left|L_n(Y) - Z_{\sigma,N}\right|^2\Big) \leq \frac{1}{\lambda}\left\|a_n - h_{\sigma,N}\right\|_{\ell^2}^2 \tag{5.4.12}$$

$$\leq \frac{2}{\lambda}\left(\left\|a_n - h\right\|_{\ell^2}^2 + \left\|h_{\sigma,N} - h\right\|_{\ell^2}^2\right) \tag{5.4.13}$$

$$\lim_{N\to\infty}\lim_{n\to\infty}\mathbf{E}\Big(\left|L_n(Y) - Z_{\sigma,N}\right|^2\Big) = \lim_{N\to\infty}\mathbf{E}\Big(\left|\mathbf{L}(X \mid Y_I) - Z_{\sigma,N}\right|^2\Big) \tag{5.4.14}$$

$$\leq \lim_{N\to\infty}\frac{2}{\lambda}\left\|h_{\sigma,N} - h\right\|_{\ell^2}^{@} \tag{5.4.15}$$

$$= 0. \tag{5.4.16}$$

□

Now the natural question is how to find h. Combining the representation of $\mathbf{L}(X \mid Y_I)$ guaranteed by teh above theorem with the orthogonality principle explicitly characterizes the optimal coefficients h as the unique solution to a system of linear equations known as the **Wiener-Hopf equations**. Under the assumptions given, they are equivalent to the orthogonality principle, and are therefore necessary and sufficient for optimality.

Theorem 5.4.2 (Wiener-Hopf Equations). Fix $I \subseteq \mathbb{Z}$ and let $Y = (Y_n)_{n\in\mathbb{Z}}$ be a zero-mean WSS process with regular covariance. The sequence $h \in \ell^2(\mathbb{Z})$ defining the best linear estimator $\mathbf{L}(X \mid Y_I)$ uniquely solves the system of equations

$$\begin{cases} \langle X, Y_n \rangle_{L^2} = (R_{YY} * h)(n), & n \in I \\ h(n) = 0, & n \notin I \end{cases}. \tag{5.4.17}$$

Proof. By the orthogonality principle, we have

$$\langle X, Y_n \rangle_{L^2} = \langle \mathbf{L}(X \mid Y_I), Y_n \rangle_{L^2}, \quad i \in I. \tag{5.4.18}$$

We necessarily have $\sum_{n\in I} h(n) \langle Y_n, Y_i \rangle_{L^2} \to \mathbf{L}(X \mid Y_I) Y_i$ in expectation, so

$$\begin{aligned} 0 &= \langle \mathbf{L}(X) Y_I, Y_n \rangle_{L^2} && (5.4.19) \\ &= \left\langle \lim_{N\to\infty} \sum_{m=-N}^{N} h(m) Y_m, Y_n \right\rangle_{L^2} && (5.4.20) \\ &= \lim_{N\to\infty} \sum_{m=-N}^{N} h(m) \langle Y_m, Y_n \rangle_{L^2} && (5.4.21) \\ &= \lim_{N\to\infty} \sum_{m=-N}^{N} h(m) R_{YY}(n-m) && (5.4.22) \\ &= \sum_{m\in\mathbb{Z}} h(m) R_{YY}(n-m) && (5.4.23) \\ &= (R_{YY} * h)(n). && (5.4.24) \end{aligned}$$

The sum converges absolutely by Cauchy-Schwarz. This shows that h satisfies the stated linear equations. It is unique by the orthogonality principle. □

When $I = \mathbb{Z}$, we can explicitly solve the Wiener-Hopf equations using Fourier transforms. A similar result can be stated when $I = \{n \in \mathbb{Z} \colon n \le k\}$ for an integer k, but this requires a more sophisticated spectral analysis.

Corollary 5.4.3. Let $Y = (Y_n)_{n\in\mathbb{Z}}$ be a zero-mean WSS stationary process with regular covariance. The sequence $h \in \ell^2(\mathbb{Z})$ defining the best linear estimator $\mathbf{L}(X \mid Y)$ in Theorem 5.4.1 has Fourier transform

$$H(\omega) = \frac{S_{YX}(\omega)}{S_{YY}(\omega)}, \quad \omega \in [-\pi, \pi). \tag{5.4.25}$$

The resulting estimation error is

$$\mathbf{E}\Big(|X - \mathbf{L}(X \mid Y)|^2 \Big) = \mathrm{Var}(X) - \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|S_{YX}(\omega)|^2}{S_{YY}(\omega)} \, \mathrm{d}\omega. \tag{5.4.26}$$

Proof. We have the Wiener-Hopf equations

$$\underbrace{\mathbf{E}(XY_n)}_{\in\ell^2(\mathbb{Z})} = \Bigg(\underbrace{R_{YY}}_{\in\ell^1(\mathbb{Z})} * \underbrace{h}_{\in\ell^2} \Bigg)(n).$$

Taking Fourier transforms,

$$S_{YX}(\omega) = S_{YY}(\omega)H(\omega). \tag{5.4.27}$$

Thus

$$H(\omega) = \frac{S_{YX}(\omega)}{S_{YY}(\omega)}. \tag{5.4.28}$$

Now it remains to show the estimation error. Indeed,

$$\begin{aligned}
\langle \mathbf{L}(X \mid Y), X\rangle_{L^2} &= \lim_{N\to\infty} \sum_{n=-N}^{N} h(n) \langle X, Y_n\rangle_{L^2} && (5.4.29)\\
&= \sum_{n\in\mathbb{Z}} h(n) \langle X, Y_n\rangle_{L^2} && (5.4.30)\\
&= \left\langle h, \langle X, Y_{(\cdot)}\rangle_{L^2} \right\rangle_{\ell^2} && (5.4.31)\\
&= \langle h, R_{YX}\rangle && (5.4.32)\\
&= \langle H, S_{YX}\rangle_{L^2} && (5.4.33)\\
&= \left\langle \frac{S_{YX}}{S_{YY}}, S_{YX} \right\rangle_{L^2} && (5.4.34)\\
&= \mathbf{E}\left(\frac{|S_{YX}|^2}{S_{YY}}\right) && (5.4.35)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|S_{YX}(\omega)|^2}{S_{YY}(\omega)}\,\mathrm{d}\omega. && (5.4.36)
\end{aligned}$$

□

5.5 Non-Causal Wiener Filter

We will now go to a slight generalization of the previous setting. We now consider the setting of two JWSS processes X and Y, where our goal is to estimate the X process from the observed Y process. We assume the entire Y process is available to us, which leads us to the non-causal Wiener filter. The key idea is that the sequence of best linear estimators can be realized by filtering the observation process. Under the mild conditions stated, the optimal filter always exists and has an explicit characterization in terms of its frequency response.

Theorem 5.5.1. Let $(X) = (X_n)_{n\in\mathbb{Z}}$ and $Y = (Y_n)_{n\in\mathbb{Z}}$ be zero-mean JWSS processes, and assume that Y has regular covariance. The process $(\mathbf{L}(X_n \mid Y))_{n\in\mathbb{Z}}$ can be realized by passing Y through an LTI system with frequency response

$$H(\omega) = \frac{S_{YX}(\omega)^*}{S_{YY}(\omega)}. \tag{5.5.1}$$

The resulting estimation error is

$$\mathbf{E}\left(|X_n - \mathbf{L}(X_n \mid Y)|^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(S_{XX}(\omega) - \frac{|S_{YX}(\omega)|^2}{S_{YY}(\omega)}\right) \mathrm{d}\omega. \tag{5.5.2}$$

Proof. We aim to find the impulse response h such that

$$\mathbf{L}(X_n \mid Y) = \sum_{k\in\mathbb{Z}} h(n-k)Y_k \quad \text{in } L^2.$$

By the JWSS assumption, we can take $n = 0$, and thus want to find h such that

$$\mathbf{L}(X_0 \mid Y) = \sum_{k\in\mathbb{Z}} h(-k)Y_k \quad \text{in } L^2 \quad . \tag{5.5.3}$$

Noting the time reversal of h, by Corollary 5.4.3 the frequency response satisfies

$$H(\omega) = \left(\frac{S_{YX}(\omega)}{S_{YY}(\omega)}\right)^* = \frac{S^*_{YX}(\omega)}{S_{YY}(\omega)}. \tag{5.5.4}$$

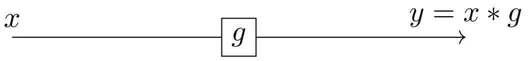
The error estimate is from Corollary 5.4.3 as well, giving

$$\mathbf{E}\Big(|X_n - \mathbf{L}(X_n \mid Y)|^2\Big) = \mathrm{Var}(X_n) - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|S_{YX}(\omega)|^2}{S_{YY}(\omega)}\,\mathrm{d}\omega. \tag{5.5.5}$$

□

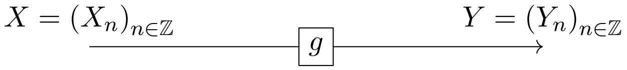
5.6 Linear Filtering for WSS Processes

Recall that an LTI system has the form



LTI systems are "stable" if bounded inputs lead to bounded outputs. Indeed, the system with impulse response g is stable if and only if $g \in \ell^1(\mathbb{Z})$, which comes from the fact that $\ell^1(\mathbb{Z})$ is dual to $\ell^\infty(\mathbb{Z})$.

Now suppose we have a WSS process $X = (X_n)_{n\in\mathbb{Z}}$ and run it through g. It would look like this:



If g is stable then (X, Y) are JWSS.

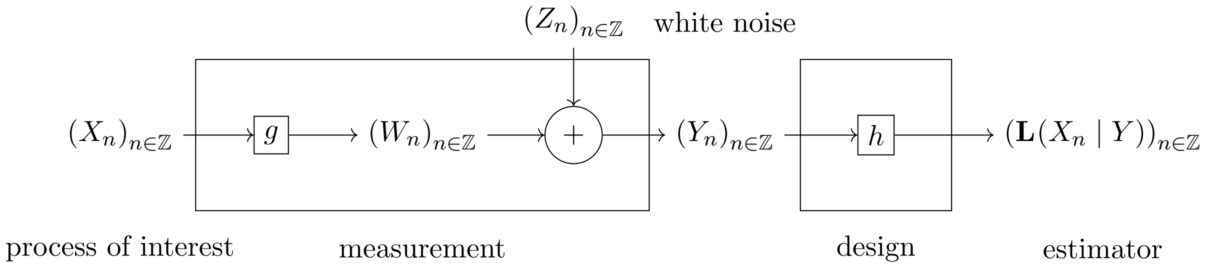
Theorem 5.6.1. Let $X = (X_n)_{n\in\mathbb{Z}}$ be a zero-mean WSS process and $g \in \ell^1(\mathbb{Z})$ be an impulse response of a stable LTI system. There exists a zero-mean WSS process $Y = (Y_n)_{n\in\mathbb{Z}}$, jointly WSS with X, satisfying

$$Y_n = \sum_{k\in\mathbb{Z}} g(n-k)X_k \quad \text{a.s. and in } L^2 \text{ for all } n \in \mathbb{Z}. \tag{5.6.1}$$

Moreover, if $R_{XX} \in \ell^2(\mathbb{Z})$ then $R_{YY} \in \ell^2(\mathbb{Z})$ and we have

$$S_{YY}(\omega) = |G(\omega)|^2\, S_{XX}(\omega), \quad S_{YX}(\omega) = G(\omega)S_{XX}(\omega). \tag{5.6.2}$$

Example 5.6.2. Many times we run into the following setup.



If $R_{XX} \in \ell^2(\mathbb{Z})$ with ess sup$(S_{XX}) < \infty$, then

$$W_n = \sum_{k\in\mathbb{Z}} g(n-k)X_k \tag{5.6.3}$$

exists as a process, and $Y_n = W_n + Z_n$ has regular covariance:

$$S_{YY}(\omega) = S_{WW}(\omega) + S_{ZZ}(\omega) = |G(\omega)|^2\, S_{XX}(\omega) + \sigma_Z^2. \tag{5.6.4}$$

The optimal filter is given by

$$\begin{aligned} H(\omega) &= \frac{S_{YX}(\omega)^*}{S_{YY}(\omega)} \qquad (5.6.5)\\ &= \frac{G^*(\omega) S_{XX}(\omega)}{|G(\omega)|^2 S_{XX}(\omega) + \sigma_Z^2} \qquad (5.6.6)\\ &= \frac{1}{G(\omega) + \frac{\sigma_Z^2}{S_{XX}(\omega) G(\omega)^*}}. \qquad (5.6.7) \end{aligned}$$

Here the term $\frac{\sigma_Z^2}{S_{XX}(\omega)G(\omega)^*}$ is a regularizer.

In practice, we model X as white noise, so this simplifies further to be

$$H(\omega) = \frac{1}{G(\omega) + \frac{\sigma_Z^2}{\sigma_X^2 G(\omega)^*}} = \frac{G(\omega)^*}{|G(\omega)|^2 + \frac{\sigma_Z^2}{\sigma_X^2}}. \tag{5.6.8}$$

6 Discrete-Time Markov Chains

We established limit theorems for i.i.d. stochastic processes. We now turn to a new class of stochastic processes known as Markov chains, which incorporate a simple model of memory. As we will see, this is a rich class of processes capable of modeling many situations of practical interest, yet it is sufficiently structured to permit characterization of long-term behavior.

6.1 Definitions

Roughly speaking, a discrete-time Markov chain $(X_n)_{n\in\mathbb{N}_0}$ is a stochastic process whose future depends on the past only through the present. The random variable X_n is called the **state** of the process at time n. We will work now with countable state spaces, which means that X_n takes values in a countable set $\mathcal{S}$. Unless specified otherwise, we generally take $\mathcal{S} = \mathbb{N}_0$.

Definition 6.1.1 (Discrete-Time Markov Chain). A **discrete-time Markov chain** is a process $(X_n)_{n\in\mathbb{N}_0}$ satisfying

$$\mathbf{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbf{P}(X_{n+1} = j \mid X_n = i) \tag{6.1.1}$$

for all $n \geq 1$ and $i, i_0, \ldots, i_{n-1}, j \in \mathcal{S}$.

Definition 6.1.2. A Markov chain is **temporally homogeneous** if there are numbers $(p_{ij})_{i,j\in\mathcal{S}}$ not depending on n such that

$$\mathbf{P}(X_{n+1} = j \mid X_n = i) = p_{ij} \tag{6.1.2}$$

for all $n \geq 0$ and all states $i, j \in \mathcal{S}$. The numbers $(p_{ij})_{i,j\in\mathcal{S}}$ are generically referred to as the **transition probabilities** of the Markov chain.

Example 6.1.3. Let $(X_n)_{n\in\mathbb{N}_0}$ be i.i.d. discrete random variables. The process $(X_n)_{n\in\mathbb{N}_0}$ is a temporally homogeneous Markov chain. Define $S_n = \frac{1}{n+1}\sum_{k=0}^{n} X_k$ for all $n \geq 0$. The process $(S_n)_{n\in\mathbb{N}_0}$ is a Markov chain, but it is not temporally homogeneous.

By virtue of their definition, the transition probabilities must satisfy

$$p_{ij} \geq 0 \quad \text{for each } i, j \in \mathcal{S} \quad , \quad \sum_{j\in\mathcal{S}} p_{ij} = 1 \quad \text{for each } i \in \mathcal{S}. \tag{6.1.3}$$

For convenience, these probabilities are collected into a square matrix (of possibly infinite dimensions) as

$$P = \left[p_{ij}\right]_{i,j\in\mathcal{S}}. \tag{6.1.4}$$

The matrix P is the **transition matrix**. The transition matrix is a *stochastic matrix*; i.e., it is a square matrix with non-negative entries, whose rows sum to one. Transition probabilities for a Markov chain always define a stochastic matrix, and vice versa.

Example 6.1.4. Markov chains as we have defined only have a "working memory" of one timestep. Here's how we extend that. Let k be a finite fixed integer, and X_{n+1} is conditionally independent of the past given $(X_n, X_{n-1}, \ldots, X_{n-k})$. The process $(X_n)_{n\in\mathbb{N}_0}$ is not a Markov chain, but we can define a new process $(Y_n)_{\substack{n\in\mathbb{N}_0\\ n\geq k}}$ with $Y_n = (Y_n, \ldots, Y_{n-k})$. The process $(Y_n)_{\substack{n\in\mathbb{N}_0\\ n\geq k}}$ is a Markov chain.

6.2 Chapman-Kolmogorov Equations

A temporally homogeneous Markov chain is defined by its transition probabilities of transitioning from a state i to a state j in one step. What can we say about the probabilities of transitioning from i to j after n steps? To start, define the **multi-step transition probabilities**

$$P_{ij}^n := \mathbf{P}(X_{n+m} = j \mid X_m = i) \quad \text{for all } m, n \in \mathbb{N}_{\geq 0}. \tag{6.2.1}$$

Note that the super-script nh ere denotes an n step transition probability. In particular, $P_{ij}^n \neq (P_{ij})^n$ in general.

The **Chapman-Kolmogorov Equations** give a recursive formula for computing the n-step transition probabilities.

Proposition 6.2.1. For all $m, n \in \mathbb{N}_0$ and states $i, j \in \mathcal{S}$,

$$P_{ij}^{m+n} = \sum_{k \in \mathcal{S}} P_{ik}^m P_{kj}^n. \tag{6.2.2}$$

In particular, we have

$$P^n = (P)^n = \left[P_{ij}^n\right]_{i,j \in \mathcal{S}}. \tag{6.2.3}$$

Note that here P is the one-step transition probability matrix and P^n is the n-step transition probability matrix.

Proof. It follows by the law of total probability:

$$\begin{aligned}
P_{ij}^{m+n} &= \mathbf{P}(X_{m+n} = j \mid X_0 = i) && (6.2.4)\\
&= \sum_{k \in \mathbb{Z}} \mathbf{P}(X_{m+n} = j, X_m = k \mid X_0 = i) && (6.2.5)\\
&= \sum_{k \in \mathbb{Z}} \mathbf{P}(X_{m+n} = j \mid X_m = k)\, \mathbf{P}(X_m = k \mid X_0 = i) && (6.2.6)\\
&= \sum_{k \in \mathbb{Z}} \mathbf{P}(X_n = j \mid X_0 = k)\, \mathbf{P}(X_m = k \mid X_0 = i) && (6.2.7)\\
&= \sum_{k \in \mathbb{Z}} P_{ik}^m P_{kj}^n. && (6.2.8)
\end{aligned}$$

From here the matrix equality follows. □

6.3 Classification of States

Definition 6.3.1 (Accessibility). We say state j is **accessible** from state i if there exists $n \geq 0$ such that $P_{ij}^n > 0$. This is denoted by $i \to j$.

Definition 6.3.2 (Communication). If i, j are states and $i \to j$ and $j \to i$, we say i and j **communicate**, and write $i \leftrightarrow j$.

Note that $i \leftrightarrow i$, if $i \leftrightarrow j$ then $j \leftrightarrow i$, and if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.

Proposition 6.3.3. The binary function $\cdot \leftrightarrow \cdot$ is an equivalence relation on $\mathcal{S}$.

Definition 6.3.4 (Classes of States). The equivalence classes of states under $\leftrightarrow$ are called **classes**.

To be more explicit, a class of states is a nonempty set of states $C \subseteq S$ such that if $i \in C$ and $j \in C$, then $i \leftrightarrow j$; if $i \in C$ and $j \notin C$, then $i \not\leftrightarrow j$.

Definition 6.3.5 (Irreducibility). If the Markov chain has only one class, then it is **irreducible**.

The reason we care about irreducibility is that we can compose any number of Markov chains into one gigantic chain with disjoint classes, which can have different behavior depending on the class (sub-chain) you start in.

Many important properties of states are *class properties.* That is, if one state in the class has this property, then all states in this class have this property. One example of this is defined now.

Definition 6.3.6 (Periodicity). For a state i, define its **period** $d(i)$ to be the greatest common divisor of the set $\{n \in \mathbb{N} \colon P_{ii}^n > 0\}$. In other words, all round trips from i to i are integer multiples of the period $d(i)$, and moreover $d(i)$ is the largest such integer. States with period 1 are called **aperiodic**. If there is *no* n such that $P_{ii}^n > 0$, we say that the period $d(i) = +\infty$.

Proposition 6.3.7. Periodicity is a class property. That is, if $i \leftrightarrow j$, then $d(i) = d(j)$.

Proof. Suppose $i \leftrightarrow j$. Using the fact that $i \leftrightarrow j$, pick $k, m \in \mathbb{N}$ such that $P_{ij}^k > 0$ and $P_{ji}^m > 0$. By Chapman-Kolmogorov equations, this implies

$$P_{ii}^{k+m} = \sum_{s \in \mathcal{S}} P_{is}^k P_{si}^m = \underbrace{P_{ij}^k P_{ji}^m}_{>0} + \sum_{\substack{s \in \mathcal{S} \\ s \neq j}} \underbrace{P_{is}^k P_{si}^m}_{\geq 0} > 0. \tag{6.3.1}$$

This implies $d(i)$ divides $k + m$.

Now choose ℓ such that $P_{jj}^\ell > 0$, which implies

$$P_{ii}^{k+m+\ell} = \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{S}} P_{is}^k P_{st}^\ell P_{ti}^m \geq \sum_{s \in \mathcal{S}} P_{is}^k P_{ss}^\ell P_{si}^m \geq P_{ij}^k P_{jj}^\ell P_{ji}^m > 0. \tag{6.3.2}$$

Thus $d(i)$ divides ℓ. Since ℓ is arbitrary, ℓ is a common divisor of the set $\left\{\ell \in \mathbb{N} \colon P_{jj}^\ell > 0\right\} = d(j)$. Hence $d(i) \leq d(j)$. By symmetry $d(j) \leq d(i)$ and hence $d(i) = d(j)$ as desired. □

Often, periodicity is a property that we exclude by hypothesis. It being a class property makes it easier. Indeed, if one state in a class is aperiodic, then the entire class will be aperiodic. In particular, if a Markov chain is irreducible then periodicity being a class property allows us to infer the periodicity of all states based on the periodicity of a single state. Note that any (positive probability) self-loop within a class renders the entire class aperiodic.

Definition 6.3.8 (Transient, Recurrent). Define the **hitting time**

$$T_j := \inf \{n \in \mathbb{N} \colon X_n = j\}. \tag{6.3.3}$$

Define also the **first passage probability**

$$F_{ij}(\infty) := \mathbf{P}(T_j < \infty \mid X_0 = i). \tag{6.3.4}$$

A state is **transient** if $F_{ii}(\infty) < 1$. A state **recurrent** if $F_{ii}(\infty) = 1$.

If j is recurrent and we start and we start the chain in state $X_0 = j$, then the resulting process will revisit j infinitely often, with probability one. On the other hand, if j is transient and we start the chain in state $X_0 = j$, then the resulting process will revisit j only finitely many times, with probability one. Since the process "resets" from a probabilistic point of view each time it re-enters state j (due to the Markov property), the number of visits to state j will be a geometric random variable with success probability $F_{jj}(\infty)$.

Proposition 6.3.9. Recurrence and transience are class properties. That is, if $i \leftrightarrow j$ and i is recurrent, then j is also recurrent. Conversely, if $i \leftrightarrow j$ and i is transient, then j is also transient.

Before proving the proposition, we establish a useful lemma that characterizes recurrence in terms of n-step transition probabilities.

Lemma 6.3.10. State i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^n = \infty$.

Proof. For the chain starting in $X_0 = i$, let N be a random variable equal to the total number of re-entries into state i. In particular, i is recurrent if and only if $\mathbf{E}(N \mid X_0 = i) = \infty$. Then

$$\mathbf{E}(N \mid X_0 = i) = \mathbf{E}\left(\sum_{n\in\mathbb{N}} 1_{X_n=i} \,\middle|\, X_0 = i\right) = \sum_{n\in\mathbb{N}} \mathbf{E}(1_{X_n=i} \mid X_0 = i) = \sum_{n\in\mathbb{N}} P_{ii}^n. \tag{6.3.5}$$

The monotone convergence theorem justifies the exchange of the sum and the expectation. □

Proof. Suppose $i \leftrightarrow j$, and let m, n be such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. For any $s \geq 1$, we have

$$P_{jj}^{m+n+s} = \sum_{\ell_1,\ell_2\in\mathcal{S}} P_{j\ell_1}^m P_{\ell_1\ell_2}^s P_{\ell_2 j}^n \geq P_{ji}^m P_{ii}^s P_{ij}^n. \tag{6.3.6}$$

Hence

$$\sum_{s\in\mathbb{N}} P_{jj}^{m+n+s} \geq P_{ij}^m P_{ij}^n \sum_{s\in\mathbb{N}} P_{ii}^s. \tag{6.3.7}$$

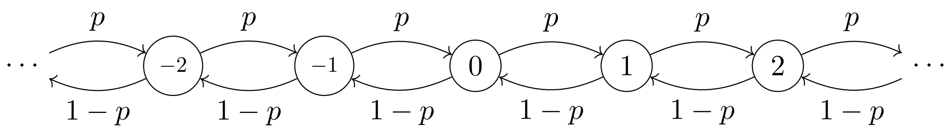
Hence by the above lemma i is recurrent if and only if j is also recurrent, and i is transient if and only if j is transient. □

If i, j are in the same recurrent class, then we will transition from state i to j in finite time with probability one. This implies the choice of starting state within a recurrent class is inconsequential in the long run.

Corollary 6.3.11. If $i \leftrightarrow j$ and j is recurrent, then $F_{ij}(\infty) = 1$.

Proof. If j is recurrent, then so is i. Then there exists $n \in \mathbb{N}$ such that $P_{ij}^n > 0$. Since i is recurrent, it will be visited infinitely often, assuming $X_0 = i$. But each time we enter i, we will enter j after n steps with positive probability P_{ij}^n. Hence, with probability 1, we need to return to i a geometric number of times until a subsequent tour passes through j. □

Example 6.3.12. Consider a random walk on the integers which moves right with probability p and left with probability $1-p$.



This Markov chain has only one class, and each state has period 2. By symmetry it suffices to look at state 0. We note that $P_{00}^{2n+1} = 0$ for $n \in \mathbb{N}_0$ due to periodicity. Moreover, every path starting and ending at state 0 must have equal numbers of left and right steps, we have

$$P_{00}^{2n} = \binom{2n}{n} p^n (1-p)^n \sim \frac{(4p(1-p))^n}{\sqrt{\pi n}} \tag{6.3.8}$$

by Stirling's formula. As a result,

$$\sum_{n\in\mathbb{N}} P_{00}^{2n} = \begin{cases} \infty & p = \frac{1}{2} \\ < \infty & p \neq \frac{1}{2} \end{cases}. \tag{6.3.9}$$

Hence all states are recurrent if $p = \frac{1}{2}$, and are transient otherwise.

6.4 Markov Limit Theorems

Define $N_j(n)$, for $n \in \mathbb{N}$, be the number of transitions into state j, up to and including time n. More precisely,

$$N_j(n) := |\{1 \le k \le n \colon X_k = j\}|\,. \tag{6.4.1}$$

Then the quantity $\frac{N_j(n)}{n}$ is the fraction of time that the process has spent in state j, from time 1 through n. Intuitively, the fraction of time spent in state j should be inversely proportional to the average number of steps it takes the process to go from state j back to j, defined as

$$\mu_{jj} := \mathbf{E}(T_j \mid X_0 = j). \tag{6.4.2}$$

Theorem 6.4.1. Let $(X_n)_{n\in\mathbb{N}_0}$ be a Markov chain starting in state i. If $i \leftrightarrow j$, then

$$\lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{\mu_{jj}} \quad \mathbf{P}\text{-a.s.} \tag{6.4.3}$$

Proof. If j is transient, then $\frac{1}{\mu_{jj}} = 0$ by definition. Moreover, $\lim_{n\to\infty} N_j(n) < \infty$ almost surely, so that $\frac{N_j(n)}{n} \to 0$ almost surely, which is what we want.

Now suppose that j is recurrent. We can assume without loss of generality that $X_0 = j$. This is because if the process starts in state $X_0 = i$, then we will enter state j after a finite amount of time almost surely. This "start-up" negligibly affects the notation $\frac{N_j(n)}{n}$ as $n \to \infty$. Since j is recurrent, it will be revisited infinitely often almost surely, so

$$\lim_{n\to\infty} N_j(n) = \infty, \quad \mathbf{P}\text{-a.s.} \tag{6.4.4}$$

Let R_k be the travel time of the k^{th} round trip from $j \to j$. Since Markov chains forget the past, $(R_k)_{k\in\mathbb{N}}$ is a sequence of non-negative i.i.d. random variables with $\mathbf{E}(R_n) = \mu_{jj}$. Define

$$S_{N_j(n)} := \sum_{k=1}^{N_j(n)} R_k. \tag{6.4.5}$$

For any $M > 0$, the strong law of large numbers yields

$$\lim_{n\to\infty} \frac{1}{N_j(n)} \sum_{k=1}^{N_j(n)} (R_k \wedge M) = \mathbf{E}(R_1 \wedge M) \quad \mathbf{P}\text{-a.s.} \tag{6.4.6}$$

By monotone convergence, we let $M \to +\infty$, and conclude

$$\lim_{n\to\infty} \frac{S_{N_j(n)}}{N_j(n)} = \mu_{jj} \quad \mathbf{P}\text{-a.s.} \tag{6.4.7}$$

Note that $S_{N_j(n)}$ is the total time spent in round trips up to time n, and therefore $S_{N_j(n)} \le n$. Similarly, $S_{N_j(n)+1} \ge n$, since it includes the duration of the trip currently underway at time n. Using these estimates, we sandwich

$$\frac{S_{N_j(n)}}{N_j(n)} \le \frac{n}{N_j(n)} \le \frac{S_{N_j(n)+1}}{N_j(n)} = \left(\frac{S_{N_j(n)+1}}{N_j(n)+1}\right)\left(\frac{N_j(n)+1}{N_j(n)}\right). \tag{6.4.8}$$

Taken together, the previous observations show that the upper and lower bounds each approach μ_{jj} almost surely. □

Corollary 6.4.2. If $(X_n)_{n\in\mathbb{N}}$ is an irreducible Markov chain, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mu_{jj}} \quad \text{for all } j \in \mathcal{S}. \tag{6.4.9}$$

Proof. Write

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n} P_{ij}^{k} &= \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}\big(1_{\{X_k=j\}} \mid X_0=i\big) \tag{6.4.10}\\
&= \mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n} 1_{\{X_k=j\}} \,\middle|\, X_0=i\right) \tag{6.4.11}\\
&= \mathbf{E}\left(\underbrace{\frac{N_j(n)}{n}}_{\leq 1} \,\middle|\, X_0=i\right). \tag{6.4.12}\\
\implies \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P_{ij}^{k} &= \lim_{n\to\infty}\mathbf{E}\left(\underbrace{\frac{N_j(n)}{n}}_{\leq 1} \,\middle|\, X_0=i\right) \tag{6.4.13}\\
&= \mathbf{E}\left(\lim_{n\to\infty}\underbrace{\frac{N_j(n)}{n}}_{\leq 1} \,\middle|\, X_0=i\right) \tag{6.4.14}\\
&= \mu_{jj}. \tag{6.4.15}
\end{aligned}$$

□

Definition 6.4.3 (Stationary Distributions). A probability distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ is a stationary distribution for a Markov chain with transition probability P if $\pi_j = \sum_{i\in\mathcal{S}} \pi_i P_{ij}$.

Note that if $X_0 \sim \pi$, then

$$\begin{aligned}
\pi_j &= \mathbf{P}(X_1=j) \tag{6.4.16}\\
&= \sum_{i\in\mathcal{S}} \mathbf{P}(X_1=j \mid X_0=i)\,\mathbf{P}(X_0=i) \tag{6.4.17}\\
&= \sum_{i\in\mathcal{S}} P_{ij}\pi_i. \tag{6.4.18}
\end{aligned}$$

By induction, $X_n \sim \pi$ for $n \in \mathbb{N}$.

Definition 6.4.4 (Types of Recurrency). A recurrent j is

1. **null recurrent** if $\mu_{jj} = +\infty$;
2. **positive recurrent** if $\mu_{jj} < \infty$.

In the context of SLLN for Markov chains, it distinguishes between states where we spend zero proportion of time in it in the long run, and states where we spend nonzero proportion of time in it in the long run.

Proposition 6.4.5. Positive and null recurrence are class properties.

Remark 6.4.6. Stationary distributions may not be unique (take $P = I$) and may not exist.

Theorem 6.4.7 (Existence and Uniqueness of Stationary Distributions). For any irreducible Markov chain $(X_n)_{n\in\mathbb{N}_0}$, exactly one of the following is true.

1. All states are transient or null recurrent. In this case, $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P_{ij}^{k} = 0$ almost surely for all states $i, j \in \mathcal{S}$, and no stationary distribution exists.

2. All states are positive recurrent. In this case, a unique stationary distribution exists and satisfies

$$\pi_j = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mu_{jj}} \tag{6.4.19}$$

for every $j \in \mathcal{S}$.

Proof. Since transience and positive/null recurrence are class properties, it follows that an irreducible Markov chain has all states with the same property of transience, positive recurrence, or null recurrence.

Case 1. Suppose all states are transient or null recurrent. We have $\mu_{jj} = 0$ by definition, and by Corollary 6.4.2 we have $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P_{ij}^k = 0$ almost surely. We argue by contradiction that a stationary distribution does not exist. To this end, suppose π is a stationary distribution. Fix $j \in \mathcal{S}$. Then

$$\pi_j = \sum_{i\in\mathcal{S}} \pi_i P_{ij}^k \quad \text{for any } k \in \mathbb{N}. \tag{6.4.20}$$

Thus averaging over $k \in [n]$,

$$\pi_j = \frac{1}{n}\sum_{k=1}^{n} \pi_j = \frac{1}{n}\sum_{k=1}^{n}\sum_{i\in\mathcal{S}} \pi_i P_{ij}^k = \sum_{i\in\mathcal{S}} \pi_i \left(\frac{1}{n}\sum_{k=1}^{n} P_{ij}^k\right) \quad \text{for any } n \in \mathbb{N}. \tag{6.4.21}$$

By Dominated Convergence Theorem, $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P_{ij}^k = 0$ almost surely, so

$$\pi_j = \lim_{n\to\infty} \sum_{i\in\mathcal{S}} \pi_i \left(\frac{1}{n}\sum_{k=1}^{n} P_{ij}^k\right) = \sum_{i\in\mathcal{S}} \pi_i \left(\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P_{ij}^k\right) = \sum_{i\in\mathcal{S}} \pi_i \cdot 0 = 0. \tag{6.4.22}$$

This is a contradiction and thus there is no stationary distribution π.

Case 2. Suppose all states are positive recurrent. Fix a labeling $\sigma\colon \mathbb{N}_0 \to \mathcal{S}$. For $M \geq 1$, Corollary 6.4.2 gives

$$\sum_{j=0}^{M} \frac{1}{\mu_{\sigma(j)\sigma(j)}} = \sum_{j=0}^{M}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P_{\sigma(i)\sigma(j)}^k\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\sum_{j=0}^{M} P_{\sigma(i)\sigma(j)}^k \leq 1. \tag{6.4.23}$$

Letting $M \to \infty$, we conclude

$$\sum_{j\in\mathbb{N}_0} \frac{1}{\mu_{\sigma(j)\sigma(j)}} = \sum_{j\in\mathcal{S}} \frac{1}{\mu_{jj}} \leq 1. \tag{6.4.24}$$

The Chapman-Kolmogorov equations give for each $M, N \geq 1$,

$$\frac{1}{N}\sum_{n=1}^{N} P_{ij}^{n+1} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k\in\mathcal{S}} P_{kj}P_{ik}^n \geq \sum_{k=0}^{M} P_{\sigma(k)j}\left(\frac{1}{N}\sum_{n=1}^{N} P_{i\sigma(k)}P_{ik}^n\right). \tag{6.4.25}$$

Take $N \to \infty$ and apply Corollary 6.4.2, and then take $M \to \infty$ to conclude

$$\frac{1}{\mu_{jj}} \geq \sum_{k\in\mathcal{S}} P_{kj}\frac{1}{\mu_{kk}}. \tag{6.4.26}$$

Suppose this inequality was strict for some $j \in \mathcal{S}$, then

$$1 \leq \sum_{j\in\mathcal{S}} \frac{1}{\mu_{jj}} > \sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{S}} P_{kj}\frac{1}{\mu_{kk}} = \sum_{k\in\mathcal{S}} \frac{1}{\mu_{kk}}\sum_{j\in\mathcal{S}} P_{kj} = \sum_{k\in\mathcal{S}} \frac{1}{\mu_{kk}}. \tag{6.4.27}$$

This is a contradiction, so we have equality:

$$\frac{1}{\mu_{jj}} = \sum_{k\in\mathcal{S}} P_{kj}\frac{1}{\mu_{kk}}. \tag{6.4.28}$$

Thus taking $\pi_j = \frac{1/\mu_{jj}}{\sum_{k\in\mathcal{S}} 1/\mu_{kk}}$ is a stationary distribution.

Having shown existence, we will show uniqueness, which will finish the proof. To this end, suppose η is another stationary distribution. As such,

$$\eta_j = \sum_{i\in\mathcal{S}} \eta_i P_{ij}^n = \sum_{i\in\mathcal{S}} \eta_i \left(\frac{1}{N} \sum_{n=1}^{N} P_{ij}^n \right). \tag{6.4.29}$$

By Dominated Convergence Theorem, we can take the limit as $N \to \infty$ to conclude

$$\eta_j = \sum_{i\in\mathcal{S}} \eta_i \frac{1}{\mu_{jj}} = \frac{1}{\mu_{jj}} \tag{6.4.30}$$

which finishes the proof.

□

Example 6.4.8 (Ergodic Theorem for Markov Chains). Consider an irreducible Markov chain $(X_n)_{n\in\mathbb{N}_0}$ with stationary distribution π. Let $r\colon \mathcal{S} \to \mathbb{R}$ be a bounded "reward" function. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r(X_k) = \sum_{j\in\mathcal{S}} \pi_j r(j) \quad \text{a.s.} \tag{6.4.31}$$

In other words,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r(X_k) = \mathbf{E}_{X\sim\pi}(r(X)) \quad \text{a.s.} \tag{6.4.32}$$

This is like a version of Strong Law of Large Numbers. But $(r(X_n))_{n\in\mathbb{N}_0}$ is not an i.i.d. process – in fact, it may not even be Markov!

This is true because

$$\frac{1}{n} \sum_{k=1}^{n} r(X_k) = \sum_{j\in\mathcal{S}} \frac{N_j(n)}{n} r(j) \tag{6.4.33}$$

and

$$\lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{\mu_{jj}} = \pi_j. \tag{6.4.34}$$

Example 6.4.9. Let $(X_n)_{n\in\mathbb{N}_0}$ be a random walk on a finite graph $G = (V, E)$. Let

$$P_{ij} = \begin{cases} \frac{1}{\deg(i)} & (i,j) \in E \\ 0 & (i,j) \notin E \end{cases}. \tag{6.4.35}$$

We claim $\pi_i = \frac{\deg(i)}{2|E|}$ is a stationary distribution. Indeed,

$$\sum_{i\in\mathcal{S}} P_{ij}\pi_i = \sum_{i\in\mathcal{S}} \frac{\deg(i)}{2\,|E|} \begin{cases} \frac{1}{\deg(i)} & (i,j) \in E \\ 0 & (i,j) \notin E \end{cases} \tag{6.4.36}$$

$$= \sum_{\substack{i\in\mathcal{S} \\ (i,j)\in E}} \frac{\deg(i)}{2\,|E|} \cdot \frac{1}{\deg(i)} \tag{6.4.37}$$

$$= \sum_{\substack{i\in\mathcal{S} \\ (i,j)\in E}} \frac{1}{2\,|E|} \tag{6.4.38}$$

$$= \frac{\deg(j)}{2\,|E|}. \tag{6.4.39}$$

Thus the expected number of steps until we return to state j is $\frac{2|E|}{\deg(j)}$.

Theorem 6.4.10. Let $(X_n)_{n\in\mathbb{N}_0}$ be irreducible, aperiodic, positive recurrent, and having stationary distribution π. Then

$$\lim_{n\to\infty}\sum_{j\in\mathcal{S}}\left|P_{ij}^n-\pi_j\right|=0. \tag{6.4.40}$$

In particular, $P_{ij}^n\to\pi_j$ for all starting states $i\in\mathcal{S}$.

Remark 6.4.11. If the Markov chain is periodic, a version still holds:

$$\lim_{n\to\infty}\sum_{j\in\mathcal{S}}\left|P_{ij}^{dn}-d\pi_j\right|=0. \tag{6.4.41}$$

for all starting states $i\in\mathcal{S}$.

Remark 6.4.12. If the Markov chain is null recurrent or transient, we may show $\lim_{n\to\infty}P_{ij}^n=0$ for all $i,j\in\mathcal{S}$, even without the assumption of aperiodicity.

Proof of Theorem 6.4.10. The proof introduces the idea of coupling two independent Markov chains together, which may sound arbitrary at first, but ultimately turns out to be a very useful idea in the analysis of convergence of Markov chains. To this end, let $(X_n)_{n\in\mathbb{N}_0}$ be an irreducible, aperiodic, and positive recurrent Markov chain, and let $(Y_n)_{n\in\mathbb{N}_0}$ be an independent Markov chain with the same transition probabilities. That is, $(X_n,Y_n)_{n\in\mathbb{N}_0}$ is a Markov chain with state space $\mathcal{S}\times\mathcal{S}$ and transition probabilities

$$\begin{aligned}
P_{(i,i')(j,j')}&=\mathbf{P}\left(X_1=j,Y_1=j'\;\middle|\;X_0=i,Y_0=i'\right) &\text{(6.4.42)}\\
&=\mathbf{P}\left(X_1=j\mid X_0=i\right)\mathbf{P}\left(Y_1=j\mid Y_0=i\right) &\text{(6.4.43)}\\
&=P_{ij}P_{i'j'}. &\text{(6.4.44)}
\end{aligned}$$

The first thing to do is check that $(X_n,Y_n)_{n\in\mathbb{N}_0}$ is irreducible; this is where the assumption of aperiodicity is needed. To this end, we claim that for each j, there is an n_0 such that $P_{jj}^n>0$ for all $n\geq n_0$. Using irreducibility of the individual chains, there are K,L such that $P_{ij}^K>0$ and $P_{i'j'}^L>0$. It follows that there is a sufficiently large M such that

$$\begin{aligned}
P_{(i,i')(j,j')}&=\mathbf{P}\left(X_{K+L+M}=j,Y_{K+L+M}=j'\;\middle|\;X_0=i,Y_0=i'\right) &\text{(6.4.45)}\\
&=\mathbf{P}\left(X_{K+L+M}=j\mid X_0=i\right)\mathbf{P}\left(Y_{K+L+M}=j'\;\middle|\;Y_0=i'\right) &\text{(6.4.46)}\\
&=P_{ij}^{K+L+M}P_{i'j'}^{K+L+M} &\text{(6.4.47)}\\
&\geq P_{ij}^KP_{jj}^{L+M}P_{i'j'}^LP_{j'j'}^{K+M} &\text{(6.4.48)}\\
&>0 &\text{(6.4.49)}
\end{aligned}$$

which establishes that $(i,i')\to(j,j')$. Since i,j,i',j' are arbitrary, all states communicate, and hence $(X_n,Y_n)_{n\in\mathbb{N}_0}$ is irreducible. It follows from definitions that $Q_{jj'}=\pi_j\pi_{j'}$ is a stationary distribution for $(X_n,Y_n)_{n\in\mathbb{N}_0}$, so by Existence and Uniqueness of Stationary Distributions all states are recurrent.

Now, for any choice of state j, define the random variables

$$T=\inf\left\{n\in\mathbb{N}_0\colon X_n=Y_n\right\},\quad T_j=\inf\left\{n\in\mathbb{N}_0\colon X_n=j,Y_n=j\right\}. \tag{6.4.50}$$

Since our process is recurrent, it holds by definition that $\mathbf{P}(T_j<\infty)=1$. Since $T\leq T_j$, it follows that $\mathbf{P}(T<\infty)=1$.

We prove the fact that, conditioned on $\{T\leq n\}$, the random variables X_n and Y_n are equal in distribution. This might sound surprising, but by Markovity, once the processes enter the same state, they are statistically indistinguishable from that point on. More precisely,

$$\mathbf{P}\left(X_n=j,T\leq n\right)=\sum_{m=0}^{n}\sum_{i\in\mathcal{S}}\mathbf{P}\left(X_n=j,T=m,X_m=i\right) \tag{6.4.51}$$

$$= \sum_{m=0}^{n} \sum_{i \in \mathcal{S}} \mathbf{P}(T = m, X_m = i)\, \mathbf{P}(X_n = j \mid X_m = i) \tag{6.4.52}$$

$$= \sum_{m=0}^{n} \sum_{i \in \mathcal{S}} \mathbf{P}(T = m, Y_m = i)\, \mathbf{P}(Y_n = j \mid Y_m = i) \tag{6.4.53}$$

$$= \sum_{m=0}^{n} \sum_{i \in \mathcal{S}} \mathbf{P}(Y_n = j, T = m, Y_m = i) \tag{6.4.54}$$

$$= \mathbf{P}(Y_n = j, T \leq n). \tag{6.4.55}$$

Now we use this to obtain

$$\begin{aligned} \mathbf{P}(X_n = j) &= \mathbf{P}(X_n = j, T \leq n) + \mathbf{P}(X_n = j, T > n) && (6.4.56)\\ &= \mathbf{P}(Y_n = j, T \leq n) + \mathbf{P}(X_n = j, T > n) && (6.4.57)\\ &\leq \mathbf{P}(Y_n = j) + \mathbf{P}(X_n = j, T > n). && (6.4.58) \end{aligned}$$

By symmetry,

$$\mathbf{P}(Y_n = j) \leq \mathbf{P}(X_n = j) + \mathbf{P}(Y_n = j, T > n), \tag{6.4.59}$$

so it holds that

$$|\mathbf{P}(X_n = j) - \mathbf{P}(Y_n = j)| \leq \mathbf{P}(X_n = j, T > n) + \mathbf{P}(Y_n = j, T > n). \tag{6.4.60}$$

Summing over $j \in \mathcal{S}$ gives

$$\begin{aligned} \sum_{j \in \mathcal{S}} |\mathbf{P}(X_n = j) - \mathbf{P}(Y_n = j)| &\leq \sum_{j \in \mathcal{S}} (\mathbf{P}(X_n = j, T > n) + \mathbf{P}(Y_n = j, T > n)) && (6.4.61)\\ &\leq 2\,\mathbf{P}(T > n). && (6.4.62) \end{aligned}$$

Now if we let $X_0 = i$ and $Y_0 \sim \pi$, then $Y_n \sim \pi$ for all $n \in \mathbb{N}_0$ by definition of the stationary distribution. Hence, the above can be specialized to

$$\sum_{j \in \mathcal{S}} \left| P_{ij}^n - \pi_j \right| \leq 2\,\mathbf{P}(T > n). \tag{6.4.63}$$

Taking the limit $n \to \infty$ and using the fact that $\mathbf{P}(T < \infty) = 1$ completes the proof. □

Remark 6.4.13. This theorem doesn't give us quantitative rates of convergence.

6.5 Reversibility and Spectral Gap

6.5.1 Detailed Balance Equations

The class of Markov chains satisfy a property known as reversibility arise frequently in practice. Although not important for our purposes, the term *reversible* comes from the fact that if $(X_n)_{n \in \mathbb{N}_0}$ is a reversible Markov chain with stationary distribution π and $X_0 \sim \pi$, then $(X_0, X_1, \ldots, X_n) = (X_n, X_{n-1}, \ldots, X_0)$ in distribution for each $n \geq 0$. Hence, the stationary Markov chain running forward in time is indistinguishable from the process running backward in time.

Definition 6.5.1. A Markov chain with transition probabilities P and stationary distribution π is **reversible** if the transition probabilities satisfy

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \text{for all } i, j \in \mathcal{S}. \tag{6.5.1}$$

We say (P, π) is reversible.

The so-called detailed balance equations (present in the definition of reversibility) are also sufficient for reversibility:

Proposition 6.5.2 (Detailed Balance Equations). If there is a probability distribution π on $\mathcal{S}$ such that

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \text{for all } i, j \in \mathcal{S} \tag{6.5.2}$$

then the Markov chain with transition probabilities P is reversible with stationary distribution π. In this case, we say (P, π) is reversible.

Proof. We only need to check that π is a stationary distribution. To see this, sum both sides of the detailed balance equations over j to find

$$\pi_i = \sum_j \pi_i p_{ij} = \sum_j \pi_j p_{ji} \tag{6.5.3}$$

which is the definition of a stationary distribution. □

Remark 6.5.3. Attempting to solve the detailed balance equations is one of the easiest ways to find a stationary distribution, and will be successful if (and only if) the chain is reversible. This is often the case in practice, especially if they model physical systems at equilibria.

Example 6.5.4. Consider a connected, undirected graph $G = (V, E)$ with finitely many vertices, and define transition probabilities

$$p_{ij} = \begin{cases} \frac{1}{\deg(i)} & (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{6.5.4}$$

By considering the detailed balance equations, the Markov chain is reversible, with stationary distribution $\pi_i = \frac{\deg(i)}{2|E|}$ for each $i \in V$.

6.5.2 Spectral Gap and Trend to Equilibrium

Reversible Markov chains provide a convenient setting for studying rates of convergence. We begin by introducing the total variation distance between probability measures.

Definition 6.5.5. For probability measures μ, ν on a measurable space $(\Omega, \mathcal{F})$, we define their **total variation** distance

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| \,. \tag{6.5.5}$$

If Ω is countable, then viewing μ, ν as sequences,

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)| = \frac{1}{2} \|\mu - \nu\|_{\ell^1(\Omega)} \,. \tag{6.5.6}$$

If we let $P^n_{i\cdot}$ denote the distribution over states at time n given that $X_0 = i$, then the conclusion of Theorem 6.4.10 says that $P^n_{i\cdot}$ converges to π in total variation. That is,

$$\lim_{n \to \infty} \|P^n_{i\cdot} - \pi\|_{\mathrm{TV}} = 0. \tag{6.5.7}$$

Toward characterizing how quickly $\|P^n_{i\cdot} - \pi\|_{\mathrm{TV}}$ vanishes, define $\mathrm{Var}_\pi(f) = \mathrm{Var}(f(X))$ for $X \sim \pi$ and $f \colon \mathcal{S} \to \mathbb{R}$. Also, define the function $Pf \colon \mathcal{S} \to \mathbb{R}$ via the balance equation

$$(Pf)(i) = \sum_{j \in \mathcal{S}} p_{ij} f(j), \quad i \in \mathcal{S}. \tag{6.5.8}$$

Definition 6.5.6. For a reversible Markov chain (P, π), define the **spectral gap** $\gamma := 1 - \lambda_2$, where $\lambda_2 \geq 0$ is the smallest number satisfying

$$\mathrm{Var}_\pi(Pf) \leq \lambda_2 \mathrm{Var}_\pi(f) \quad \text{for all } f \colon \mathcal{S} \to \mathbb{R} \text{ with } \mathrm{Var}_\pi(f) < \infty. \tag{6.5.9}$$

The term *spectral gap* comes from the fact that λ_2 is an upper bound on the modulus of the second largest eigenvalue of P. The operator P has eigenvalue 1 since it is stochastic, so the distance between the modulus of the two largest eigenvalues is the spectral gap γ. There are some caveats here: P is a contraction so 1 is the largest eigenvalue, and we are finding the second eigenvector using $\mathrm{Var}_\pi(\cdot)$ instead of $\mathbf{E}_\pi\left(\cdot^2\right)$ in the Rayleigh quotient because 1 is an eigenvector and we want to subtract constants so as to find the largest eigenvector f that is orthogonal to 1.

Proposition 6.5.7. P is self-adjoint as a linear map $f \mapsto Pf$ on the space $L^2\left(\mathcal{S}, 2^{\mathcal{S}}, \pi\right)$.

Proof. Take $f, g \in L^2\left(\mathcal{S}, 2^{\mathcal{S}}, \pi\right)$. Then the detailed balance equations give the identity

$$\begin{aligned}\langle Pf, g\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)} &= \sum_{j\in\mathcal{S}}\left(\sum_{k\in\mathcal{S}} P_{jk} f(k)\right) g(j)\pi_j && (6.5.10)\\ &= \sum_{k\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}} P_{kj} g(j)\right) f(k)\pi_k && (6.5.11)\\ &= \langle f, Pg\rangle\,. && (6.5.12)\end{aligned}$$

Here the exchange of sums is justified by Fubini-Tonelli (and absolute convergence of the series is assured by Cauchy-Schwarz). □

By the spectral theorem, the spectrum of P is real (i.e., $\mathrm{Spec}(P) \subseteq \mathbb{R}$), and $\mathrm{Spec}(P) \subseteq [-\lambda_2, \lambda_2] \cup \{1\}$.

Theorem 6.5.8. If (P, π) is a reversible Markov chain with spectral gap γ, then

$$\|P^n_{i\cdot} - \pi\|^2_{\mathrm{TV}} \le \frac{(1-\gamma)^n}{4\pi(i)} \quad \text{for each } n \in \mathbb{N}_0, i \in \mathcal{S}. \tag{6.5.13}$$

Moreover, if the Markov chain is irreducible, aperiodic, and has finite state space, then $\gamma > 0$.

What's really happening here is that as we multiply a vector by a contractive operator, the component of the vector in the direction of the eigenvector with eigenvalue 1 is preserved, while all other components decay exponentially fast.

Definition 6.5.9. Let μ, π be probability measures on $\mathcal{S}$. We say that μ has density h with respect to π (written $\mathrm{d}\mu = h\,\mathrm{d}\pi$) if

$$\mu(j) = h(j)\pi(j) \quad \text{for all states } j. \tag{6.5.14}$$

Lemma 6.5.10. Let (P, π) be a reversible Markov chain. If $\mathrm{d}\mu = h\,\mathrm{d}\pi$, then $P^n h$ is the density of μP^n with respect to π (i.e., $\mathrm{d}(\mu P^n) = (P^n h)\,\mathrm{d}\pi$).

Proof. It suffices to prove the statement for $n = 1$, and the rest follows by induction. Hence, using reversibility, observe that

$$(\mu P)_j = \sum_{i\in\mathcal{S}} \mu(i) P_{ij} = \sum_{i\in\mathcal{S}} h(i)\pi(i) P_{ij} = \sum_{i\in\mathcal{S}} h(i)\pi(j) P_{ji} = \pi(j)(Ph)(j). \tag{6.5.15}$$

The general case goes exactly the same. □

Remark 6.5.11. The notation μP^n means the distribution at time n given that $X_0 \sim \mu$. If μ is a row vector and P is a matrix then this turns into a vector-matrix product.

Proof of Theorem 6.5.8. Suppose $\mathrm{d}\mu = h\,\mathrm{d}\pi$, with $h \in L^2\left(\mathcal{S}, 2^{\mathcal{S}}, \pi\right)$. Using the self-adjoint property of P, we have

$$\langle 1, P^n(h-1)\rangle = \langle P^n 1, h-1\rangle = \langle 1, h-1\rangle = \sum_{j\in\mathcal{S}} (\mu(j) - \pi(j)) = 0. \tag{6.5.16}$$

Thus, we can write norms of $P^n(h-1)$ in terms of variance, and by definition of spectral gap:

$$\begin{aligned}
\|P^n h - 1\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)} &= \|P^n\,(h-1)\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)} && (6.5.17)\\
&= \operatorname{Var}_\pi(P^n h) && (6.5.18)\\
&\le (1-\gamma)^n \operatorname{Var}_\pi(h) && (6.5.19)\\
&= (1-\gamma)^n \|h-1\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)}\,. && (6.5.20)
\end{aligned}$$

Jensen's inequality controls total variation distance via

$$\|\mu P^n - \pi\|^2_{\mathrm{TV}} \le \frac{1}{4}\,\|P^n h - 1\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)} \le \frac{(1-\gamma)^n}{4}\,\|h-1\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)}\,. \tag{6.5.21}$$

To finish, take $\mu = \delta_i$, so that $\mu P^n = P^n_{i\cdot}$, $h(j) = \frac{\delta_{ij}}{\pi(i)}$, and

$$\|h-1\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},\pi)} = \frac{1-\pi(i)}{\pi(i)} \le \frac{1}{\pi(i)}. \tag{6.5.22}$$

Hence

$$\|P^n_{i\cdot} - \pi\|^2_{\mathrm{TV}} \le \frac{(1-\gamma)^n}{4\pi(i)}. \tag{6.5.23}$$

□

Remark 6.5.12. If $\mathcal{S}$ is finite, then $\gamma > 0$ automatically, and we get exponentially fast convergence to π. This is a special case of the *Perron-Frobenius theorem.*

7 Martingales

7.1 Definitions and Examples

Definition 7.1.1. Let $(X_n)_{n\in\mathbb{N}_0}$ be a stochastic process. A process $(M_n)_{n\in\mathbb{N}_0}$ is **adapted to** $(X_n)_{n\in\mathbb{N}_0}$ if for all $n \geq 0$, M_n is a measurable function of $X_0, X_1, \ldots, X_n$.

Definition 7.1.2 (Martingale). The stochastic process $(M_n)_{n\in\mathbb{N}_0}$ is a **martingale** with respect to the stochastic process $(X_n)_{n\in\mathbb{N}_0}$ if

(i) $(M_n)_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$;

(ii) for each $n \in \mathbb{N}_0$, $M_n \in L^1$;

(iii) for each $n \in \mathbb{N}_0$, $\mathbf{E}(M_{n+1} \mid X_0, \ldots, X_n) = M_n$.

If the equality in (iii) is replaced by $\geq$ or $\leq$, then the process is said to be a **submartingale** or **supermartingale**, respectively.

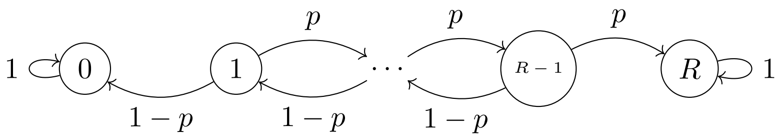
Remark 7.1.3. The requirement $M_n \in L^1$ is how we ensure that the conditional expectations exist.

Remark 7.1.4. The adaptivity is ensured by (iii) in the case of equality; the requirement is only really needed in the case of sub/supermartingales.

Remark 7.1.5. The variable X_n should be taken to represent "the information obtained at time n". Sometimes we condition on σ-algebras, i.e., M_n is a measurable function with respect to $\mathcal{F}_n$ for an sequence of σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ called a **filtration**, and $\mathbf{E}(M_{n+1} \mid \mathcal{F}_n) = M_n$. The σ-algebra $\mathcal{F}_n$ is directly the information we know at time n.

Remark 7.1.6. A martingale is both a submartingale and a supermartingale. If $(M_n)_{n\in\mathbb{N}_0}$ is a submartingale, then $(-M_n)_{n\in\mathbb{N}_0}$ is a supermartingale, and so on.

Example 7.1.7. Gambler's ruin:



is a martingale (resp. submartingale, supermartingale) with respect to itself if $p = \frac{1}{2}$ (resp. $p \geq \frac{1}{2}$, $p \leq \frac{1}{2}$).

Example 7.1.8. Let $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of independent integrable random variables. Then

$$M_n = \sum_{i=0}^{n} (X_i - \mathbf{E}(X_i)) \tag{7.1.1}$$

is a martingale.

Example 7.1.9. Suppose $(X_n)_{n\in\mathbb{N}_0}$ is independent with $\mathbf{E}(X_n) = 1$. Then

$$M_n = \prod_{i=0}^{n} X_i. \tag{7.1.2}$$

Example 7.1.10. Let $(X_n)_{n\in\mathbb{N}_0}$ be a branching process. Let X_n be the population at time n, and μ be the expected number of offspring generated at time n. We claim

$$M_n = \frac{X_n}{\mu^n} \tag{7.1.3}$$

is a martingale.

Let $Z_1, \dots, Z_{X_n}$ be the number of offspring generated at time n. Then

$$\mathbf{E}(M_{n+1} \mid X_0, \dots, X_n) = \frac{1}{\mu^{n+1}} \mathbf{E}\left(\sum_{i=1}^{X_n} Z_i \;\middle|\; X_0, \dots, X_n \right) = \frac{X_n}{\mu^n} = M_n. \tag{7.1.4}$$

Example 7.1.11 (Doob's Martingale). Let $Y \in L^1$ and $(X_n)_{n\in\mathbb{N}_0}$ be any stochastic process. Then

$$M_n = \mathbf{E}(Y \mid X_0, \dots, X_n) \tag{7.1.5}$$

is a martingale. The typical use case is $Y = f(X_0, \dots, X_N)$ for some finite large N; eventually $M_n = Y$ almost surely. If Y is a function of all the X_n's, then $\lim_{n\to\infty} M_n = Y$ almost surely.

Proposition 7.1.12. If $(M_n)_{n\in\mathbb{N}_0}$ is a sub-martingale and $m > n$, then

$$\mathbf{E}(M_m) \geq \mathbf{E}(M_n) \quad \text{and} \quad \mathbf{E}(M_m \mid X_0, \dots, X_n) \geq M_n. \tag{7.1.6}$$

Proof. Indeed,

$$\begin{aligned}
\mathbf{E}(M_m \mid X_0, \dots, X_n) &= \mathbf{E}(\mathbf{E}(M_m \mid X_0, \dots, X_{m-1}) \mid X_0, \dots, X_n) \\
&\geq \mathbf{E}(M_{m-1} \mid X_0, \dots, X_n) \\
&\geq \quad \vdots \\
&\geq \mathbf{E}(M_n \mid X_0, \dots, X_n) \\
&= M_n.
\end{aligned}$$

Then by iterated expectation,

$$\mathbf{E}(M_m) = \mathbf{E}(\mathbf{E}(M_m \mid X_0, \dots, X_n)) \geq \mathbf{E}(M_n). \tag{7.1.7}$$

□

Corollary 7.1.13. If $(M_n)_{n\in\mathbb{N}_0}$ is a super-martingale and $m > n$, then

$$\mathbf{E}(M_m) \leq \mathbf{E}(M_n) \quad \text{and} \quad \mathbf{E}(M_m \mid X_0, \dots, X_n) \leq M_n. \tag{7.1.8}$$

Proof. $(-M_n)_{n\in\mathbb{N}_0}$ is a super-martingale; use the above proposition. □

Corollary 7.1.14. If $(M_n)_{n\in\mathbb{N}_0}$ is a martingale and $m > n$, then

$$\mathbf{E}(M_m) = \mathbf{E}(M_n) \quad \text{and} \quad \mathbf{E}(M_m \mid X_0, \dots, X_n) = M_n. \tag{7.1.9}$$

Proof. It follows from the previous two results, since a martingale is both a sub-martingale and super-martingale. □

Remark 7.1.15. In theorems involving martingales, we usually prove the theorem for sub-martingales, because the consequences for super-martingales and martingales follow trivially from there.

7.2 Stopping Times

We digress from martingales for a moment and introduce a special class of random variables known as **stopping times**. Stopping times are not specific to martingales, but they do play an important role in their study.

Definition 7.2.1. A non-negative integer-valued random variables T is a **stopping time** with respect to $(X_n)_{n\in\mathbb{N}_0}$ if the event $\{T \le n\}$ is determined by $X_0, \ldots, X_n$ for all $n \in \mathbb{N}_0$. In particular, $1_{\{T\le n\}}$ is a function of $X_0, \ldots, X_n$.

Theorem 7.2.2 (Wald's Identity). Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with $\mathbf{E}(X_n) = \mu$ for all $n \in \mathbb{N}$ and $\sup_{n\in\mathbb{N}} \mathbf{E}(|X_n|) < \infty$. If $T \ge 1$ is a stopping time with $\mathbf{E}(T) < \infty$, then

$$\mathbf{E}\left(\sum_{n=1}^{T} X_n\right) = \mu\,\mathbf{E}(T). \tag{7.2.1}$$

Proof. The key observation is that $\{T < n\} = \{T \le n-1\}$ since T is integer-valued. Then $\{T \le n-1\}$ is a measurable function of $X_1, \ldots, X_{n-1}$, and thus independent of X_n. Since $\{T \ge n\} = \{T < n\}^c$, it follows that $\{T \ge n\}$ is also independent of X_n. Hence since X_n is integrable,

$$\mathbf{E}\big(X_n 1_{\{T\ge n\}}\big) = \mathbf{E}(X_n)\,\mathbf{E}\big(1_{\{T\ge n\}}\big) = \mu\,\mathbf{P}(T \ge n). \tag{7.2.2}$$

Now we can formally manipulate

$$\begin{aligned}
\mathbf{E}\left(\sum_{n=1}^{T} X_n\right) &= \mathbf{E}\left(\sum_{n\in\mathbb{N}} X_n 1_{\{T\ge n\}}\right) \\
&= \sum_{n\in\mathbb{N}} \mathbf{E}\big(X_n 1_{\{T\ge n\}}\big) \\
&= \mu \sum_{n\in\mathbb{N}} \mathbf{P}(T \ge n) \\
&= \mu\,\mathbf{E}(T),
\end{aligned}$$

where we use the identity $\mathbf{E}(T) = \sum_{n\in\mathbb{N}} \mathbf{P}(T \ge n)$ (tail sum equality).

The exchange of the sum and expectation above requires justification. To do this, fix k and write

$$\sum_{n\in\mathbb{N}} X_n 1_{\{T\ge n\}} \le \sum_{n=1}^{k} X_n 1_{\{T\ge n\}} + \sum_{n=k+1}^{\infty} |X_n|\, 1_{|T\ge n|}. \tag{7.2.3}$$

We can exchange a finite sum and expectation, we have

$$\begin{aligned}
\mathbf{E}\left(\sum_{n\in\mathbb{N}} X_n 1_{\{T\ge n\}}\right) &\le \sum_{n=1}^{k} \mathbf{E}\big(X_n 1_{\{T\ge n\}}\big) + \mathbf{E}\left(\sum_{n=k+1}^{\infty} |X_n|\, 1_{\{T\ge n\}}\right) \\
&= \sum_{n=1}^{k} \mathbf{E}\big(X_n 1_{\{T\ge n\}}\big) + \sum_{n=k+1}^{\infty} \mathbf{E}(|X_n|\, 1_{T\ge n}),
\end{aligned}$$

where the exchange of the expectation and the sum in the last line is justified by the monotone convergence theorem. Now, by independence of X_n and $\{T \ge n\}$,

$$\begin{aligned}
\sum_{n=k+1}^{\infty} \mathbf{E}\big(|X_n|\, 1_{\{T\ge n\}}\big) &= \sum_{n=k+1}^{\infty} \mathbf{E}(|X_n|)\,\mathbf{P}(T \ge n) \\
&\le \sum_{n=k+1}^{\infty} \mathbf{P}(T \ge n)
\end{aligned}$$

$$\begin{aligned}
\lim_{k\to\infty}\sum_{n=k+1}^{\infty}\mathbf{E}\big(|X_n|\,1_{\{T\geq n\}}\big) &\leq \left(\sup_{n\in\mathbb{N}}\mathbf{E}(|X_n|)\right)\lim_{k\to\infty}\sum_{n=k+1}^{\infty}\mathbf{P}(T\geq n)\\
&= \left(\sup_{n\in\mathbb{N}}\mathbf{E}(|X_n|)\right)\lim_{k\to\infty}\left(\mathbf{E}(T)-\sum_{n=1}^{k}\mathbf{P}(T\geq n)\right)\\
&= \left(\sup_{n\in\mathbb{N}}\mathbf{E}(|X_n|)\right)(\mathbf{E}(T)-\mathbf{E}(T))\\
&= \left(\sup_{n\in\mathbb{N}}\mathbf{E}(|X_n|)\right)\cdot 0\\
&= 0.
\end{aligned}$$

Thus

$$\mathbf{E}\left(\sum_{n\in\mathbb{N}}X_n 1_{\{T\geq n\}}\right)\leq\sum_{n\in\mathbb{N}}\mathbf{E}\big(X_n 1_{\{T\geq n\}}\big). \tag{7.2.4}$$

The reverse argument follows by a similar justification, and this completes the proof. □

Stopping times allow us to define *stopped processes*. The idea of a stopped process is that once the stopping time is reached, the process is frozen at that value. Recall that $a\wedge b=\min\{a,b\}$.

Definition 7.2.3. Let $(X_n)_{n\in\mathbb{N}}$ be a process, and T be a stopping time. The process $(X_{T\wedge n})_{n\in\mathbb{N}}$ is called the **stopped process**. Note that the stopped process satisfies $X_{T\wedge n}=X_n$ for $n\leq T$, and $X_{T\wedge n}=X_T$ for $n>T$.

The stopping time T above does not necessarily have to be a stopping time with respect to the process $(X_n)_{n\in\mathbb{N}}$.

7.2.1 Stopping Times and Martingales

Stopping times play a prominent role in the theory of martingales. The starting point is that a stopped martingale inherits the martingale property.

Proposition 7.2.4. If $(M_n)_{n\in\mathbb{N}_0}$ is a submartingale (resp. super-martingale, martingale) and T is a stopping time, both with respect to $(X_n)_{n\in\mathbb{N}_0}$, then the stopped process $(M_{T\wedge n})_{n\in\mathbb{N}_0}$ is also a submartingale (resp. supermartingale, martingale) with respect to $(X_n)_{n\in\mathbb{N}_0}$.

Proof. We assume that $(M_n)_{n\in\mathbb{N}_0}$ is a submartingale. This is sufficient since if $(M_n)_{n\in\mathbb{N}_0}$ is a supermartingale, then $(-M_n)_{n\in\mathbb{N}_0}$ is a submartingale, and a martingale is both a submartingale and a supermartingale.

Indeed, write

$$M_{T\wedge n}=M_n 1_{\{T\geq n\}}+M_t 1_{T\leq n}=M_n\left(1-1_{\{T\leq n\}}\right)+M_T 1_{\{T\leq n\}} \tag{7.2.5}$$

and indeed this shows that $M_{T\wedge n}$ is a function of $X_0,\ldots,X_n$, hence $(M_{T\wedge n})_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$. And

$$\begin{aligned}
\mathbf{E}(|M_{T\wedge n}|) &= \mathbf{E}(|M_{T\wedge n}|\mid T\leq n)\,\mathbf{P}(T\leq n)+\mathbf{E}(|M_{T\wedge n}|\mid T\geq n)\,\mathbf{P}(T\geq n)\\
&= \sum_{i=0}^{n-1}\mathbf{E}(|M_{T\wedge n}|\mid T=i)\,\mathbf{P}(T=i)+\mathbf{E}(|M_{T\wedge n}|\mid T\geq n)\,\mathbf{P}(T\geq n)\\
&= \sum_{n=0}^{n-1}\underbrace{\mathbf{E}(|M_i|\mid T=i)}_{<\infty}\mathbf{P}(T=i)+\underbrace{\mathbf{E}(|M_n|\mid T\geq n)}_{<\infty}\mathbf{P}(T\geq n)\\
&< \infty.
\end{aligned}$$

where the inequality holds since $\mathbf{E}(|M_n|)<\infty$ for each $n\in\mathbb{N}_0$.

Now we want to confirm the main sub-martingale inequality. We use the fact that $\{T \geq n+1\} = \{T \leq n\}^c$ is determined by $X_0, \ldots, X_n$, we have

$$\begin{aligned}
\mathbf{E}\big(M_{T\wedge(n+1)} \mid X_0, \ldots, X_n\big) &= \mathbf{E}\big(M_{n+1} 1_{\{T\geq n+1\}} \mid X_0, \ldots, X_n\big) + \mathbf{E}\big(M_{T\wedge n} 1_{\{T\leq n\}} \mid X_0, \ldots, X_n\big) \\
&= \mathbf{E}(M_{n+1} \mid X_0, \ldots, X_n) 1_{\{T\geq n+1\}} + \mathbf{E}(M_{T\wedge n} \mid X_0, \ldots, X_n) 1_{T\leq n} \\
&\geq M_n 1_{\{T\geq n+1\}} + M_{T\wedge n} 1_{\{T\leq n\}} \\
&= M_{T\wedge n}.
\end{aligned}$$

□

Proposition 7.2.5. If $(M_n)_{n\in\mathbb{N}_0}$ is a sub-martingale and T is a stopping time, both with respect to $(X_n)_{n\in\mathbb{N}_0}$, then

$$\mathbf{E}(M_0) \leq \mathbf{E}(M_{T\wedge n}) \leq \mathbf{E}(M_n), \quad n \in \mathbb{N}_0. \tag{7.2.6}$$

The inequalities are reversed for super-martingales, and met with equality for martingales.

Proof. As before, it suffices to consider only the case of a sub-martingale. Then $(M_{T\wedge n})_{n\in\mathbb{N}_0}$ is a sub-martingale, so it follows that

$$\mathbf{E}(M_0) = \mathbf{E}(M_{T\wedge 0}) \leq \mathbf{E}(M_{T\wedge n}), \quad n \in \mathbb{N}_0. \tag{7.2.7}$$

Now we show that $(M'_n)_{n\in\mathbb{N}_0}$ is a sub-martingale with respect to $(X_n)_{n\in\mathbb{N}_0}$, where $M'_n = M_n - M_{T\wedge n}$. To this end, integrability of M'_n follows from integrability of $M_0, \ldots, M_n$ and similarly $(M'_n)_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$ since T is a stopping time for $(X_n)_{n\in\mathbb{N}_0}$. now

$$\begin{aligned}
M'_{n+1} &= M_{n+1} - M_{T\wedge(n+1)} \\
&= M_{n+1} - 1_{\{T\leq n\}} M_{T\wedge n} - \big(1 - 1_{\{T\leq n\}}\big) M_{n+1} \\
&= 1_{\{T\leq n\}} M_{n+1} - 1_{T\leq n} M_{T\wedge n} \\
&= 1_{\{T\leq n\}} (M_{n+1} - M_n) - 1_{T\leq n} (M_{T\wedge n} - M_n) \\
&= 1_{\{T\leq n\}} (M_{n+1} - M_n) - (M_{T\wedge n} - M_n) \\
&= 1_{\{T\leq n\}} (M_{n+1} - M_n) - \big(-M'_n\big) \\
&= 1_{\{T\leq n\}} (M_{n+1} - M_n) + M'_n \\
\mathbf{E}(M_{n+1} \mid X_0, \ldots, X_n) &= \mathbf{E}\big(1_{\{T\leq n\}} (M_{n+1} - M_n) + M'_n \mid X_0, \ldots, X_n\big) \\
&= M'_n + 1_{\{T\leq n\}} \underbrace{\mathbf{E}(M_{n+1} - M_n \mid X_0, \ldots, X_n)}_{\geq 0} \\
&\geq M'_n.
\end{aligned}$$

Then we use monotonicity of expectation to conclude

$$\mathbf{E}(M_n) - \mathbf{E}(M_{T\wedge n}) = \mathbf{E}\big(M'_n\big) \geq \mathbf{E}\big(M'_0\big) = 0, \tag{7.2.8}$$

which completes the proof. □

Theorem 7.2.6 (Optional Stopping Theorem). Let $(M_n)_{n\in\mathbb{N}_0}$ be a sub-martingale and T be a stopping time, both with respect to $(X_n)_{n\in\mathbb{N}_0}$. If there is a constant $k < \infty$ such that any one of the following hold:

(i) $T \leq k$ a.s.; or

(ii) $|M_n| \leq k$ a.s. for each n, and $\mathbf{P}(T < \infty) = 1$; or

(iii) $\mathbf{E}(T) < \infty$ and $|M_n - M_{n-1}| \leq k$ a.s. for each $n \geq 1$, then

$$\mathbf{E}(M_0) \leq \mathbf{E}(M_T).$$

The inequality above is reversed when $(M_n)_{n\in\mathbb{N}_0}$ is a super-martingale, and an equality when $(M_n)_{n\in\mathbb{N}_0}$ is a martingale.

Proof. It suffices to prove the statement for a sub-martingale. (i) follows immediately from Proposition 7.2.5 by taking $n = k$. (ii) follows from (i) since under the assumption of $\mathbf{P}(T < \infty) = 1$, we have

$$\lim_{n\to\infty} M_{T\wedge n} = M_T \quad \text{almost surely.} \tag{7.2.9}$$

Hence, Dominated Convergence Theorem gives

$$\mathbf{E}(M_0) \leq \lim_{n\to\infty} \mathbf{E}(M_{T\wedge n}) = \mathbf{E}(M_T). \tag{7.2.10}$$

For (iii) we have

$$|M_{T\wedge n} - M_0| = \left|\sum_{k=1}^{T\wedge n} (M_k - M_{k-1})\right| \leq kT \quad \text{almost surely.} \tag{7.2.11}$$

Thus, Dominated Convergence Theorem and Proposition 7.2.5 gives

$$0 \leq \lim_{n\to\infty} \mathbf{E}(M_{T\wedge n} - M_0) = \mathbf{E}\Big(\lim_{n\to\infty} (M_{T\wedge n} - M_0)\Big) = \mathbf{E}(M_T - M_0) \tag{7.2.12}$$

which completes the proof. □

Remark 7.2.7. Probability brings causality by this notion of stopping times. The optional stopping theorem is in some sense a way to test causal filters.

7.2.2 Absorption Probabilities and Hitting Times

Definition 7.2.8 (Absorbing States). For a Markov chain with state space $\mathcal{S}$ and transition probabilities $(p_{ij})_{i,j\in\mathcal{S}}$, an **absorbing state** $j \in \mathcal{S}$ is such that $p_{jj} = 1$. The probability that one reaches an absorbing state is an **absorption probability**.

Martingales can simplify computations involving absorption probabilities (i.e., probabilities of reaching given absorbing states.)

Theorem 7.2.9. Consider a Markov chain with transition probabilities $(p_{ij})_{i,j\in\mathcal{S}}$. Let $a, b \in \mathcal{S}$ be two distinct absorbing states, and define

$$V_a = \inf\{n \in \mathbb{N}_0 \colon X_n = a\}, \quad V_b = \inf\{n \in \mathbb{N}_0 \colon X_n = b\}, \quad T = V_a \wedge V_b. \tag{7.2.13}$$

If $\mathbf{P}(T < \infty \mid X_0 = s) = 1$ and all states are accessible from state s, then there exists a function $h \colon \mathcal{S} \to [0, 1]$ satisfying $h(a) = 1$, $h(b) = 0$, and

$$h(i) = \sum_{j\in\mathcal{S}} p_{ij} h(j), \quad i \in \mathcal{S}. \tag{7.2.14}$$

In particular, for any such h, we have

$$\mathbf{P}(V_a < V_b \mid X_0 = s) = h(s). \tag{7.2.15}$$

Proof. We first show existence of h. To this end, note that definitions imply

$$\mathbf{P}(V_a < V_b \mid X_0 = a) = 1 \quad \text{and} \quad \mathbf{P}(V_a < V_b \mid X_0 = b) = 0. \tag{7.2.16}$$

Moreover, for s such that $\mathbf{P}(T < \infty \mid X_0 = s) = 1$ and all states are accessible from s, the probabilities $\mathbf{P}(V_a < V_b \mid X_0 = j)$ are well-defined for $j \in \mathcal{S}$ (since the comparison $+\infty < +\infty$ is avoided almost surely). Hence, the Markov property implies

$$\mathbf{P}(V_a < V_b \mid X_0 = i) = \sum_{j\in\mathcal{S}} p_{ij}\, \mathbf{P}(V_a < V_b \mid X_0 = j), \quad i \in \mathcal{S}. \tag{7.2.17}$$

Letting $h(i) = \mathbf{P}(V_a < V_b \mid X_0 = i)$ demonstrates the existence of an h satisfying the given linear system.

Assume now that there exists an h satisfying the given system of linear equations. We claim that $(h(X_n))_{n\in\mathbb{N}_0}$ is a martingale with respect to $(X_n)_{n\in\mathbb{N}_0}$. Clearly $(h(X_n))_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$, and we have $\mathbf{E}(|h(X_n)|) \leq 1 < \infty$ for all n by definition. Now, by the Markov property and definition of h,

$$\mathbf{E}(h(X_{n+1}) \mid X_0, \ldots, X_n) = \mathbf{E}(h(X_{n+1}) \mid X_n) = \sum_{j\in\mathcal{S}} p_{X_n j} h(j) = h(X_n). \tag{7.2.18}$$

Thus $(h(X_n))_{n\in\mathbb{N}_0}$ is a martingale with respect to $(X_n)_{n\in\mathbb{N}_0}$. Since T is a stopping time with respect to $(X_n)_{n\in\mathbb{N}_0}$, if s is such that $\mathbf{P}(T < \infty \mid X_0 = s) = 1$, then Optional Stopping Theorem tells us that for $X_0 = s$,

$$h(s) = \mathbf{E}(X_0) = \mathbf{E}(X_T) = \underbrace{h(a)}_{=1}\mathbf{P}(V_a < V_b \mid X_0 = s) + \underbrace{h(b)}_{=0}\mathbf{P}(V_b < V_a \mid X_0 = s) \tag{7.2.19}$$

$$= \mathbf{P}(V_a < V_b \mid X_0 = s). \tag{7.2.20}$$

This confirms uniqueness as desired. □

The above tells us how to compute absorption probabilities, but it does not tell us how long we will need to wait until absorption takes place. This is referred to as a **hitting time**, and the next result addresses how to compute expected hitting times.

Theorem 7.2.10. Consider a finite-state Markov chain with transition probabilities $(p_{ij})_{i,j\in\mathcal{S}}$. Let $a \in \mathcal{S}$ be an absorbing state, define

$$V_a = \inf\{n \in \mathbb{N}_0 \colon X_n = a\} \tag{7.2.21}$$

and assume $i \to a$ for each $i \in \mathcal{S}$. There exists a unique $g \colon \mathcal{S} \to \mathbb{R}$ with $g(a) = 0$ satisfying

$$g(i) = 1 + \sum_{j\in\mathcal{S}} p_{ij} g(j), \quad i \in \mathcal{S} \setminus \{a\}. \tag{7.2.22}$$

Moreover,

$$\mathbf{E}(V_a \mid X_0 = i) = g(i) \quad \text{for all } i \in \mathcal{S}. \tag{7.2.23}$$

Proof. We claim that such a function g exists. Note that

$$\mathbf{E}(V_a \mid X_0 = a) = 0 \tag{7.2.24}$$

and the Markov property implies

$$\mathbf{E}(V_a \mid X_0 = i) = 1 + \sum_{j\in\mathcal{S}} p_{ij}\, \mathbf{E}(V_a \mid X_0 = j), \quad i \in \mathcal{S} \setminus \{a\}. \tag{7.2.25}$$

Now we will show that $\mathbf{E}(V_a \mid X_0 = i) < \infty$ for each $i \in \mathcal{S}$. Note that $i \to a$ implies $\mathbf{P}(V_a < \infty \mid X_0 = i) > 0$. Since this holds for all $i \in \mathcal{S}$ and the state space is finite, there are $\alpha > 0$ and $k \geq 1$ such that

$$\mathbf{P}(V_a \leq k \mid X_0 = i) \geq \alpha > 0, \quad i \in \mathcal{S}. \tag{7.2.26}$$

Now, using the Markov property, for any $n \in \mathbb{N}_0$,

$$\mathbf{P}(V_a > nk \mid X_0 = i) \leq (1-\alpha)^n, \quad i \in \mathcal{S}. \tag{7.2.27}$$

Thus

$$\mathbf{E}(V_a \mid X_0 = i) = \sum_{m\in\mathbb{N}_0} \mathbf{P}(V_a > m \mid X_0 = i) \le k \sum_{n\in\mathbb{N}_0} \mathbf{P}(V_a > nk \mid X_0 = i) < \infty. \tag{7.2.28}$$

Thus $g(i) = \mathbf{E}(V_a \mid X_0 = i)$ establishes existence of the desired function $g\colon \mathcal{S} \to \mathbb{R}$.

Now we show that g is unique. Let g be a function with the given constraints. Then V_a is a stopping time adapted to $(X_n)_{n\in\mathbb{N}_0}$. We will show that $(g(X_n) + (V_a \wedge n))_{n\in\mathbb{N}_0}$ is a martingale. It is adapted to $(X_n)_{n\in\mathbb{N}_0}$, and is integrable for each n due to the assumption of a finite number of states. Note that

$$\mathbf{E}(g(X_{n+1}) \mid X_n) = g(X_n) - 1_{\{X_n \ne a\}} = g(X_n) - 1_{V_a > n}, \tag{7.2.29}$$

and

$$\mathbf{E}((V_a \wedge (n+1)) \mid X_0, \dots, X_n) = V_a 1_{\{V_a \le n\}} + (n+1)\, 1_{\{V_a > n\}}. \tag{7.2.30}$$

Hence, by the Markov property, for $X_n \ne a$,

$$\mathbf{E}(g(X_{n+1}) + (V_a \wedge (n+1)) \mid X_0, \dots, X_n) \tag{7.2.31}$$
$$= \mathbf{E}(g(X_{n+1}) \mid X_n) + \mathbf{E}(V_a \wedge (n+1) \mid X_0, \dots, X_n) \tag{7.2.32}$$
$$= g(X_n) - 1_{\{V_a > n\}} + V_a 1_{\{V_a \le n\}} + (n+1)\, 1_{\{V_a > n\}} \tag{7.2.33}$$
$$= g(X_n) + (V_a \wedge n)\,. \tag{7.2.34}$$

We have already shown that $\mathbf{E}(V_a \mid X_0 = i) < \infty$ for each $i \in \mathcal{S}$. Our martingale has bounded increments, so Optional Stopping Theorem gives for $X_0 = i$,

$$g(i) = \mathbf{E}(g(i) + (V_a \wedge 0)) = \mathbf{E}(g(X_{V_a}) + (V_a \wedge V_a)) = \mathbf{E}(V_a \mid X_0 = i). \tag{7.2.35}$$

Thus g is unique. □

7.3 Concentration Inequalities

Note that the definition of a martingale implies $\mathbf{E}(M_n) = \mathbf{E}(M_0)$ for all $n \in \mathbb{N}_0$. Thus we might suspect that the random variable M_n is close to M_0 in some sense. Under certain assumptions, this indeed becomes the case, and is the focus of this section. In particular, we establish a few of the most well-known and powerful concentration inequalities.

Corollary 7.3.1. Let $(M_n)_{n\in\mathbb{N}_0}$ be a sub-martingale. For any $\alpha > 0$ and $n \in \mathbb{N}_0$,

$$\mathbf{P}\left(\max_{0\le k\le n} M_k \ge \alpha\right) \le \frac{\mathbf{E}(M_n^+)}{\alpha} \tag{7.3.1}$$

where

$$M_n^+ = M_n \vee 0 = \max\{M_n, 0\}\,. \tag{7.3.2}$$

Proof. By Jensen's inequality,

$$\mathbf{E}\left(M_{n+1}^+ \,\middle|\, M_0, \dots, M_n\right) \ge \mathbf{E}(M_{n+1} \mid M_0, \dots, M_n)^+ \ge M_n^+. \tag{7.3.3}$$

Hence $(M_n^+)_{n\in\mathbb{N}_0}$ is also a sub-martingale, so we can assume without loss of generality that $(M_n)_{n\in\mathbb{N}_0}$ is non-negative. Define the stopping time

$$T = \inf\{n \in \mathbb{N}_0 \colon M_n \ge \alpha\}\,. \tag{7.3.4}$$

By Markov's inequality and Proposition 7.2.5,

$$\mathbf{P}\left(\max_{0\le k\le n} M_k \ge \alpha\right) = \mathbf{P}\left(M_{T\wedge n}^+ \ge \alpha\right) \le \frac{\mathbf{E}\left(M_{T\wedge n}^+\right)}{\alpha} \le \frac{\mathbf{E}(M_n^+)}{\alpha}. \tag{7.3.5}$$

□

Example 7.3.2. Let $(X_n)_{n\in\mathbb{N}_0} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2^{-m})$. Define

$$M_n = \sum_{i=1}^{n} X_i, \quad M_0 = 0. \tag{7.3.6}$$

Then

$$\begin{aligned}\mathbf{P}\left(\max_{0\le k\le 2^m} M_k \ge C\right) &= \mathbf{P}\left(\max_{0\le k\le 2^m} \mathrm{e}^{\lambda M_k} \ge \mathrm{e}^{\lambda C}\right) &(7.3.7)\\ &\le \mathrm{e}^{-\lambda C}\,\mathbf{E}\left(\mathrm{e}^{\lambda M_{2^m}}\right) = \mathrm{e}^{-\lambda C}\mathrm{e}^{\lambda^2/2} &(7.3.8)\\ &= \mathrm{e}^{-C^2/2}. &(7.3.9)\end{aligned}$$

This is a concentration inequality for samples of Brownian motion at dyadic rational timesteps, in the sense that

$$\mathbf{P}\left(\max_{0\le k\le 2^m} B_{k2^{-m}} \ge C\right) = \mathbf{P}\left(\max_{0\le k\le 2^m} M_k \ge C\right). \tag{7.3.10}$$

Thus by continuity of Brownian motion,

$$\mathbf{P}\left(\sup_{0\le t\le 1} B_t \ge C\right) \le \mathrm{e}^{-C^2/2}. \tag{7.3.11}$$

Example 7.3.3. If $(M_n)_{n\in\mathbb{N}_0}$ is a non-negative martingale, then for any $\alpha > 0$,

$$\mathbf{P}\left(\sup_{k\in\mathbb{N}_0} M_k \ge \alpha\right) \le \frac{\mathbf{E}(M_0)}{\alpha}. \tag{7.3.12}$$

Theorem 7.3.4 (Azuma-Hoeffding Inequality). Let $(M_n)_{n\in\mathbb{N}_0}$ be a sub-martingale adapted to $(X_n)_{n\in\mathbb{N}_0}$. For each $k \in \mathbb{N}$, suppose there are random variables L_k, U_k, each measurable functions of $X_0, \ldots, X_{k-1}$ (**predictable**), such that

$$L_k \le M_{k-1} - M_k \le U_k \quad \text{almost surely}, \quad U_k - L_k \le c_k \quad \text{almost surely}, \tag{7.3.13}$$

for some constant $c_k \ge 0$. Then for any $n \in \mathbb{N}_0$ and $t > 0$,

$$\mathbf{P}(M_n \le M_0 - t) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\right). \tag{7.3.14}$$

Corollary 7.3.5. If $(M_n)_{n\in\mathbb{N}_0}$ is a martingale that otherwise fulfills the conditions of the Azuma-Hoeffding Inequality, then

$$\mathbf{P}(|M_n - M_0| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\right) \tag{7.3.15}$$

Example 7.3.6. Suppose $(X_n)_{n\in\mathbb{N}_0} \overset{\text{i.i.d.}}{\sim} \text{Rademacher}\left(\frac{1}{2}\right)$. Define M_n by the following:

$$M_0 = 0, \quad M_n = \sum_{i=1}^{n} X_i \quad \text{for all } n \in \mathbb{N}. \tag{7.3.16}$$

Then by Azuma-Hoeffding Inequality,

$$\mathbf{P}(|M_n| \ge t) \le 2\exp\left(-\frac{t^2}{2n}\right). \tag{7.3.17}$$

Picking $t = \sqrt{2n\left(1+\epsilon\right)\log\left(n\right)}$, we get

$$\mathbf{P}\left(\left|\frac{M_n}{\sqrt{n\log\left(n\right)}}\right| \geq \sqrt{2\left(1+\epsilon\right)}\right) \leq \frac{2}{n^{1+\epsilon}}. \tag{7.3.18}$$

By First Borel-Cantelli Lemma,

$$\limsup_{n\to\infty} \frac{\left|M_n\right|}{\sqrt{n\log\left(n\right)}} \leq \sqrt{2} \quad \text{a.s.} \tag{7.3.19}$$

Now we will try to prove Azuma-Hoeffding Inequality. We will need a lemma.

Lemma 0.1.
If Y is real-valued, $\mathbf{E}(Y) \leq 0$, and $a \leq Y \leq b$ almost surely, then

$$\mathbf{E}\left(\mathrm{e}^{\alpha Y}\right) \leq \exp\left(\frac{\alpha^2}{8}\left(b-a\right)^2\right). \tag{7.3.20}$$

Proof. Tedious. □

Proof of Azuma-Hoeffding Inequality. Define

$$Y_k = M_{k-1} - M_k, \quad k \in \mathbb{N}. \tag{7.3.21}$$

Then there exists L_k, U_k such that

$$L_k \leq Y_k \leq U_k \quad \text{a.s.} \tag{7.3.22}$$

And

$$\mathbf{E}(Y_k \mid X_0, \ldots, X_{k-1}) = M_{k-1} - \mathbf{E}(M_k \mid X_0, \ldots, X_{n-1}) \leq 0. \tag{7.3.23}$$

Plugging this into the lemma,

$$\mathbf{E}\left(\mathrm{e}^{\alpha Y_n} \mid X_0, \ldots, X_{k-1}\right) \leq \exp\left(\frac{\alpha^2}{8}\left(U_k - L_k\right)^2\right) \leq \exp\left(\frac{\alpha^2}{8} c_k^2\right). \tag{7.3.24}$$

Then

$$\mathbf{E}\left(\mathrm{e}^{\alpha\left(M_0 - M_n\right)}\right) = \mathbf{E}\left(\prod_{k=1}^{n} \mathrm{e}^{\alpha Y_k}\right) \tag{7.3.25}$$

$$= \mathbf{E}\left(\mathbf{E}\left(\prod_{k=1}^{n} \mathrm{e}^{\alpha Y_k} \,\middle|\, X_0, \ldots, X_{n-1}\right)\right) \tag{7.3.26}$$

$$= \mathbf{E}\left(\prod_{k=1}^{n-1} \mathrm{e}^{\alpha Y_k}\, \mathbf{E}\left(\mathrm{e}^{\alpha Y_n} \mid X_0, \ldots, X_{n-1}\right)\right) \tag{7.3.27}$$

$$\leq \exp\left(\alpha^2 c_k^2 / 8\right) \mathbf{E}\left(\prod_{k=1}^{n-1} \mathrm{e}^{\alpha Y_k}\right) \tag{7.3.28}$$

$$\leq \quad \vdots \tag{7.3.29}$$

$$\leq \quad \vdots \tag{7.3.30}$$

$$\leq \exp\left(\frac{\alpha^2}{8} \sum_{k=1}^{n} c_k^2\right). \tag{7.3.31}$$

By Chernoff bound,

$$\mathbf{P}\left(M_0 - M_n \geq t\right) = \mathbf{P}\left(\mathrm{e}^{\alpha\left(M_0 - M_n\right)} \geq \mathrm{e}^{\alpha t}\right) \tag{7.3.32}$$

$$\leq \exp\left(\frac{\alpha^2}{8}\sum_{k=1}^{n} c_k^2 - \alpha t\right) \tag{7.3.33}$$

$$\leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\right). \qquad \left(\alpha = \tfrac{4t}{\sum_{k=1}^{n} c_k^2}\right)$$

□

Theorem 7.3.7 (McDiarmid's Inequality). Let $X_1, \dots, X_n$ be independent random variables. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function which satisfies

$$\sup_{x_1,\dots,x_n,x_n'} \left|f(x_1,\dots,x_k,\dots,x_n) - f\left(x_1,\dots,x_k',\dots,x_n\right)\right| \leq c_k \quad \text{for } 1 \leq k \leq n. \tag{7.3.34}$$

Then

$$\mathbf{P}(f(X_1,\dots,X_n) \geq \mathbf{E}(f(X_1,\dots,X_n)) + t) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\right). \tag{7.3.35}$$

Corollary 7.3.8. Suppose the conditions of McDiarmid's Inequality are met. Then

$$\mathbf{P}(|f(X_1,\dots,X_n) - \mathbf{E}(f(X_1,\dots,X_n))| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\right). \tag{7.3.36}$$

Example 7.3.9. Let f be L-Lipschitz. If $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}((0,1))$. Then

$$\mathbf{P}(|f(X_1,\dot{;}sX_n) - \mathbf{E}(f(X_1,\dots,X_n))| > t) \leq 2\exp\left(-\frac{2t^2}{L^2 n}\right). \tag{7.3.37}$$

This is because

$$\left|f(X_1,\dots,X_n) - f\left(X_1,\dots,X_k',\dots,X_n\right)\right| \leq L\left|X_k - X_k'\right| \leq L. \tag{7.3.38}$$

It remains to compute $\mathbf{E}(f(X_1,\dots,X_n))$. The brute force method is exponential in n. Define

$$\overline{f}_M\left(X_1^{(1)},\dots,X_n^{(M)}\right) := \frac{1}{M}\sum_{i=1}^{M} f\left(X_1^{(i)},\dots,X_n^{(i)}\right) \tag{7.3.39}$$

where the $X_j^{(i)}$ are random i.i.d. X_j. Then $\overline{f}_M$ is $\frac{L}{M}$-Lipschitz, and

$$\mathbf{E}\left(\overline{f}_M\left(X_1^{(1)},\dots,X_n^{(M)}\right)\right) = \mathbf{E}(f(X_1,\dots,X_n)). \tag{7.3.40}$$

Then

$$\mathbf{P}\left(\left|\overline{f}_M\left(X_1^{(1)},\dots,X_n^{(M)}\right) - \mathbf{E}(f(X_1,\dots,X_n))\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\left(\frac{L^2}{M^2}\right)nM}\right) \tag{7.3.41}$$

$$= 2\exp\left(-\frac{2Mt^2}{L^2 n}\right). \tag{7.3.42}$$

Taking $M \approx nL^2$ examples suffices to ensure $\overline{f}_M \approx \mathbf{E}(f)$.

This reduces the complexity from exponential to linear in the number of samples, and quadratic in the Lipschitz constant.

Example 7.3.10. For a set $A \subseteq [-B,B]^n$, define the "Rademacher complexity"

$$\operatorname{Rad}(A) = \mathbf{E}\left(\frac{1}{n}\sup_{a\in A}\sum_{i=1}^{n} a_i X_i\right) \tag{7.3.43}$$

where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Rademacher}\left(\frac{1}{2}\right)$.

Let

$$f(X_1, \dots, X_n) := \frac{1}{n} \sup_{a \in A} \sum_{i=1}^{n} a_i X_i \tag{7.3.44}$$

is $\frac{2B}{n}$-Lipschitz.

Let $\overline{f}_M$ be defined as before, i.e., for $X_1^{(1)}, \dots, X_n^{(M)} \overset{\text{i.i.d.}}{\sim} \text{Rademacher}\left(\frac{1}{2}\right)$, then

$$\mathbf{P}\left(\left|\overline{f}_M - \text{Rad}(A)\right| \geq t\right) \leq 2 \exp\left(\frac{nMt^2}{2B^2}\right). \tag{7.3.45}$$

We just need $M \approx \frac{2B^2}{n}$ to get $\overline{f}_M \approx \text{Rad}(A)$.

Example 7.3.11. Let $\mathcal{G}(n,p)$ be the "Erdös-Rényi" ensemble of random graphs. To be precise, let $G \sim \mathcal{G}(n,p)$; then G is a graph on n vertices, where edges appear i.i.d. with probability p.

Define $\chi(G)$ to be the chromatic number of G. Then

$$\mathbf{P}\left(|\chi(G) - \mathbf{E}(\chi(G))| > t\sqrt{n}\right) \leq 2\mathrm{e}^{-2t^2} \quad \text{for } G \sim G(n,p). \tag{7.3.46}$$

We can prove this statement. Label the vertices $1, \dots, n$, and let X_i be the edges between vertex i and vertices $i, \dots, i-1$. Then $X_i \in \{0,1\}^{i-1}$. Then G is equivalent to the vector $(X_1, \dots, X_n)$, and χ has 1-bounded differences in each coordinate (for instance adding a new color for a perturbed graph). Applying McDiarmid's Inequality gives the claim.

Proof of McDiarmid's Inequality. We use Doob's martingale. More precisely, let

$$M_0 := 0, \quad M_k := \mathbf{E}(f(X_1, \dots, X_n) \mid X_1, \dots, X_k), \quad k \in \mathbb{N}. \tag{7.3.47}$$

Then apply Azuma-Hoeffding Inequality. □

7.4 The Martingale Convergence Theorem

Our goal is to establish and interpret the following result.

Theorem 7.4.1 (Doob's Martingale Convergence Theorem). Let $(X_n)_{n \in \mathbb{N}_0}$ be a sub-martingale (or super-martingale, or martingale) with respect to $(X_n)_{n \in \mathbb{N}_0}$. If $\sup_{n \in \mathbb{N}_0} \mathbf{E}(|M_n|) < \infty$, then $M := \lim_{n \to \infty} M_n$ exists almost surely, and satisfies $\mathbf{E}(|M|) < \infty$.

The stnadard proof of Doob's Martingale Convergence Theorem is via **Doob's Upcrossing Lemma**. For an interval $[a,b]$ and a sample ω, we define the number of upcrossings $U_N(\omega)$ made by the sample path $\omega \mapsto (M_n(\omega))_{n \in \mathbb{N}_0}$ before time N as

$$U_N(\omega) = \sup\left\{k \in \mathbb{N}_0 \colon \begin{array}{c} \exists 0 \leq m_1 < n_1 < \dots < m_k < n_k \leq N \\ \text{s.t. } M_{m_i}(\omega) < a, M_{n_i}(\omega) > b, 1 \leq i \leq k \end{array}\right\}. \tag{7.4.1}$$

Lemma 7.4.2 (Doob's Upcrossing Lemma). Let $(M_n)_{n \in \mathbb{N}_0}$ be a super-martingale. For any interval $[a,b]$, the number of upcrossings by time N satisfies

$$(b-a)\,\mathbf{E}(U_N) \leq \mathbf{E}\left((M_N - a)^-\right) \tag{7.4.2}$$

where $x^- = -\min\{x, 0\}$.

Proof. Define a new process $(K_n)_{n\in\mathbb{N}}$ as follows: $K_1 = 1_{\{M_0<a\}}$, and

$$K_{n+1} = 1_{\{K_n=1\}}1_{\{M_n\le b\}} + 1_{\{K_n=0\}}1_{\{M_n<a\}}, \quad n\in\mathbb{N}. \tag{7.4.3}$$

Since $(M_n)_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$, it follows that K_n is predictable with respect to $(X_n)_{n\in\mathbb{N}_0}$. Define

$$Y_0 := 0, \quad Y_n = \sum_{k=1}^{n} K_k\left(M_k - M_{k-1}\right), \quad n\in\mathbb{N}. \tag{7.4.4}$$

We claim that $(Y_n)_{n\in\mathbb{N}_0}$ is a supermartingale adapted to $(X_n)_{n\in\mathbb{N}_0}$. Indeed, $(Y_n)_{n\in\mathbb{N}_0}$ is adapted to $(X_n)_{n\in\mathbb{N}_0}$ by the above discussion. Moreover, $|K_k|\le 1$ for all $k\in\mathbb{N}$, integrability of Y_n follows from integrability of $(M_k)_{0\le k\le n}$. Now using the super-martingale property of $(M_n)_{n\in\mathbb{N}_0}$,

$$\begin{aligned}\mathbf{E}(Y_{n+1}\mid X_0,\dots,X_n) &= Y_n + \mathbf{E}(K_{n+1}\left(M_{n+1}-M_n\right)\mid X_0,\dots,X_n) && (7.4.5)\\ &= Y_n + K_{n+1}\left(\mathbf{E}(M_{n+1}\mid X_0,\dots,X_n) - M_n\right) && (7.4.6)\\ &\le Y_n. && (7.4.7)\end{aligned}$$

The construction of Y_n implies

$$Y_N \ge (b-a)\,U_N - (M_N-a)^{-}. \tag{7.4.8}$$

Since expectations are non-increasing for super-martingales,

$$0 = \mathbf{E}(Y_0) \ge \mathbf{E}(Y_N) \ge \mathbf{E}\big((b-a)\,U_N - (M_N-a)^{-}\big) \tag{7.4.9}$$

which completes the proof. □

Proof of Doob's Martingale Convergence Theorem. It suffices to assume $(M_n)_{n\in\mathbb{N}_0}$ is a super-martingale. For $a<b$, define

$$A_{a,b} = \left\{\liminf_{n\to\infty} M_n(\omega) < a < b < \limsup_{n\to\infty} M_n(\omega)\right\} \tag{7.4.10}$$

and $U_{[a,b]} = \lim_{N\to\infty} U_N$, where U_N is the number of upcrossings before time N with respect to the interval $[a,b]$; this limit exists since $N\mapsto U_N(\omega)$ is non-decreasing. By Doob's Upcrossing Lemma, we have

$$(b-a)\,\mathbf{E}(U_N) \le |a| + \sup_{N\in\mathbb{N}} \mathbf{E}(|M_N|) < \infty. \tag{7.4.11}$$

By Monotone Convergence Theorem, this implies $\mathbf{E}\big(U_{[a,b]}\big)<\infty$, and so $\mathbf{P}\big(U_{[a,b]}=\infty\big)=0$. Since $A_{a,b}\subset \big\{U_{[a,b]}=\infty\big\}$, we also have $\mathbf{P}(A_{a,b})=0$.

Now observe that

$$\left\{\lim_{n\to\infty} M_n \text{ does not exist}\right\} = \bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} A_{a,b}. \tag{7.4.12}$$

Since this is a countable union of sets,

$$\mathbf{P}\left(\lim_{n\to\infty} M_n \text{ does not exist}\right) \le \bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} \mathbf{P}(A_{a,b}) = 0. \tag{7.4.13}$$

Hence the limit $m = \lim_{n\to\infty} M_n$ exists almost surely. To finish the proof, apply Fatou's Lemma to conclude

$$\mathbf{E}(|M|) \le \liminf_{n\to\infty} \mathbf{E}(|M_n|) \le \sup_{n\in\mathbb{N}_0} \mathbf{E}(|M_n|) < \infty. \tag{7.4.14}$$

□

8 Poisson Processes

Having studied discrete-time Markov chains previously, we now start laying the groundwork for continuous-time Markov processes. This leads us to Poisson processes. Much like Gaussian processes play a special role in the study of second-order processes, Poisson processes play a special role within the class of so-called counting processes.

8.1 The Exponential Distribution

The basic building block for the Poisson process is the exponential random variable, which enjoys many nice properties.

Definition 8.1.1. The **exponential distribution with rate** $\lambda > 0$, denoted $\text{Exp}(\lambda)$, has density

$$f(t) = \begin{cases} \lambda \mathrm{e}^{-\lambda t} & t \geq 0 \\ 0 & t < 0. \end{cases} \tag{8.1.1}$$

For $T \sim \text{Exp}(\lambda)$, the distribution function is given by

$$\mathbf{P}(T \leq t) = \begin{cases} 1 - \mathrm{e}^{-\lambda t} & t \geq 0 \\ 0 & t < 0. \end{cases} \tag{8.1.2}$$

Proposition 8.1.2. If $T \sim \text{Exp}(\lambda)$, then

$$\mathbf{E}(T) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(T) = \frac{1}{\lambda^2}. \tag{8.1.3}$$

Proposition 8.1.3. One of the most useful and remarkable properties of exponential random variables is the **memoryless property**. In particular, if $T \sim \text{Exp}(\lambda)$, then

$$\mathbf{P}(T > t + s \mid T > t) = \mathbf{P}(T > s). \tag{8.1.4}$$

Proof.

$$\begin{aligned} \mathbf{P}(T > t + s \mid T > t) &= \frac{\mathbf{P}(T > t + s, T > t)}{\mathbf{P}(T > t)} && (8.1.5) \\ &= \frac{\mathbf{P}(T > t + s)}{\mathbf{P}(T > t)} && (8.1.6) \\ &= \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda t}} && (8.1.7) \\ &= \mathrm{e}^{-\lambda s} = \mathbf{P}(T > s). && (8.1.8) \end{aligned}$$

□

Proposition 8.1.4. The exponential distribution is the only memoryless continuous-time distribution.

Proof. Let $t, s \geq 0$ and

$$\mathbf{P}(T > t + s \mid T > t) = \mathbf{P}(T > s) \tag{8.1.9}$$

$$\implies \frac{\mathbf{P}(T > t+s, T > t)}{\mathbf{P}(T > t)} = \mathbf{P}(T > s) \tag{8.1.10}$$

$$\implies \frac{\mathbf{P}(T > t+s)}{\mathbf{P}(T > t)} = \mathbf{P}(T > s) \tag{8.1.11}$$

$$\implies \mathbf{P}(T > t+s) = \mathbf{P}(T > t)\,\mathbf{P}(T > s) \tag{8.1.12}$$

$$\tag{8.1.13}$$

Let $g(x) = \mathbf{P}(T > x)$. Then the factorization is given by

$$g(t+s) = g(t)g(s). \tag{8.1.14}$$

The only non-negative monotone solutions to this functional equation are of the form $g(x) = e^{-\lambda x}$ for $x \geq 0$ and some $\lambda \geq 0$. But this is just the CDF of the exponential distribution. □

Example 8.1.5. Suppose we wait at a bus stop and the duration between bus arrivals are modeled as exponential random variables, with λ buses arriving per minute on average. Even if we've already waited t minutes, we expect to wait another $\frac{1}{\lambda}$ minutes, the same as if we had just arrived to the bus stop.

Proposition 8.1.6. If $T_1, \ldots, T_k \overset{\perp\!\!\!\perp}{\sim} \mathrm{Exp}(\lambda_1), \ldots, \mathrm{Exp}(\lambda_k)$ and $T = \min\{T_1, \ldots, T_k\}$, then $T \sim \mathrm{Exp}\left(\sum_{i=1}^{k} \lambda_i\right)$.

Proof.

$$\mathbf{P}(T > t) = \mathbf{P}\left(\bigcap_{i=1}^{k} \{T_i > t\}\right) \tag{8.1.15}$$

$$= \prod_{i=1}^{k} \mathbf{P}(T_i > t) \tag{8.1.16}$$

$$= \exp\left(-t \sum_{i=1}^{k} \lambda_i\right). \tag{8.1.17}$$

□

Proposition 8.1.7. If $T_1, \ldots, T_k \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$ and $T = \sum_{i=1}^{k} T_i$, then $T \sim \mathrm{Erlang}(k, \lambda)$, with density

$$f(t) = \begin{cases} \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} & t \geq 0 \\ 0 & t < 0. \end{cases} \tag{8.1.18}$$

Proof. We prove it by induction. The base case is just $k = 1$, and indeed $\mathrm{Erlang}(1, \lambda) = \mathrm{Exp}(\lambda)$. By inductive hypothesis, the density for T is then given by the convolution

$$f(t) = \int \lambda e^{-\lambda(t-s)} \frac{(\lambda(t-s))^{k-2}}{(k-2)!} 1_{t-s\geq 0} \lambda e^{-\lambda s} 1_{s \geq 0}\, \mathrm{d}s \tag{8.1.19}$$

$$= \int_0^t \lambda e^{-\lambda(t-s)} \frac{(\lambda(t-s))^{k-2}}{(k-2)!} \lambda e^{-\lambda s}\, \mathrm{d}s \tag{8.1.20}$$

$$= \lambda \frac{\lambda^{k-1} e^{-\lambda t}}{(k-2)!} \int_0^t (t-s)^{k-2}\, \mathrm{d}s \tag{8.1.21}$$

$$= \lambda \frac{\lambda^{k-1} e^{-\lambda t}}{(k-2)!} \frac{t^{k-1}}{k-1} \tag{8.1.22}$$

$$= \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!}. \tag{8.1.23}$$

□

8.2 Poisson Processes

In the last section, we explained how exponential random variables can be thought of as modeling inter-arrival times for certain arrival processes (e.g. buses arriving at a bus stop). The Poisson process is a continuous time random process that simply counts those arrivals. In particular, if $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process, then the random variable N_t is the number of arrivals observed between time 0 and t (inclusive). The following definition makes this precise.

Definition 8.2.1 (Poisson Process). Let $(\tau_n)_{n\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$. For $n \in \mathbb{N}$, define $T_n = \sum_{i=1}^n \tau_i$ with the convention $T_0 = 0$. For each $t \in \mathbb{R}_{\geq 0}$, define teh random variable $N_t = \max\{n \geq 0 \colon T_n \leq t\}$. The process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is called a (homogeneous) **Poisson process** with rate λ.

The poisson process is a canonical example of a counting process. A **counting process** is a random process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$, such that

(i) N_t is a non-negative integer for each time $t \in \mathbb{R}_{\geq 0}$;

(ii) the sample paths $t \mapsto N_t(\omega)$ are non-decreasing in t;

(iii) the sample paths $t \mapsto N_t(\omega)$ are right-continuous.

Definition 8.2.2. A random variable X is said to be Poisson distributed with mean $\lambda \geq 0$ if X has the probability mass function

$$\mathbf{P}(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0. \tag{8.2.1}$$

THis is abbreviated as $X \sim \mathrm{Pois}(\lambda)$.

Proposition 8.2.3. If $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process with rate $\lambda > 0$, then for each $t \in \mathbb{R}_{\geq 0}$, we have $N_t \sim \mathrm{Pois}(\lambda t)$.

Proof. Note that $N_t = n$ if and only if

$$T_n \leq t < T_{n+1} = T_n + \tau_{n+1}. \tag{8.2.2}$$

Hence, by the properties of the exponential distribution,

$$\begin{aligned}
\mathbf{P}(N_t = n) &= \mathbf{P}(T_n \leq t < T_n + \tau_{n+1}) && (8.2.3)\\
&= \mathbf{E}\big(1_{\{T_n \leq t < T_n + \tau_{n+1}\}}\big) && (8.2.4)\\
&= \mathbf{E}\big(1_{\{T_n \leq t\} \cap \{t < T_n + \tau_{n+1}\}}\big) && (8.2.5)\\
&= \mathbf{E}\big(1_{\{T_n \leq t\}} 1_{\{t < T_n + \tau_{n+1}\}}\big) && (8.2.6)\\
&= \int_0^\infty f_{T_n}(s) 1_{s \leq t}\, \mathbf{E}\big(1_{\{t < s + \tau_{n+1}\}}\big)\, \mathrm{d}s && (8.2.7)\\
&= \int_0^t f_{T_n}(s)\, \mathbf{E}(\tau_{n+1} > t - s)\, \mathrm{d}s && (8.2.8)\\
&= \int_0^t \lambda \mathrm{e}^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!} \cdot \mathrm{e}^{-\lambda(t-s)}\, \mathrm{d}s && (8.2.9)\\
&= \mathrm{e}^{-\lambda t}\frac{\lambda^n}{(n-1)!}\int_0^t s^{n-1}\, \mathrm{d}s && (8.2.10)\\
&= \mathrm{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}. && (8.2.11)
\end{aligned}$$

Thus $N_t \sim \mathrm{Pois}(\lambda t)$ as desired. □

Corollary 8.2.4. Poisson processes have the following **Markov property**: $(N_{t+s} - N_s)_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process and is independent of past $(N_r)_{0\leq r<s}$.

Corollary 8.2.5. Poisson processes have **independent and stationary increments**. That is, for any finite sequence of distinct time instants $t_0 < t_1 < t_2 < \cdots < t_k$, the random variables $N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ are independent. Moreover, $N_{t_{n+1}} - N_{t_n} \sim \mathrm{Pois}(\lambda\,(t_{n+1} - t_n))$.

Theorem 8.2.6 (Characterization of Poisson Processes). If $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process, then the following hold:

1. $N_0 = 0$;
2. $N_t \sim \mathrm{Pois}(\lambda t)$ for all $t \in \mathbb{R}_{\geq 0}$.
3. $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ has independent increments.

Conversely, if these properties hold for a counting process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$, then it is a Poisson process.

Proof. The fact that (1) – (3) hold is a direct consequence of the above discussion. Since (1) – (3) define the joint distribution of any finite collection of samples $N_{t_1}, \ldots, N_{t_k}$, the Kolmogorov extension theorem can be invoked to uniquely define the law of the process for all rational time instants $t \in \mathbb{Q}_{\geq 0}$, and right-continuity of sample paths uniquely extends the sample paths to all $t \in \mathbb{R}_{\geq 0}$. □

Remark 8.2.7. Together, (1) – (3) imply that $N_{t+s} - N_s \sim \mathrm{Pois}(\lambda t)$ for all $t, s \geq 0$. Indeed, by the independent increments property, N_{t+s} can be written as the sum of independent random variables

$$N_{t+s} = (N_s - N_0) + (N_{t+s} - N_s)\,. \tag{8.2.12}$$

Since $N_{t+s} \sim \mathrm{Pois}(\lambda\,(t+s))$ and $N_s - N_0 = N_s \sim \mathrm{Pois}(\lambda s)$ by (2) and (1) – (2) respectively, it can be shown that $N_{t+s} - N_s \sim \mathrm{Pois}(\lambda t)$ (e.g. by characteristic functions).

8.3 Transformations of Poisson Processes

Similar to Gaussian processes, the class of Poisson processes are closed under certain operations.

8.3.1 Superposition

Theorem 8.3.1 (Superposition of Poisson Processes). If $(N_{i,t})_{t\in\mathbb{R}_{\geq 0}}$, $i = 1, \ldots, k$ are independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_k$, then

$$N_t = \sum_{i=1}^{k} N_{i,t}, \quad t \geq 0. \tag{8.3.1}$$

is a Poisson process with rate $\sum_{i=1}^{k} \lambda_i$.

Proof. We use Characterization of Poisson Processes. (1) holds because

$$N_0 = \sum_{i=1}^{k} N_{i,0} = \sum_{i=1}^{k} 0 = 0. \tag{8.3.2}$$

(2) holds because $N_{i,t} \sim \mathrm{Pois}(t\lambda_i)$, so

$$N_t = \sum_{i=1}^{k} N_{i,t} \sim \mathrm{Pois}\left(t = \sum_{i=1}^{k} \lambda_i\right). \tag{8.3.3}$$

(3) holds because each $N_{i,t}$ has independent stationary increments, and the sum of independent stationary increments $N_{i,t} - N_{i,s}$ is an independent stationary increment $N_t - N_s$. □

8.3.2 Thinning

This is in some sense the opposite of super-position.

Theorem 8.3.2. Let p be a probability measure over $\left([k], 2^{[k]}\right)$. Let $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ be a Poisson process with rate λ. Suppose that the arrivals counted by $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ are distributed among k classes with probability $p(\{i\})$ that a given arrival is in class i, independent of all other arrivals. Let $(N_{i,t})_{t\in\mathbb{R}_{\geq 0}}$ be the counting process that counts arrivals of class i. Then $(N_{i,t})_{t\in\mathbb{R}_{\geq 0}}$, for $1 \leq i \leq k$, are independent Poisson processes with rates $\lambda_i = p(\{i\})\lambda$.

Proof. It suffices to prove the theorem for $k = 2$, and then one can use the result with $k = 2$ to repeatedly divide $[k]$, getting

$$\begin{aligned}
[k] &= \{1, \ldots, k\} && (8.3.4)\\
&= \{1\} \cup \{2, \ldots, k\} && (8.3.5)\\
&= \{1\} \cup \{2\} \cup \{3, \ldots, k\} && (8.3.6)\\
&\;\;\vdots && (8.3.7)\\
&= \{1\} \cup \{2\} \cup \cdots \cup \{k\} && (8.3.8)
\end{aligned}$$

where each of the sets represent a different thinned independent Poisson process.

Now let us prove the result for $k = 2$. By an abuse of notation, $p := p(\{1\})$, so that $1 - p = p(\{2\})$.

Note that $N_t = N_{1,t} + N_{2,t}$. Then

$$\begin{aligned}
\mathbf{P}(N_{1,t} = m, N_{2,t} = n) &= \sum_{k\in\mathbb{N}_0} \mathbf{P}(N_{1,t} = m, N_{2,t} = n \mid N_t = k) && (8.3.9)\\
&= \mathbf{P}(N_{1,t} = m, N_{2,t} = n \mid N_t = m + n) && (8.3.10)\\
&= \frac{(m+n)!}{m!n!} p^m (1-p)^n \mathrm{e}^{-\mathsf{P}\lambda t} \mathrm{e}^{-(1-p)\lambda t} \frac{(\lambda t)^{m+n}}{(m+n)!} && (8.3.11)\\
&= \mathrm{e}^{-\lambda p t} \frac{(p\lambda t)^m}{m!} \cdot \mathrm{e}^{-(1-p)\lambda t} \frac{((1-p)\,\lambda t)^n}{n!} && (8.3.12)
\end{aligned}$$

so that $N_{1,t} \sim \mathrm{Pois}(p\lambda t) \perp\!\!\!\perp N_{2,t} \sim \mathrm{Pois}((1-p)\,\lambda t)$.

Now we want to show that $(N_{1,t})_{t\in\mathbb{R}_{\geq 0}}$ and $(N_{2,t})_{t\in\mathbb{R}_{\geq 0}}$ are Poisson processes. We prove that $(N_{1,t})_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process and say that $(N_{2,t})_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process by symmetry.

Indeed, $N_{1,0} = 0$. And we have already shown that $N_{1,t} \sim \mathrm{Pois}(p\lambda t)$. Finally $(N_{1,t})_{t\in\mathbb{R}_{\geq 0}}$ has independent increments, inherited from the parent Poisson process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$.

We now need to show that the Poisson processes $(N_{1,t})_{t\in\mathbb{R}_{\geq 0}}$ and $(N_{2,t})_{t\in\mathbb{R}_{\geq 0}}$ are independent. Fix $0 = t_0 < t_1 < t_2 < \cdots < t_k$. We want to show that

$$\begin{aligned}
& (N_{1,t_1}, \ldots, N_{1,t_k}) \perp\!\!\!\perp (N_{2,t_1}, \ldots, N_{2,t_k}) && (8.3.13)\\
\iff & \left(N_{1,t_1} - N_{1,0}, \ldots, N_{1,t_k} - N_{1,t_{k-1}}\right) \perp\!\!\!\perp \left(N_{2,t_1} - N_{2,0}, \ldots, N_{2,t_k} - N_{2,t_{k-1}}\right) && (8.3.14)\\
\iff & N_{1,t_i} - N_{1,t_{i-1}} \perp\!\!\!\perp N_{2,t_i} - N_{2,t_{i-1}}, \quad i \in [k]. && (8.3.15)
\end{aligned}$$

From the first step to the second step we use the fact that $N_{1,0} = N_{2,0} = 1$. From the second step to the third step we use that $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ has independent increments and that the classifications are independent.

Now define $M_s = N_{s+t_{i-1}} - N_{t_{i-1}}$. Then $(M_s)_{s\in\mathbb{R}_{\geq 0}}$ is a Poisson process, and $M_{1,s} \perp\!\!\!\perp M_{2,s}$. Take $s = t_i - t_{i-1} > 0$, giving

$$M_{1,t_i - t_{i-1}} = N_{1,t_i} - N_{1,t_{i-1}}. \qquad (8.3.16)$$

This gives independence of increments of N_1 and N_2.

Thus we can conclude that N_1 and N_2 are independent Poisson processes. □

8.3.3 Inhomogeneous Poisson Processes

Definition 8.3.3 (Inhomogeneous Poisson Process). A counting process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is an inhomogeneous Poisson process with rate function $\lambda\colon t \mapsto \lambda(t) \geq 0$ if the following hold:

1. $N_0 = 0$;
2. $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ has independent increments.
3. $N_t - N_s \sim \mathrm{Pois}\left(\int_s^t \lambda(r)\,\mathrm{d}r\right)$ for $t \geq s \geq 0$.

Note that the inter-arrival times are generally not increments. Nevertheless, by the assumption of independent increments, inhomogeneous Poisson processes still enjoy the Markov property.

If the rate function is bounded from above, then we can construct an inhomogeneous Poisson process by thinning a homogeneous one.

Theorem 8.3.4. Let $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ be a Poisson process with rate λ, and let $p\colon [0,\infty) \to [0,1]$ be Riemann integrable on any finite interval. If we count arrivals at time t with probability $p(t)$, independent of all other arrivals, then we obtain a non-homogeneous Poisson process with rate $\lambda p(t)$.

Proof. By construction, the resulting process (call it $\left(N_t^h\right)_{t\in\mathbb{R}_{\geq 0}}$) inherits $N_0^h = 0$ and independent increments from the Poisson process $(N_t)_{t\in\mathbb{R}_{\geq 0}}$. Now, we can model the thinning process as follows. For each arrival T_i, generate an independent uniform random variable $U_i \sim \mathcal{U}([0,1])$ and accept the arrival if $U_i \leq p(T_i)$. Now the number of arrivals accepted in an interval (s,t) is the same as the number of pairs (T_i, U_i) that fall in the region

$$A = \{(\tau, u) : s \leq \tau \leq t, 0 \leq u \leq p(t)\}\,.$$

Now if A is a finite disjoint union of rectangles, say

$$A = \bigcup_{j=1}^{m} A_j, \quad A_j = \{(\tau, u) : s_j \leq \tau \leq t_j, q_j \leq u \leq p_j\}$$

then by thinning, the number of points falling in rectangle A_j is Poisson with mean $(t_j - s_j)\,(p_j - q_j)$. By superposition of independent Poisson processes, the total number of points is Poisson with mean

$$\sum_{j=1}^{m} (t_j - s_j)\,(p_j - q_j) = \lambda_2(A) = \int_s^t p(r)\,\mathrm{d}r. \tag{8.3.17}$$

The general case follows by approximating A as a union of disjoint rectangles, which is valid by the assumption of Riemann integrability. □

Example 8.3.5 (M/G/∞ Queue). In queueing theory, the M/G/∞ queue refers to a system with a memoryless (M) arrival process (i.e., Poisson), a generic (G) distribution of service times, and infinite (∞) number of servers.

More precisely, suppose customers arrive to a service counter with an infinite number of cashiers according to a Poisson process with rate λ. Each customer remains at their cashier for a service time (a.s. finite) with distribution function G, independent of all other customers. The main question to address is: what is the distribution of X_t, the number of customers in service at time t?

The basic approach is to thin the arrival process according to the following rules:

- Mark the arrival as “Type I” if the customer is finished at time t.
- Mark the arrival as “Type II” if the customer is not finished at time t.

Specifically, consider a customer who enters the service station at time $s \leq t$. We label the customer "Type I" if the service time for that customer satisfies $T \leq t - s$. Otherwise, we label the customer "Type II". Thus we have

$$\mathbf{P}(\text{customer is "Type I"}) = \mathbf{P}(T \leq t - s) = G(t - s). \tag{8.3.18}$$

$$\mathbf{P}(\text{customer is "Type II"}) = \mathbf{P}(T > t - s) = 1 - G(t - s). \tag{8.3.19}$$

The number of customers in service at time t is precisely the number of arrivals labeled as "Type II". Theorem 8.11 implies that this process is a Poisson process of rate $\lambda \cdot (1 - G(t - s))$, for $0 \leq s \leq t$, since all monotone functions are Riemann integrable. In particular,

$$X_t \sim \operatorname{Pois}\left(\lambda \int_0^t (1 - G(t - s)) \, \mathrm{d}s\right). \tag{8.3.20}$$

Moreover, if the expected service time is finite, then as $t \to +\infty$, the distribution of X_t tends to Poisson with rate $\lambda \int_0^\infty (1 - G(s)) \, \mathrm{d}s = \lambda \, \mathbf{E}(T)$. Hence

$$X_t \to \operatorname{Pois}(\lambda \, \mathbf{E}(T)) \quad \text{in distribution.} \tag{8.3.21}$$

8.4 Conditioning on Arrivals

One final property of Poisson processes is that if we condition on the number of arrivals in a given interval, the individual arrivals appear to be independent nad uniformly distributed within the interval.

Definition 8.4.1. Let $X_1, X_2, \ldots, X_k$ be a collection of random variables. The **order statistics** $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$ are the random variables defined by sorting the realizations of $X_1, X_2, \ldots, X_k$ into increasing order.

Theorem 8.4.2. Let $(N_t)_{t \in \mathbb{R}_{\geq 0}}$ be a Poisson process. Condition on the event $\{N_t = n\}$, and let T_i denote the i^{th} arrival time with $T_1 \leq \cdots \leq T_n$. The vector of ordered arrival times $(T_1, \ldots, T_n)$ has the same distribution of order statistics $\big(U_{(1)}, \ldots, U_{(n)}\big)$, where $U_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0, t])$ for $1 \leq i \leq n$.

Proof. Let $V_1 \leq V_2 \leq \cdots \leq V_n$ are the order statistics of $U_1, \ldots, U_n$, then their density is

$$f_V(v_1, \ldots, v_n) = f_U(v_1, \ldots, v_n) \cdot n! \cdot 1_{0 \leq v_1 \leq \cdots \leq v_n \leq t} = \frac{n!}{t^n} 1_{0 \leq v_1 \leq \cdots \leq v_n \leq t}. \tag{8.4.1}$$

Let λ denote the rate of $(N_t)_{t \in \mathbb{R}_{\geq 0}}$. For $0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1} = t$ and $\delta > 0$ sufficiently small, the independent increments of the Poisson process allows us to write

$$\mathbf{P}\left(\bigcap_{i=1}^n \{T_i \in (t_i, t_i + \delta]\} \,\middle|\, N_t = n\right) \tag{8.4.2}$$

$$= \frac{\mathbf{P}(N_{t_1} = 0) \prod_{i=1}^n \mathbf{P}(N_{t_i + \delta} - N_{t_i} = 1) \, \mathbf{P}\big(N_{t_{i+1}} - N_{t_i + \delta} = 0\big)}{\mathbf{P}(N_t = n)} \tag{8.4.3}$$

$$= \frac{\mathrm{e}^{-\lambda t_1} \prod_{i=1}^n \delta \lambda \mathrm{e}^{-\lambda \delta} \mathrm{e}^{-\lambda(t_{i+1} - t_i - \delta)}}{(\lambda t)^n \, \mathrm{e}^{\lambda t} / n!} \tag{8.4.4}$$

$$= \delta^n \frac{n!}{t^n}. \tag{8.4.5}$$

Dividing through by δ^n and taking $\delta \downarrow 0$, we find that $T_1, T_2, \ldots, T_n$ conditioned on $\{N_t = n\}$ admit a density

$$f_{T|N_t}(t_1, \ldots, t_n \mid n) = \frac{n!}{t^n} 1_{0 \leq t_1 \leq \cdots \leq t_n \leq t} \tag{8.4.6}$$

which is precisely the density of the order statistics. □

Example 8.4.3. Given $\{N_t = n\}$, we want to find the distance from $x \in [0,t]$ to the nearest arrival. Let's call this distance $d(x, N)$. Then

$$\mathbf{P}(d(x,N) > \delta \mid N_t = n) = \left(\frac{t - \lambda(B_\delta(x) \cap [0,t])}{t}\right)^n \tag{8.4.7}$$

$$\approx \mathrm{e}^{-\frac{2\delta n}{t}} \tag{8.4.8}$$

which implies that

$$d(x,N) \approx \operatorname{Exp}\left(\frac{2n}{t}\right). \tag{8.4.9}$$

Remark 8.4.4. For inhomogeneous Poisson processes, the construction is the same except for U_i. In particular U_i are i.i.d. and have density $f_U(s) = \frac{\lambda(s)}{\int_0^t \lambda(u)\,\mathrm{d}u}$.

9 Continuous-Time Markov Chains

9.1 Definitions and Constructions

Continuous-time Markov chains are closely related to discrete-time Markov chains and Poisson processes. A continuous-time Markov chain $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is a sequence of random variables indexed by a continuous-time parameter $t \in \mathbb{R}_{\geq 0}$ with X_t called the state of the process at time t. As with discrete-time Markov chains, we will consider countable state spaces, meaning that each X_t takes values in a countable set $\mathcal{S}$. Unless otherwise specified, we generally take $\mathcal{S} = \mathbb{N}_0$.

Definition 9.1.1. A process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is a temporally homogeneous **continuous-time Markov chain** if

(i) given any initial state, with probability 1, the sample paths $t \mapsto X_t(\omega)$ are right-continuous with finitely many discontinuities in any finite time interval; and

(ii) for any choice of discrete time instants $0 \leq t_1 < \cdots < t_k < t \leq s$ and states $i_1, i_2, \ldots, i_k, i, j \in \mathcal{S}$, we have the Markov property

$$\mathbf{P}(X_s = j \mid X_t = i, X_{t_k} = i_k, \ldots, X_{t_1} = i_1) = \mathbf{P}(X_{s-t} = j \mid X_0 = i). \tag{9.1.1}$$

The right-continuity of sample paths is simply convention; it implies that the process assumes the value of the new state at the time of a transition. The assumption of finitely many jumps in any finite interval implies that the Markov process is **regular**. This is often taken as an assumption separate from the definition of a Markov chain, but we adopt regularity as part of our definition since there is not much one can easily say without it.

Example 9.1.2. Consider a process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ taking values in $\mathbb{N}$ where the process spends $T_j \sim \operatorname{Exp}\bigl(j^2\bigr)$ seconds in state j before moving to state $j+1$. By the memoryless property of the exponential distribution, this process satisfies the Markov property:

$$\mathbf{P}(X_{t+h} = j \mid X_t = i, X_{t_k} = i_k, \ldots, X_{t_1} = i_1) = \mathbf{P}(X_h = j \mid X_0 = i), \quad h > 0. \tag{9.1.2}$$

Assuming the process starts in state $X_0 = 1$, let T_n denote the time at which the n^{th} transition takes place. Note that

$$\mathbf{E}(T_n) = \sum_{j=1}^{n} j^{-2} \leq \frac{\pi^2}{6} < 2. \tag{9.1.3}$$

By Markov's inequality,

$$\mathbf{P}(T_n < 4) \geq \frac{1}{2}, \quad n \in \mathbb{N}. \tag{9.1.4}$$

Hence, taking limits,

$$\mathbf{P}\left(\bigcap_{n\in\mathbb{N}} \{T_n < 4\}\right) = \lim_{n\to\infty} \mathbf{P}(T_n < 4) \geq \frac{1}{2}. \tag{9.1.5}$$

Thus with probability at least $\frac{1}{2}$, the process makes an infinite number of transitions in finite time, and as such does not satisfy the assumed regularity conditions of a Markov chain. This pathological behavior is referred to as an **exploding** Markov chain. There do exist methods for dealing with explosive chains, but we do not consider them here.

Example 9.1.3. Let $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ be a homogeneous Poisson process. It is a continuous-time Markov chain due to the memoryless property of the interarrival times and the fact that the number of jumps in any interval is Poisson, and therefore finite with probability one.

Similar to discrete-time Markov chains, we define the transition probabilities for a continuous-time Markov chain as

$$P_{ij}^t := \mathbf{P}(X_t = j \mid X_0 = i), \quad t \geq 0 \tag{9.1.6}$$

where the time parameter t now varies continuously on $[0,\infty)$. The transition probabilities are summarized by the corresponding transition matrix

$$P^t = \left[P_{ij}^t\right]_{i,j\in\mathcal{S}}, \quad t \in \mathbb{R}_{\geq 0}. \tag{9.1.7}$$

We adopt the convention that $P^0 = I$.

Similar to the discrete case, we have the Chapman-Kolmogorov equations:

Theorem 9.1.4. Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ be a continuous-time Markov chain. The transition probabilities satisfy

$$P^{s+t} = P^s P^t, \quad s,t \in \mathbb{R}_{\geq 0} \tag{9.1.8}$$

and $\lim_{t\downarrow 0} P^t = I$. In other words, the transition probabilities $\left(P^t\right)_{t\in\mathbb{R}_{\geq 0}}$ form a **Markov semigroup**.

Proof. The identity $P^{s+t} = P^s P^t$ follows from the Markov property, identically to the Chapman-Kolmogrov equations in the discrete case. To establish the continuity property, we note that the regularity assumption implies $\lim_{t\downarrow 0} X_t = X_0$ almost surely. Since almost sure convergence implies convergence in distribution, $\lim_{t\downarrow 0} \mathbf{P}(X_t = j \mid X_0 = i) = \delta_{ij}$. □

Similar to discrete-time Markov chains, we define an absorbing state i as one where $P_{ii}^t = 1$ for all $t \in \mathbb{R}_{\geq 0}$. If i is not an absorbing state, we say i is non-absorbing.

Theorem 9.1.5. Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ be a continuous-time Markov chain with initial non-absorbing state $X_0 = i$. The holding time $T = \inf\{t \in \mathbb{R}_{\geq 0} \colon X_t \neq i\}$ has distribution $T \sim \operatorname{Exp}(\lambda_i)$ for $\lambda_i > 0$ satisfying

$$P_{ii}^h = 1 - h\lambda_i + o(h). \tag{9.1.9}$$

Moreover, the next state X_T is independent of T and has distribution

$$p_{ij} := \mathbf{P}(X_T = j \mid X_0 = i) = \lim_{h\downarrow 0} \frac{P_{ij}^h}{1 - P_{ii}^h}, \quad j \neq i. \tag{9.1.10}$$

It should be emphasized that the above theorem is a consequence of the definition of a continuous-time Markov chain. However, the assertion means that the following construction is essentially the only way to realize a continuous-time Markov chain. This is because a Markov chain $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ satisfies the **strong Markov property**: $(X_{t+T})_{t\in\mathbb{R}_{\geq 0}}$ is a Markov chain starting at $X_T = j$, independent of $(X_s)_{s\leq T}$, where the random variable T denotes the time of the first transition by $(X_t)_{t\in\mathbb{R}_{\geq 0}}$.

Starting the process in state $X_0 = i$, we wait for time $T_i \sim \operatorname{Exp}(\lambda_i)$ before jumping to a new state $j \neq i$ with probability p_{ij}. Then assuming j is non-absorbing, we wait for time $T_j \sim \operatorname{Exp}(\lambda_j)$ before jumping to a new state $k \neq j$ with probability p_{jk}, and so on. The transition probabilities $\left(p_{ij}\right)_{i,j\in\mathcal{S}}$ (with $p_{ii} = 0$) define a discrete-time Markov chain, known as the **embedded chain**. Stated another way, the sequence of states visited by a continuous-time Markov chain is realized by the discrete-time embedded chain. However, instead of making transitions at regular time instants, the continuous-time chain remains in each state i for an exponentially distributed length of time (with rate λ_i) before transitioning. The parameters $(\lambda_i)_{i\in\mathcal{S}}$ are called the **transition rates** for the Markov chain; indeed, by the above result, λ_i is precisely the rate at which the process transitions out of state i. If i is an absorbing state, we set $\lambda_i = 0$ by convention.

Lemma 9.1.6. Let $(p_{ij})_{i,j\in\mathcal{S}}$ be transition probabilities for a discrete-time Markov chain $(X_n)_{n\in\mathbb{N}_0}$ starting in a non-absorbing state $X_0 = i$.

(i) The random variable $N := \inf\{n \in \mathbb{N}_0 \colon X_n \neq i\}$ is geometric with distribution

$$\mathbf{P}(N = k \mid X_0 = i) = p_{ii}^{k-1}(1 - p_{ii}), \quad k \in \mathbb{N}. \tag{9.1.11}$$

(ii) The random variable X_N is independent of N, and has distribution

$$\mathbf{P}(X_N = j \mid X_0 = i) = \frac{p_{ij}}{1 - p_{ii}}, \quad j \neq i. \tag{9.1.12}$$

Proof. The only non-trivial thing to prove is that N and X_N are independent, given $X_0 = i$. This follows by using Markovity to express the joint distribution:

$$\begin{aligned}
\mathbf{P}(X_N = j, N = k \mid X_0 = i) &= \mathbf{P}(X_k = j, X_{k-1} = i, \ldots, X_1 = i \mid X_0 = i) && (9.1.13)\\
&= p_{ii}^{k-1} p_{ij} && (9.1.14)\\
&= \mathbf{P}(N = k \mid X_0 = i)\,\mathbf{P}(X_N = j \mid X_0 = i). && (9.1.15)
\end{aligned}$$

□

Proof of Theorem 9.1.5. Let $(h_k)_{k\in\mathbb{N}_0}$ be any sequence of positive numbers satisfying $h_k \downarrow 0$. For each $k \in \mathbb{N}_0$, the sampled process $(X_{nh_k})_{n\in\mathbb{N}_0}$ is a discrete-time Markov chain with transition probabilities $P_{ij}^{h_k}$, $i, j \in \mathcal{S}$. Define $N_k = \inf\{n \in \mathbb{N}_0 \colon X_{nh_k} \neq i\}$, which is geometric with parameter $1 - P_{ii}^{h_k}$. In particular,

$$\mathbf{P}(N_k > \lfloor t/h_k \rfloor \mid X_0 = i) = \left(P_{ii}^{h_k}\right)^{\lfloor t/h_k \rfloor} = \exp\left(\lfloor t/h_k \rfloor \log\left(P_{ii}^{h_k}\right)\right), \quad t \in \mathbb{R}_{\geq 0}. \tag{9.1.16}$$

Using the regularity of sample paths, $\lim_{k\to\infty} N_k/h_k = T$ almost surely, so it follows that

$$\mathbf{P}(T > t \mid X_0 = i) = \lim_{k\to\infty} \mathbf{P}(N_k > \lfloor t/h_k \rfloor \mid X_0 = i) = \exp\left(t \lim_{k\to\infty} \frac{1}{h_k} \log\left(P_{ii}^{h_k}\right)\right) \tag{9.1.17}$$

at all continuity points of the distribution of T. In particular, the latter limit must exist and be strictly negative, else $T = \infty$ with probability one, implying that i is absorbing. Further, by right-continuity of sample paths, we must have $\mathbf{P}(T = 0 \mid X_i = 0) = 0$, so that the limit must also be finite. So, we define

$$\lambda_i := -\lim_{k\to\infty} \frac{1}{h_k} \log\left(P_{ii}^{h_k}\right) \in (0, \infty), \tag{9.1.18}$$

and conclude that $T \sim \mathrm{Exp}(\lambda_i)$ as desired. Now we use the bound $-\log(x) \geq 1 - x$ to see that

$$\limsup_{h\downarrow 0} \frac{1 - P_{ii}^h}{h} \leq \lambda_i, \tag{9.1.19}$$

where passage to the continuum limit follows by arbitrariness of the sequence $(h_k)_{k\in\mathbb{N}_0}$. On the other hand, since $P_{ii}^h \to 1$ by the continuity of $t \mapsto P^t$ at $t = 0$, we use the bound $-\log(t) \leq 1 - t + (1-t)^2$, valid for t sufficiently close to 1, to conclude

$$\liminf_{h\downarrow 0} \frac{1 - P_{ii}^h}{h} + \limsup_{h\downarrow 0} h\left(\frac{1 - P_{ii}^h}{h}\right)^2 \geq \liminf_{h\downarrow 0} \frac{1 - P_{ii}^h + \left(1 - P_{ii}^h\right)^2}{h} \geq \lambda_i. \tag{9.1.20}$$

The lim sup term must vanish since the squared term is bounded by $\lambda_i^2 < \infty$ for sufficiently small h, so we conclude that $P_{ii}^h = 1 - h\lambda_i + o(h)$.

Now, using the assumption that sample paths are right-continuous with finitely many jumps in any finite interval, we have $X_{N_k h_k} \to X_T$ almost surely. This implies convergence in distribution, so

$$\mathbf{P}(X_T = j \mid X_0 = i) = \lim_{k\to\infty} \mathbf{P}(X_{N_k h_k} = j \mid X_0 = i) = \lim_{k\to\infty} \frac{P_{ij}^{h_k}}{1 - P_{ii}^{h_k}}, \quad j \neq i. \tag{9.1.21}$$

The only thing to show is independence, which similarly follows since

$$(N_k h_k, X_{N_k h_k}) \to (T, X_T) \quad \text{almost surely,} \tag{9.1.22}$$

and hence by the lemma and the above observations

$$\begin{aligned}
\mathbf{P}(X_T = j, T > t \mid X_0 = i) &= \lim_{k\to\infty} \mathbf{P}(X_{N_k h_k} = j, N_k > \lfloor t/h_k \rfloor \mid X_0 = i) && (9.1.23)\\
&= \lim_{k\to\infty} \mathbf{P}(X_{N_k h_k} = j \mid X_0 = i)\, \mathbf{P}(N_k > \lfloor t/h_k \rfloor \mid X_0 = i) && (9.1.24)\\
&= \mathbf{P}(X_T = j \mid X_0 = i)\, \mathbf{P}(T > t \mid X_0 = i). && (9.1.25)
\end{aligned}$$

□

With the structure of a continuous-time Markov chain articulated by Theorem 9.1.5, we can upgrade continuity of $t \mapsto P^t$ from Theorem 9.1.4 to all time $t \in [0, \infty)$.

Corollary 9.1.7. The map $t \mapsto P^t$ is continuous on $[0, \infty)$.

Proof. Given that $X_0 = i$ (which we can assume to be non-absorbing), let $T_1 \sim \mathrm{Exp}(\lambda_i)$ denote the time of the first transition, and let $(p_{ij})_{i,j\in\mathcal{S}}$ denote the transition probabilities of the embedded chain. We have

$$\begin{aligned}
P_{ij}^t &= \mathbf{P}(X_t = j \mid X_0 = i) && (9.1.26)\\
&= \mathbf{P}(X_t = j, T_1 > t \mid X_0 = i) + \mathbf{P}(X_t = j, T_1 \leq t \mid X_0 = i) && (9.1.27)\\
&= \mathrm{e}^{-\lambda t}\delta_{ij} + \sum_{\substack{k\in\mathcal{S}\\ k\neq i}} \mathbf{P}(X_t = j, X_{T_1} = k, T_1 \leq t \mid X_0 = i) && (9.1.28)\\
&= \mathrm{e}^{-\lambda t}\delta_{ij} + \sum_{\substack{k\in\mathcal{S}\\ k\neq i}} \int_0^t \lambda_i \mathrm{e}^{-\lambda_i s} p_{ik} P_{kj}^{t-s}\, \mathrm{d}s. && (9.1.29)\\
&= \mathrm{e}^{-\lambda t}\delta_{ij} + \lambda_i \mathrm{e}^{-\lambda_i t} \int_0^t \mathrm{e}^{\lambda_i u} \sum_{\substack{k\in\mathcal{S}\\ k\neq i}} p_{ik} P_{kj}^{u}\, \mathrm{d}u. && (9.1.30)
\end{aligned}$$

where the last line follows by a change of variables $u = t - s$ and monotone convergence. The integrand is bounded by $\mathrm{e}^{\lambda_i u}$, so that $t \mapsto P_{ij}^t$ is continuous, as claimed. □

9.2 The Kolmogorov Differential Equations

Altogether, this allows us to compute P^t as the solution to a differential equation.

Definition 9.2.1. The **infinitesimal generator** for a continuous-time Markov chain $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ with transition rates $(\lambda_i)_{i\in\mathcal{S}}$ is a matrix Q with entries

$$q_{ij} := \begin{cases} \lambda_i p_{ij} & j \neq i \\ -\lambda_i & j = i \end{cases}, \tag{9.2.1}$$

where $(p_{ij})_{i,j\in\mathcal{S}}$ are the transition probabilities for the embedded chain. In particular, $\lambda_i = \sum_{\substack{j\in\mathcal{S}\\ j\neq i}} q_{ij}$.

The numbers $(q_{ij})_{i,j\in\mathcal{S}}$ are called the **jump rates** for the Markov chain; the reason for this terminology is that q_{ij} describes the rate at which the Markov chain with infinitesimal generator Q transitions from state i to state j $(i \neq j)$. Indeed, if we recall the construction described following Theorem 9.1.5, then we can view the transition out of state i as coincident with the first arrival of a Poisson process with rate λ_i. By independence of the subsequent state and the holding time, the characterization of q_{ij} as the transition rate from state i to state j follows by the thinning property of Poisson processes.

The term *infinitesimal generator* (sometimes called the *generator*) comes from the fact that the matrix Q encodes all information about the process in terms of how the probabilities change on an infinitesimal scale. Indeed, Theorem 9.1.5 implies the following.

Corollary 9.2.2. For a continuous-time Markov chain with infinitesimal generator Q, the transition probabilities satisfy

$$P^h_{ii} = 1 - h\lambda_i + o(h), \quad i \in \mathcal{S} \tag{9.2.2}$$

$$P^h_{ij} = hq_{ij} + o(h), \quad i, j \in \mathcal{S}, i \neq j. \tag{9.2.3}$$

Proof. The first statement was proved in Theorem 9.1.5. The second follows since for $j \neq i$,

$$p_{ij} = \lim_{h\downarrow 0} \frac{P^h_{ij}}{1 - P^h_{ii}} = \lim_{h\downarrow 0} \lambda_i^{-1} \frac{P^h_{ij}}{h}. \tag{9.2.4}$$

□

Together with the semigroup property, this suggests that we can express P^t in terms of a differential equation involving Q. This is made precise by the **Kolmogorov Differential Equations**.

Theorem 9.2.3. For a continuous-time Markov chain with infinitesimal generator Q, the map $t \mapsto P^t$ is continuously differentiable on $[0, \infty)$, and satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P^t = QP^t \tag*{(Kolmogorov Backward Equation.)}$$

Moreover, if $\sup_{i\in\mathcal{S}} \lambda_i < \infty$[1], then

$$\frac{\mathrm{d}}{\mathrm{d}t} P^t = P^t Q \tag*{(Kolmogorov Forward Equation.)}$$

Moreover, $P^t = \mathrm{e}^{tQ} = \sum_{k\in\mathbb{N}_0} t^k \frac{Q^k}{k!}$ for all $t \in \mathbb{R}_{\geq 0}$.

Sketch. We use the asymptotics to write

$$P^h = I + hQ + o(h) \tag{9.2.5}$$

giving the formal expressions

$$P^{t+h} = P^t P^h = P^t + hP^t Q + o(h) \tag{9.2.6}$$

$$= P^h P^t = P^t + hQP^t + o(h). \tag{9.2.7}$$

These are the Kolmogorov forward and backward equations, respectively. Asserting convergence is a bit tedious from the perspective of real analysis. □

Example 9.2.4. Let $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ be a Poisson process with rate λ. Then $(N_t)_{t\in\mathbb{R}_{\geq 0}}$ is a continuous-time Markov chain with jump rates $q_{k,k+1} = \lambda$ and transition rates $\lambda_k = \lambda$ for all $k \in \mathbb{R}_{\geq 0}$ (other jump rates are all zero). The transition probabilities satisfy

$$P^t_{k,k+n} = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad t \in \mathbb{R}_{\geq 0}, \quad k, n \in \mathbb{N}_0, \tag{9.2.8}$$

and $P^t_{k,m} = 0$ for $m < k$.

[1] A sufficient condition for the chain to not be explosive.

Example 9.2.5 ($M/M/s$ Queue). Let a continuous-time Markov chain $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ model the number of customers in a system (i.e., in queue and in service) at time $t \in \mathbb{R}_{\geq 0}$. We assume there are s servers; if they are all busy, those customers not being served wait in the queue until a server opens up. If customers arrive to the system at rate λ, then

$$q_{n,n+1} = \lambda. \tag{9.2.9}$$

If customers spend an average time $\frac{1}{\mu}$ in service, then we have the jump rates

$$q_{n,n-1} = \begin{cases} n\mu & 1 \leq n \leq s \\ s\mu & s < n \end{cases} = \min\{n, s\}\,\mu. \tag{9.2.10}$$

If we want to introduce balking, we can assume that if n people are in the system, then an arriving customer joins with probability a_n. Then the jump rates above are modified as

$$q_{n,n+1} = \lambda a_n, \tag{9.2.11}$$

with others remaining the same.

Example 9.2.6 (Birth-Death Process). Our discrete-time model of a birth-death process is potentially a poor model for biological models since births and deaths were forced to occur at a rate of one per discrete time increment. In many cases, it is more natural to model the population as a continuous-time Markov chain where a population of n individuals experiences births at rate λ_n and deaths at rate μ_n. In this case, our jump rates are

$$q_{n,n+1} = \lambda_n \tag{9.2.12}$$
$$q_{n,n-1} = \mu_n. \tag{9.2.13}$$

If individuals give birth at rate λ and expire at rate μ, then

$$q_{n,n+1} = \lambda n \tag{9.2.14}$$
$$q_{n,n-1} = \mu n. \tag{9.2.15}$$

Note that the transition probabilitiies for the embedded chain are given in this case by

$$p_{n,n+1} = \frac{\lambda}{\lambda + \mu} \tag{9.2.16}$$
$$p_{n,n-1} = \frac{\mu}{\lambda + \mu}. \tag{9.2.17}$$

Thus the embedded jump process behaves exactly like a discrete-time birth-death process. The additional information encoded in the continuous-time process is given by the timing of events.

9.3 Continuous-Time Markov Limit Theorems

9.3.1 Characterization of Stationary Distributions

Definition 9.3.1. A continuous-time Markov chain is **irreducible** if the embedded chain is irreducible and has at least two states.

The assumption that there are at least two states rules out the degenerate situation of a constant-valued process. Under this assumption, an irreducible Markov chain has the property that all transition rates are strictly positive.

Definition 9.3.2. A **stationary distribution** for a continuous-time Markov chain with transition probabilities $\left(P^t\right)_{t\in\mathbb{R}_{\geq 0}}$ is a probability (row) vector p satisfying $p = pP^t$ for all $t \in \mathbb{R}_{\geq 0}$.

We refrain from using π to denote the stationary distribution for continuous-time Markov chains, since we reserve it for when we invoke the stationary distribution of the embedded chain. In the discrete-time case, the stationary distribution is characterized as a left eigenvector of the transition matrix. In the continuous-time case, we require that $p = pP^t$ for all $t \in \mathbb{R}_{\geq 0}$. However, under the further assumption that $\sum_{i\in\mathcal{S}} p_i\lambda_i < \infty$, we have a simple characterization given by the following theorem.

Theorem 9.3.3. Consider a continuous-time Markov chain with infinitesimal generator Q. A vector p satisfying $\sum_{i\in\mathcal{S}} p_i\lambda_i < \infty$ is a stationary distribution if and only if $pQ = 0$. Moreover, if the chain is irreducible, then p is the unique stationary distribution.

Proof. Write

$$\frac{\mathrm{d}\left(pP^t\right)}{\mathrm{d}t} = p\frac{\mathrm{d}}{\mathrm{d}t}P^t = pQP^t. \tag{9.3.1}$$

If p is stationary, then $\frac{\mathrm{d}\left(pP^t\right)}{\mathrm{d}t} = 0$, so $pQP^t = 0$ for all $t \in \mathbb{R}_{\geq 0}$. Taking $t = 0$ implies that $pQ = 0$.

Conversely if $pQ = 0$ then $\frac{\mathrm{d}}{\mathrm{d}t}\left(pP^t\right) = 0$ and p is a stationary distribution.

To establish the claim of uniqueness, note that p must also be a stationary distribution for the discrete-time Markov chain with transition matrix $P = P^h$ for any $h > 0$. By irreducibility of the continuous-time Markov chain, the discrete-time Markov chain is also irreducible, and thus the stationary distribution on the discrete-time Markov chain is unique. Taking p to be this stationary distribution proves uniqueness. □

9.3.2 Stationary Distributions and Embedded Chains

The stationary distributions for the continuous-time Markov chain and the discrete-time embedded chain are in close correspondence. It is possible to deduce one from the other, provided some regularity.

Theorem 9.3.4. Consider an irreducible continuous-time Markov chain with generator Q. The following are equivalent.

1. The continuous-time chain has stationary distribution p satisfying $\sum_{i\in\mathcal{S}} p_i\lambda_i < \infty$.
2. The embedded chain has stationary distribution π satisfying $\sum_{i\in\mathcal{S}} \pi_i/\lambda_i < \infty$.

Moreover, if either are true, then the stationary distributions are unique, and $\pi_k = p_k\lambda_k / \sum_{i\in\mathcal{S}} p_i\lambda_i$.

Remark 9.3.5. The continuous-time condition rules out exploding chains; the embedded condition rules out so-called "sticky chains" which stay in the same state for infinitely long.

9.3.3 Ergodic Theorems and Asymptotic Convergence

Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ be an irreducible continuous-time Markov chain. Define

$$T_j = \inf\left\{t \in \mathbb{R}_{\geq 0} \colon X_t = j, \exists s \in (0,t) \text{ s.t. } X_s \neq j\right\}, \quad m_j = \mathbf{E}(T_j). \tag{9.3.2}$$

Note that it is possible for $\mathbf{P}(T_j = \infty) > 0$, i..e, if the embedded chain is transient.

State j is positive recurrent if $m_j < \infty$; null recurrent if $m_j = \infty$. As before, transience and (positive/null) recurrence are class properties.

Theorem 9.3.6. An irreducible, continuous-time Markov chain satisfies exactly one of the following:

1. All states are transient, or all states are null recurrent. In this case

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t P_{ij}^s \,\mathrm{d}s = 0 \tag{9.3.3}$$

and no stationary distribution exists.

2. All states are positive recurrent. In this case, a unique stationary distribution exists and is given by

$$p_j = \frac{1}{m_j \lambda_j} = \lim_{t\to\infty} \frac{1}{t}\int_0^t P_{ij}^s \,\mathrm{d}s. \tag{9.3.4}$$

Corollary 9.3.7. Consider an irreducible, positive-recurrent continuous-time Markov chain with stationary distribution p. If $\sum_{i\in\mathcal{S}} p_i\lambda_i < \infty$, then the embedded chain is also positive recurrent, and has stationary distribution π defined by

$$\pi_j = \frac{p_j\lambda_j}{\sum_{i\in\mathcal{S}} p_i\lambda_i}, \quad j \in \mathcal{S}. \tag{9.3.5}$$

In particular,

$$m_j = \frac{\sum_{i\in\mathcal{S}} \pi_i/\lambda_i}{\pi_j}. \tag{9.3.6}$$

Theorem 9.3.8. Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ be an irreducible continuous-time Markov chain with initial state $X_0 = i$. Define

$$f_j(t) = \frac{1}{t}\int_0^t 1_{\{X_s=j\}} \,\mathrm{d}s \tag{9.3.7}$$

to be the fraction of time that the chain spends in state j, up to time $t \geq 0$. It holds that

$$\lim_{t\to\infty} f_j(t) = \frac{1}{\lambda_j m_j} \quad \text{a.s.} \tag{9.3.8}$$

Theorem 9.3.9 (Ergodic Theorem for Continuous-Time Markov Chains). Consider an irreducible, positive-recurrent continuous-time Markov chain, and let p denote its stationary distribution. If $r\colon \mathcal{S} \to \mathbb{R}$ is bounded, then for any initial state $X_0 = i$,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t r(X_s)\,\mathrm{d}s = \mathbf{E}_{X\sim p}(r(X)) \quad \text{a.s.} \tag{9.3.9}$$

Proof. We can assume without loss of generality that $0 \leq r \leq 1$. By Fubini-Tonelli,

$$\frac{1}{t}\int_0^t r(X_s)\,\mathrm{d}s = \frac{1}{t}\int_0^t \sum_{j\in\mathcal{S}} r(j) 1_{\{X_s=j\}}\,\mathrm{d}s \tag{9.3.10}$$

$$= \sum_j r(j) f_j(t). \tag{9.3.11}$$

The difficulty is that although the $f_j(t)$ converge almost surely, the infinite sum may not converge almost surely, a priori.

But this turns out to be the case. Take $\mathcal{J} \subseteq \mathcal{S}$ to be finite. Then $\sum_{j\in\mathcal{J}} f_j(t) \to \sum_{j\in\mathcal{J}} p_j$ almost surely. Then $\sum_{j\in\mathcal{J}^c} f_j(t) \to \sum_{j\in\mathcal{J}^c} p_j$ almost surely. To finish, we choose $\mathcal{J}$ large enough that $\sum_{j\in\mathcal{J}^c} p_j \leq \epsilon$ and bound the convergent series. □

9.4 Reversibility

An irreducible positive recurrent Markov chain will always converge to the stationary distribution:

Theorem 9.4.1. Suppose $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is an irreducible positive-recurrent Markov chain with transition probabilities $\left(P^t\right)_{t\in\mathbb{R}_{\geq 0}}$ and stationary distribution p. Then

$$\lim_{t\to\infty} \left\|P_{i\cdot}^t - p\right\|_{\mathrm{TV}} = 0 \quad \text{for each } i \in \mathcal{S}.$$

Our goal is to characterize the rate of convergence.

9.4.1 Definition and Examples

Definition 9.4.2. A continuous-time Markov chain with generator Q is said to be **reversible** if there is a probability vector p satisfying $\sum_{i\in\mathcal{S}} p_i\lambda_i < \infty$ and the detailed balance condition

$$p_i q_{ij} = p_j q_{ji} \quad \text{for all } i, j \in \mathcal{S}. \tag{9.4.1}$$

In this case, we say (Q, p) is reversible.

Proposition 9.4.3. Similarly to the discrete case, p in the definition of reversibility is also a stationary distribution for the Markov chain.

Proof. Looking at the j^{th} entry of pQ, we see

$$(pQ)_j = \sum_{i\in\mathcal{S}} p_i q_{ij} = \sum_{i\in\mathcal{S}} p_j q_{ji} = p_j \underbrace{\sum_{i\in\mathcal{S}} q_{ji}}_{=0} \tag{9.4.2}$$

$$= 0. \tag{9.4.3}$$

Hence $pQ = 0$ so p is stationary for the Markov chain. □

Proposition 9.4.4. The embedded chain of a reversible continuous-time Markov chain is reversible.

Proof. Let π be the stationary distribution for the embedded chain.

$$p_i q_{ij} = p_i \lambda_i p_{ij} = \left(\sum_{k\in\mathcal{S}} p_k\lambda_k\right) \pi_i p_{ij} \tag{9.4.4}$$

$$p_j q_{ji} = p_j \lambda_j p_{ji} = \left(\sum_{k\in\mathcal{S}} p_k\lambda_k\right) \pi_j p_{ji} \tag{9.4.5}$$

but these are equal so that

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{9.4.6}$$

Hence detailed balance equations are fulfilled for the embedded chain. □

Remark 9.4.5. Similar to the discrete case, the term “reversible” comes from the fact that if $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ and $X_0 \sim p$, then for any $t \in \mathbb{R}_{\geq 0}$,

$$(X_s)_{0\leq s\leq t} = (X_{t-s})_{0\leq s\leq t} \quad \text{in distribution.} \tag{9.4.7}$$

In particular, when the Markov chain is stationary, the forward and reversed processes are equal in distribution.

Remark 9.4.6. Many Markov chains encountered in practice are reversible, and hence solving the detailed balance equations provides an easy way to simultaneously compute a stationary distribution and verify reversibility.

Example 9.4.7. Consider a birth-death process with immigration. That is, birth rates $q_{n,n+1} = n\lambda + \gamma$, $n \in \mathbb{N}_0$, and death rates $q_{n,n-1} = n\mu$, $n \in \mathbb{N}$. The detailed balance equations read

$$p_1 = p_0 \frac{\gamma}{\mu} \quad \text{and} \quad p_{n+1} = p_n \frac{n}{n+1} \frac{\lambda + \gamma/n}{\mu}, \quad n \in \mathbb{N}. \tag{9.4.8}$$

In particular, iterating the last, we find

$$p_n = p_0 \frac{\gamma}{\lambda} \frac{1}{n} \left(\frac{\lambda}{\mu}\right)^n \prod_{k=1}^{n-1} \left(1 + \frac{\gamma}{k\lambda}\right), \quad n \in \mathbb{N}. \tag{9.4.9}$$

Hence, if $\gamma > 0$ and $\lambda < \mu$, the chain is reversible and admits a stationary distribution, since then the RHS is summable over $n \in \mathbb{N}$. Indeed, note that

$$\prod_{k=1}^{n-1}\Big(1+\frac{\gamma}{k\lambda}\Big) \le \exp\left(\frac{\gamma}{\lambda}\sum_{k=1}^{n-1}\frac{1}{k}\right) = \exp\Big(\frac{\gamma}{\lambda}H_{n-1}\Big) \le \exp\Big(\frac{\gamma}{\lambda}\log(n)\Big) = n^{\gamma/\lambda}. \tag{9.4.10}$$

Example 9.4.8. Consider the $M/M/\infty$ queue with arrival rate γ and service rate μ. This is really just the previous example with $\lambda = 0$. The system is reversible with the stationary probabilities

$$p_n = \mathrm{e}^{-\gamma/\mu}\frac{(\gamma/\mu)^n}{n!}, \quad n \in \mathbb{N}_0. \tag{9.4.11}$$

A more substantial example is given by the $M/M/1$ queue, which demonstrates the remarkable behavior that customers leaving the system in equilibrium also follow a Poisson process, provided that the queue is stable.

Theorem 9.4.9 (Burke's Theorem). Consider a $M/M/1$ queue with arrival rate λ and service rate μ. If $\lambda < \mu$, then in equilibrium,

(i) the departure process is Poisson with rate λ; and

(ii) the number of customers in the queue at time $t > 0$ is independent of departures prior to t.

9.4.2 Spectral Gap and Trend to Equilibrium

For a reversible continuous-time Markov chain (Q, p), we define the **Dirichlet form**

$$\mathcal{E}(f,f) := \frac{1}{2}\sum_{i,j\in\mathcal{S}} |f(i)-f(j)|^2 q_{ij}p_i, \quad f \in L^2\big(\mathcal{S}, 2^{\mathcal{S}}, p\big) = -\frac{1}{2}\langle f, Qf\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},p)}. \tag{9.4.12}$$

The **spectral gap** for the Markov chain is defined to be the largest $\lambda \ge 0$ such that

$$\mathrm{Var}_p(f) \le \frac{1}{\lambda}\mathcal{E}(f,f) \quad \text{for all } f \in L^2\big(\mathcal{S}, 2^{\mathcal{S}}, p\big), \tag{9.4.13}$$

where $\mathrm{Var}_p(f) := \mathrm{Var}(f(X))$ for $X \sim p$. The term *spectral gap* refers to the spectrum of Q, which by definition of reversibility can be shown to be an (unbounded) self-adjoint linear operator on $L^2\big(\mathcal{S}, 2^{\mathcal{S}}, \mathbf{P}\big)$. We know that $Q \cdot 1 = 0$, so that the spectrum of Q contains the point $\{0\}$. It can be shown that $-Q$ is positive semidefinite. So if a reversible Markov chain (Q, p) admits a spectral gap $\lambda > 0$, then it means that the spectrum of Q is contained in $(-\infty, -\lambda] \cup \{0\}$ (i.e., the spectrum of Q has a nontrivial gap).

Similar to the discrete-time setting, if a reversible continuous-time Markov chain has positive spectral gap, then it enjoys exponentially fast convergence to equilibrium. This section is devoted to proving the following result.

Theorem 9.4.10. If a reversible continuous-time Markov chain (Q, p) admits spectral gap λ, then

$$\big\|P^t_{i\cdot} - p\big\|_{\mathrm{TV}^2} \le \frac{\mathrm{e}^{-2\lambda t}}{p_i}, \quad t \in \mathbb{R}_{\ge 0}. \tag{9.4.14}$$

Proof Sketch. The basic idea is to compute the derivative

$$\frac{\partial}{\partial t}\big\|P^t f\big\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},p)} = -2\mathcal{E}\big(P^t f, P^t f\big) \le -2\lambda\,\mathrm{Var}_p(f) = -2\lambda\,\|f\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},p)}. \tag{9.4.15}$$

for $f \in L^2\big(\mathcal{S}, 2^{\mathcal{S}}, p\big)$ and $\langle f, 1\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},p)} = 0$. Integrating both sides with respect to t, we get

$$\big\|P^t f\big\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},p)} \le \mathrm{e}^{-2\lambda t}\,\|f\|^2_{L^2(\mathcal{S},2^{\mathcal{S}},p)} \quad \text{for } t \in \mathbb{R}_{\ge 0}. \tag{9.4.16}$$

If $h = \frac{\mathrm{d}\mu}{\mathrm{d}p}$, then $\langle h-1, 1\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},p)} = 0$ and the above implies

$$\left\|\mu P^t - p\right\|_{\mathrm{TV}}^2 \le \left\|P^t\left(h-1\right)\right\|_{L^2(\mathcal{S},2^{\mathcal{S}},p)}^2 \le \mathrm{e}^{-2\lambda t}\left\|h-1\right\|_{L^2(\mathcal{S},2^{\mathcal{S}},p)}^2 \tag{9.4.17}$$

In particular, if $\mu = \delta_i$, then $h_j = \frac{1}{\pi_i}\delta_{ij}$ and

$$\left\|h-1\right\|_{L^2(\mathcal{S},2^{\mathcal{S}},p)}^2 = \frac{1-p_i}{p_i} \le \frac{1}{p_i}. \tag{9.4.18}$$

□

Example 9.4.11. The $M/M/\infty$ queue with arrival rate γ and service rate μ admits spectral gap μ. The proof follows by expanding the test function f in the definition of the spectral gap in terms of an orthogonal basis corresponding to the eigenvectors of the generator Q. In this case, we can use the **Poisson-Charlier polynomials**. For real parameter $a > 0$ and integer $n \in \mathbb{N}_0$, these polynomials are defined recursively via

$$c_{n+1}(x;a) = \frac{x}{a}c_n(x-1;a) - c_n(x;a), \quad x \in \mathbb{R} \tag{9.4.19}$$

where $c_0(x,a) = 1$. These polynomials are orthogonal with respect to a Poisson kernel in the sense that

$$\langle c_m, c_n\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a))} = a^{-n}n!\delta_{mn} \tag{9.4.20}$$

and form an orthogonal basis for the space $L^2\big(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a)\big)$. So, for $f \in L^2\big(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a)\big)$ with $\langle f,1\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a))} =$ 0, we can write $f = \sum_{n\in\mathbb{N}}\alpha_n c_n(\cdot;a)$ for some real coefficients $(\alpha_n)_{n\in\mathbb{N}}$. Now we can use orthogonality to write

$$\mathrm{Var}_{\mathrm{Pois}(a)}(f) = \langle f,f\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a))} = \sum_{n\in\mathbb{N}}\alpha_n^2 a^{-n}n!. \tag{9.4.21}$$

The Poisson-Charlier polynomials satisfy the identity

$$\left(k+a-n\right)c_n(k;a) = kc_n(k-1);a + ac_n(k+1;a), \quad k \in \mathcal{S}, n \in \mathbb{N}_0. \tag{9.4.22}$$

For $a = \gamma/\mu$ this is equivalent to

$$n\mu c_n(\cdot;a) = -Qc_n(\cdot;a). \tag{9.4.23}$$

In other words, $c_n(\cdot;a)$ is an eigenvector of $-Q$ with eigenvalue $n\mu$. Since $\mathcal{E}(f,f) = -\langle f,Qf\rangle_{L^2(\mathcal{S},2^{\mathcal{S}},\mathrm{Pois}(a))}$, orthogonality and the above eigenvector equation together give

$$\mathcal{E}(f,f) = \mu\sum_{n\in\mathbb{N}}n\alpha_n^2 a^{-n}n! \ge \mu\sum_{n\in\mathbb{N}}\alpha_n^2 a^{-n}n! = \mu\,\mathrm{Var}_{\mathrm{Pois}(a)}(f). \tag{9.4.24}$$

Hence the $M/M/\infty$ queue admits spectral gap equal to μ as claimed.

Example 9.4.12. Suppose an $M/M/\infty$ queue with arrival rate γ and service rate μ starts empty. If X_t denotes the number of people in the system at time t, then

$$d_{\mathrm{TV}}(X_t,\mathrm{Pois}(\gamma/\mu)) \le \mathrm{e}^{-\mu t+\gamma/(2\mu)}, \quad t \in \mathbb{R}_{\ge 0}. \tag{9.4.25}$$

In particular, at time $t(\epsilon) = \gamma/\left(2\mu^2\right) + \log(1/\epsilon)/\mu$, the total variation distance between the distribution of X_t and the stationary distribution is no more than ϵ.

10 Renewal Processes

10.1 Definition and Examples

In this chapter, we give a brief introduction to renewal processes. A renewal process can be thought of as a generalization of the Poisson process; it is a counting process with holding times that are i.i.d., but no longer assumed to be exponential. This means that renewal processes do not generally satisfy the Markov property enjoyed by Poisson processes. For example, processes that count the number of times a Markov chain re-enters its starting state is a renewal process.

Definition 10.1.1. Let $(\tau_i)_{i\in\mathbb{N}}$ be i.i.d. non-negative random variables with $0 < \mathbf{E}(\tau_1) < \infty$ and, for $n \in \mathbb{N}$, define $T_n = \sum_{i=1}^{n} \tau_i$, with the convention that $T_0 = 0$. For each $t \in \mathbb{R}_{\geq 0}$, define the random variable $X_t = \sup\{n \in \mathbb{N}_0 \colon T_n \leq t\}$. The process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is called a **renewal process** with holding times $(\tau_i)_{i\in\mathbb{N}}$.

Example 10.1.2. A homogeneous Poisson process is a renewal process.

Example 10.1.3. If $(X_t)_{t\geq 0}$ is a positive-recurrent (continuous-time or discrete-time) Markov chain with $X_0 = i$, then the process $(N_i(t))_{t\geq 0}$ that counts re-entries into state i up to and including time t is a renewal process. Indeed, the times between re-entries are i.i.d. by the Markov property.

10.2 Asymptotics of Renewal Processes

Similar to what we encountered for Markov chains, renewal processes enjoy a strong law of large numbers.

Theorem 10.2.1 (Strong Law for Renewal Processes). For a renewal process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ with i.i.d. holding times $(\tau_i)_{i\in\mathbb{N}}$,

$$\lim_{t\to\infty} \frac{X_t}{t} = \frac{1}{\mathbf{E}(\tau_1)} \quad \text{a.s.} \quad . \tag{10.2.1}$$

Proof. First, since $\mathbf{E}(\tau_1) < \infty$, we have $X_t \to \infty$ a.s.. Now, since $T_{X_t} \leq t \leq T_{X_t+1}$, we have

$$\frac{X_t}{X_{t+1}} \cdot \frac{X_t + 1}{T_{X_t+1}} \leq \frac{X_t}{t} \leq \frac{X_t}{T_{X_t}}. \tag{10.2.2}$$

Using $X_t \to \infty$ a.s., we apply the strong law of large numbers to conclude

$$\lim_{t\to\infty} \frac{X_t}{X_t + 1} = 1 \quad \text{a.s.} \tag{10.2.3}$$

and

$$\lim_{t\to\infty} \frac{T_{X_t}}{X_t} = \lim_{t\to\infty} \frac{1}{X_t} \sum_{i=1}^{X_t} \tau_i = \mathbf{E}(\tau_1) \quad \text{a.s.} \tag{10.2.4}$$

which completes the proof. □

Example 10.2.2. The strong laws for positive-recurrent discrete-time and continuous-time Markov chains can be proved as corollaries of the strong law for renewal processes.

For a renewal process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$, we define the **renewal function** $m(t) = \mathbf{E}(X_t)$ for $t \geq 0$. The following result is called the elementary renewal theorem. It states that the strong law for renewal processes also holds in expectation.

Theorem 10.2.3. For a renewal process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ with holding times $(\tau_i)_{i\in\mathbb{N}}$, the renewal function satisfies

$$\lim_{t\to\infty}\frac{m(t)}{t} = \frac{1}{\mathbf{E}(\tau_1)}. \tag{10.2.5}$$

Proof. If τ_1 is deterministic, then the result is trivial and follows by the previous theorem. by scaling time if needed, we can assume without loss of generality that $\mathbf{P}(\tau_1 \leq 1) = p \in (0,1)$. Now consider the renewal process $(\overline{X}_t)_{t\in\mathbb{R}_{\geq 0}}$ with holding times $\overline{\tau}_i = 1_{\{\tau_i \geq 1\}}$. Since $\overline{\tau}_i \leq \tau_i$ for each $i \in \mathbb{N}$, it holds that $\overline{X}_t \leq X_t$ for all $t \geq 0$. By construction, $(\overline{X}_t)_{t\in\mathbb{R}_{\geq 0}}$ only has jumps at integer values of $t \in \mathbb{R}_{\geq 0}$ and

$$\frac{1}{n}\overline{X}_n = \frac{1}{n}\sum_{i=0}^{n} G_i, \tag{10.2.6}$$

where $(G_i)_{i\in\mathbb{N}_0}$ are i.i.d. $\mathrm{Geo}(p)$ random variables supported on $\mathbb{N}_0$. By direct computation,

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\big(\overline{X}_n\big) = \lim_{n\to\infty}\frac{n+1}{n}\mathbf{E}(G_0) = \mathbf{E}(G_0). \tag{10.2.7}$$

Also by the strong law of large numbers,

$$\lim_{n\to\infty}\frac{\overline{X}_n}{n} = \mathbf{E}(G_0) \quad \text{a.s.} \tag{10.2.8}$$

Since $X_t \leq \overline{X}_t = \overline{X}_{\lfloor t\rfloor}$ by construction, it holds that

$$\frac{X_t}{t} \leq \frac{\overline{X}_{\lfloor t\rfloor}}{\lfloor t\rfloor}, \tag{10.2.9}$$

so by the generalized dominated convergence theorem, we can justify the exchange of the integral and the sum

$$\lim_{t\to\infty}\frac{\mathbf{E}(X_t)}{t} = \mathbf{E}\left(\lim_{t\to\infty}\frac{X_t}{t}\right) = \frac{1}{\mathbf{E}(\tau_1)}. \tag{10.2.10}$$

□

Remark 10.2.4. The strong law and elementary renewal theorem continue to hold if $\mathbf{E}(\tau_1) = +\infty$. To see this, consider teh renewal process $(\overline{X}_t)_{t\in\mathbb{R}_{\geq 0}}$ with holding times $\overline{\tau}_i = \tau_i \wedge N$ for fixed $N > 0$. It holds that $0 \leq X_t \leq \overline{X}_t$, so

$$0 \leq \limsup_{t\to\infty}\frac{X_t}{t} \leq \lim_{t\to\infty}\frac{\overline{X}_t}{t} = \frac{1}{\mathbf{E}(\tau_1 \wedge N)} \qquad \text{a.s.} \tag{10.2.11}$$

Letting $N \to \infty$, monotone covergence establishes $\mathbf{E}(\tau_1 \wedge N) \to +\infty$, so that $\frac{X_t}{t} \to 0$ a.s., as desired. Similarly, since $\mathbf{E}(X_t) \leq \mathbf{E}\big(\overline{X}_t\big)$, we have $m(t)/t \to 0$. This situation is relevant when considering null-recurrent Markov chains, where the time between re-entries into a given state are infinite in expectation (by definition of null recurrence).

10.3 Renewal-Reward Processes

Let $(W_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbf{E}(|W_1|) < \infty$, which we call *rewards*. For a renewal process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$, the associated **renewal-reward** process is defined by

$$R_t = \sum_{i=1}^{X_t} W_i, \quad t \in \mathbb{R}_{\geq 0}. \tag{10.3.1}$$

The idea is that we collect some reward W_i at the i^{th} renewal, and R_t tracks the cumulative reward earned as a function of time $t \in \mathbb{R}_{\geq 0}$. Note that the rewards $(W_i)_{i\in\mathbb{N}}$ are independent, but they may depend on the sequence of holding times $(\tau_i)_{i\in\mathbb{N}}$. Indeed, a typical example of a renewal-reward process is obtained when $W_i = \tau_i^2/2$, for $i \in \mathbb{N}$. The resulting renewal-reward process is called the *residual life process*, because the quantity R_t/t corresponds to the average time a person arriving at a uniform point in the interval $[0,t)$ would need to wait until the next renewal.

For a renewal-reward process $(R_t)_{t\in\mathbb{R}_{\geq 0}}$, the corresponding **reward function** is defined as $r(t) = \mathbf{E}(R_t)$. The following is both the strong law and the elementary theorem for renewal-reward processes.

Theorem 10.3.1. For a renewal-reward process $(R_t)_{t\in\mathbb{R}_{\geq 0}}$ with holding times $(\tau_i)_{i\in\mathbb{N}}$ and rewards $(W_i)_{i\in\mathbb{N}}$,

$$\lim_{t\to\infty} \frac{R_t}{t} = \frac{\mathbf{E}(W_1)}{\mathbf{E}(\tau_1)} \quad \text{a.s.} \tag{10.3.2}$$

Moreover, the reward function satisfies

$$\lim_{t\to\infty} \frac{r(t)}{t} = \frac{\mathbf{E}(W_1)}{\mathbf{E}(\tau_1)}. \tag{10.3.3}$$

Proof. Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ denote the underlying renewal process. The strong law follows from that for renewal processes; just write

$$\frac{R_t}{t} R_t = \frac{X_t}{t} \frac{1}{X_t} \sum_{i=1}^{X_t} W_i, \tag{10.3.4}$$

and apply the strong law for renewal processes together with the strong law of large numbers.

To prove the second statement, observe that for any fixed $t \in \mathbb{R}_{\geq 0}$, $X_t + 1$ is a stopping time with respect to the sequence $(\tau_i, W_i)_{i\in\mathbb{N}}$. Indeed, $X_t + 1 = \inf\{n \in \mathbb{N}_0 \colon \sum_{i=1}^n \tau_i > t\}$, so the event $\{X_t + 1 = n\}$ is determined entirely by $\tau_1, \ldots, \tau_n$ for each $n \in \mathbb{N}$. The Elementary Renewal Theorem ensures that $\mathbf{E}(X_t + 1) < \infty$, so Wald's identity applies to give

$$r(t) + \mathbf{E}(W_1) = \mathbf{E}\left(\sum_{i=1}^{X_t+1} W_i\right) = (\mathbf{E}(X_t) + 1)\,\mathbf{E}(W_1). \tag{10.3.5}$$

As a result, $r(t) = m(t)\,\mathbf{E}(W_1)$, and the claim follows by the Elementary Renewal Theorem. □

Example 10.3.2. Let $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ be an arrival process with interarrival times $(\tau_i)_{i\in\mathbb{N}}$ having finite variance, and let $T_n = \sum_{i=1}^n \tau_i$ be the time of the n^{th} arrival, with $T_0 = 0$. If a person arrives at a random time, uniformly within the interval $[0,t)$, then the expected time they wait until the next arrival is

$$\overline{W}_t = \frac{1}{t}\int_0^t W(s)\,\mathrm{d}s, \tag{10.3.6}$$

where for $t \in [T_{X_t}, T_{X_t+1})$, $W(t) := T_{X_t+1} - t$. Note that

$$W_t \sim \frac{1}{t}\sum_{i=1}^{X_t} \frac{\tau_i^2}{2} \quad \text{a.s.} \tag{10.3.7}$$

Thus

$$\lim_{t\to\infty} \overline{W}_t = \frac{1}{2}\cdot\frac{\mathbf{E}(\tau_1^2)}{\mathbf{E}(\tau_1)} = \frac{1}{2}\left(\mathbf{E}(\tau_1) + \frac{\operatorname{Var}(\tau_1)}{\mathbf{E}(\tau_1)}\right) \quad \text{a.s. and in expectation.} \tag{10.3.8}$$

This may seem like a paradox. Indeed, we might initially expect that the average waiting time is just $\frac{1}{2}\,\mathbf{E}(\tau_1)$. However, the above calculation shows that one waits, on average, for a larger amount of time whenever the interarrival times are non-deterministic. The resolution of this apparent contradiction comes from realizing that a randomly arriving passenger is more likely to arrive in renewal periods corresponding to holding times of longer duration. As a result, the expected waiting time is biased upward.

10.4 Little's Law

Little's law is a remarkable characterization of the long-term behavior in a $G/G/1$ queue. The notation $G/G/1$ means that customers enter a system with a single server according to an arrival process with generic interarrival times (i.e., a renewal process), and the service times follow some other generic distribution. We do not assume any service ordering, such as first-come, first-served. However, once a customer enters service, they remain in service until completion and we further assume the server never sits idle if there are customers in the system.

To make the setup precise, let $(A_t)_{t\in\mathbb{R}_{\geq 0}}$ count the number of arrivals up to time t, which is assumed to be a renewal process with holding times $(\tau_i)_{i\in\mathbb{N}}$. Let $(D_t)_{t\in\mathbb{R}_{\geq 0}}$ count the number of departures out of the system up to time t. Evidently, $N_t = A_t - D_t$ is non-negative, and equal to the number of people in teh system at time $t \in \mathbb{R}_{\geq 0}$. Let W_i denote the time the i^{th} customer spends in the system.

Theorem 10.4.1 (Little's Law). For the $G/G/1$ queue described above, if the average service time μ satisfies $\mu < \mathbf{E}(\tau_1) < \infty$, then the limits

$$\overline{N} = \lim_{t\to\infty} \frac{1}{t} \int_0^t N_s \,\mathrm{d}s \quad \text{and} \quad \overline{W} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i \tag{10.4.1}$$

exist almost surely, and satisfy

$$\overline{N} = \frac{\overline{W}}{\mathbf{E}(\tau_1)}. \tag{10.4.2}$$

11 Hypothesis Testing

All chapters have adopted a probabilistic point of view, wherein we fix a model (i.e., a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and/or a sequence of random variables $(Y_n)_{n\in\mathbb{N}}$), and characterize what behavior that model would exhibit.

For this final chapter, we turn things upside down and adopt a statistical point of view. That is, if nature reveals to us a sample $\omega \in \Omega$, to what extent can we distinguish whether $(\Omega, \mathcal{F}, \mathbf{P}_0)$ or $(\Omega, \mathcal{F}, \mathbf{P}_1)$ is the better model for the underlying probability space? Or, how confident can we be in accepting (or rejecting) $(\Omega, \mathcal{F}, \mathbf{P}_0)$ as a reasonable model?

These problems fall under the umbrella of **hypothesis testing**, aptly named since we aim to discriminate between two or more hypotheses given empirical data. We will consider two hypothesis testing scenarios in this chapter. The first is binary hypothesis testing, where we attempt to discriminate between two given hypotheses. The second is the multiple comparisons problem, where we seek to accept or reject each of multiple hypotheses.

11.1 Binary Hypothesis Testing

Consider a measurable space $(\Omega, \mathcal{F})$ which can be equipped with either of two probability measures we call P_0 and P_1. On the basis of observing a sample $\omega \in \Omega$, we would like to decide whether $(\Omega, \mathcal{F}, \mathbf{P}_0)$ or $(\Omega, \mathcal{F}, \mathbf{P}_1)$ is the better model. The former is called the **null hypothesis** H_0, and the latter is called the **alternate hypothesis** H_1.

A **test** is a function $\widehat{H}\colon \Omega \to \{H_0, H_1\}$ that is measurable in the sense that $\widehat{H}^{-1}(H_0) \in \mathcal{F}$ (this automatically implies $\widehat{H}^{-1}(H_1) \in \mathcal{F}$ also). Associated with any test $\widehat{H}$ are two fundamental error probabilities:

- The **Type I error rate** (or false positive probability); and
- The **Type II error rate** (or false negative probability).

More precisely,

$$\text{Type I error rate} \quad := \quad \mathbf{P}_0\Big(\widehat{H} = H_1\Big) \tag{11.1.1}$$

$$\text{Type II error rate} \quad := \quad \mathbf{P}_1\Big(\widehat{H} = H_0\Big). \tag{11.1.2}$$

The terms Type I and Type II are conventional, and are not terribly descriptive. However, the term "false positive probability" is jsutified because our test $\widehat{H}$ rejects the null hypothesis if the revealed sample ω is not in $\widehat{H}^{-1}(H_0)$. If the null hypothesis is correct, then the probability of this event happening is $\mathbf{P}_0\Big(\widehat{H} = H_1\Big)$. The term "false negative probability" is similarly justified.

The **power** of a test $\widehat{H}$ is the probability of avoiding a Type II error, and is therefore equal to $\mathbf{P}_1\Big(\widehat{H} = H_1\Big)$.

11.1.1 Likelihood Ratio

To avoid unnecessary complications[1], we assume henceforth that $\mathbf{P}_1 \ll \mathbf{P}_0$, which means that $\mathbf{P}_0(A) = 0 \implies \mathbf{P}_1(A) = 0$. Indeed, if this is not the case, then there is some set $A \in \mathcal{F}$ such that $\mathbf{P}_0(A) = 0$, but $\mathbf{P}_1(A) > 0$, and any reasonable test should automatically accept hypothesis H_1 whenever a sample $\omega \in A$ is observed. So, using the assumption that $\mathbf{P}_1 \ll \mathbf{P}_0$, the Radon-Nikodym theorem ensures that there is a measurable function $\Lambda\colon \Omega \to \mathbb{R}_{\geq 0}$ satisfying the "change of measure" identity

$$\mathbf{E}_{\mathbf{P}_1}(1_A) = \mathbf{E}_{\mathbf{P}_0}(\Lambda 1_A) \quad \text{for all } A \in \mathcal{F}. \tag{11.1.3}$$

[1] One can handle the case where $\mathbf{P}_1 \not\ll \mathbf{P}_0$ by first invoking the Lebesgue decomposition theorem to express $\mathbf{P}_1$ as a mixture, with one component absolutely continuous with respect to $\mathbf{P}_0$, and the other component singular with respect to $\mathbf{P}_0$.

The function Λ is called a Radon-Nikodym derivative (usually denoted by $\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0}$), but in the context of hypothesis testing it is generally referred to as the **likelihood ratio** since it can be thought of as the relative likelihood of observing a sample ω under the different hypotheses H_1 and H_0.

Defining Λ as a Radon-Nikodym derivative allows us to formulate the hypothesis testing problem for general probability measures. However, when $\mathbf{P}_0$ and $\mathbf{P}_1$ have densities in the usual sense, then Λ is simply the ratio of densities. Similarly, if Ω is discrete, then Λ is the ratio of the probability mass functions. This can be checked by verifying the change of measure identity.

Example 11.1.1. If $\mathbf{P}_i$ has density p_i, then $\Lambda(\omega) = \frac{p_1(\omega)}{p_0(\omega)}$. Indeed, with this choice of Λ,

$$\begin{aligned} \mathbf{E}_{\mathbf{P}_1}(1_A) &= \mathbf{E}_{\mathbf{P}_0}(\Lambda 1_A) && (11.1.4)\\ \int_A \mathbf{P}_1(\mathrm{d}\omega) &= \int_A \Lambda(\omega)\,\mathbf{P}_0(\mathrm{d}\omega) && (11.1.5)\\ \int_A p_0(\omega)\,\mathrm{d}\omega &= \int_A \Lambda(\omega) p_1(\omega)\,\mathrm{d}\omega && (11.1.6)\\ &= \int_A \frac{p_1(\omega)}{p_0(\omega)} p_0(\omega)\,\mathrm{d}\omega && (11.1.7)\\ &= \int_A p_1(\omega). && (11.1.8) \end{aligned}$$

Example 11.1.2. If Ω is discrete, $\mathcal{F} = 2^{\Omega}$, and $\mathbf{P}_i$ has probability mass function p_i, then

$$\Lambda(\omega) = \frac{p_1(\omega)}{p_0(\omega)} \tag{11.1.9}$$

for the same reason.

Example 11.1.3. Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}}$, and $\mathbf{P}_1 = \mathcal{N}(0,1)$. Let $\mathbf{P}_0$ be defined via

$$\mathbf{P}_0(A) = \frac{1}{2}\mathbf{P}_1(A) + \frac{1}{2} 1_A(0), \quad A \in \mathcal{F}. \tag{11.1.10}$$

Thus $\mathbf{P}_0$ is an equal measure of the standard normal distribution and a mass point at 0. In this case,

$$\Lambda(\omega) = \begin{cases} 2 & \omega \neq 0 \\ 0 & \omega = 0 \end{cases}. \tag{11.1.11}$$

11.1.2 Threshold Tests and the Error Curve

An important class of tests are the likelihood-ratio **threshold tests**, for which decision regions are defined based on the simple rule of thresholding the likelihood ratio.

Definition 11.1.4. Assume $\mathbf{P}_1 \ll \mathbf{P}_0$ and let $\eta \geq 0$. The threshold test with threshold η, denoted $\widehat{H}_\eta$, is defined according to

$$\widehat{H}_\eta(\omega) = \begin{cases} H_1 & \Lambda(\omega) \geq \eta \\ H_0 & \Lambda(\omega) < \eta \end{cases}. \tag{11.1.12}$$

Example 11.1.5. Suppose $\mathbf{P}_1 \ll \mathbf{P}_0$. The maximum likelihood test is a threshold test with $\eta = 1$.

More interesting is the maximum a posteriori test. Suppose we have a prior belief that H_0 is true with probability $\pi_0 < 1$ and H_1 is true with probability $\pi_1 = 1 - \pi_0$. The maximum a posteriori (MAP) test is the threshold test with threshold $\eta = \pi_0/\pi_1$, and has the property that it minimizes the total error probability among all tests. Indeed, let $\widehat{H}_{\mathrm{MAP}}$ denote the threshold test with threshold $\eta = \pi_0/\pi_1$. Under the prior π, the total error probability for any test $\widehat{H}$ satisfies

$$\mathbf{P}\Big(\widehat{H} \text{ is wrong}\Big) = \pi_0\,\mathbf{P}_0\Big(\widehat{H} = H_1\Big) + \pi_1\,\mathbf{P}_1\Big(\widehat{H} = H_0\Big) \tag{11.1.13}$$

$$= \pi_0 \, \mathbf{E}_{\mathbf{P}_0}\Big(1_{\{\widehat{H}=H_1\}}\Big) + \pi_1 \, \mathbf{E}_{\mathbf{P}_0}\Big(1_{\{\widehat{H}=H_0\}}\Big) \tag{11.1.14}$$

$$= \pi_0 + \pi_1 \, \mathbf{E}_{\mathbf{P}_0}\left(\left(\Lambda - \frac{\pi_0}{\pi_1}\right) 1_{\{\widehat{H}=H_0\}}\right) \tag{11.1.15}$$

$$\geq \pi_0 + \pi_1 \, \mathbf{E}_{\mathbf{P}_0}\left(\left(\Lambda - \frac{\pi_0}{\pi_1}\right) 1_{\{\widehat{H}_{\mathrm{MAP}}=H_0\}}\right) \tag{11.1.16}$$

$$= \mathbf{P}\Big(\widehat{H}_{\mathrm{MAP}} \text{ is wrong}\Big). \tag{11.1.17}$$

The inequality follows by the inequality on the integrands, which itself follows by definition of $\widehat{H}_{\mathrm{MAP}}$.

The threshold tests $(\widehat{H}_\eta)_{\eta \in \mathbb{R}_{\geq 0}}$ define the **error curve**, which plays a fundamental role in characterizing the best tradeoff between Type I and Type II error rates.

Definition 11.1.6. Assume $\mathbf{P}_1 \ll \mathbf{P}_0$ and let Λ denote the likelihood ratio. The error curve $u\colon [0,1] \to \mathbb{R}$ is defined via

$$u(\theta) := \sup_{\eta \in \mathbb{R}_{\geq 0}} \left[\mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big) + \eta \left(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big) - \eta\right)\right], \quad 0 \leq \theta \leq 1. \tag{11.1.18}$$

Note that as the pointwise supremum of affine functions, u is a convex function on $[0,1]$.

11.1.3 The Neyman-Pearson Lemma

We say that a test $\widehat{H}$ **lies above the error curve** if

$$\mathbf{P}_1\Big(\widehat{H} = H_0\Big) \geq u\Big(\mathbf{P}_0\Big(\widehat{H} = H_1\Big)\Big), \tag{11.1.19}$$

and we say that $\widehat{H}$ **lies on the error curve** if this is met with equality.

Theorem 11.1.7 (Neyman-Pearson Lemma). Assume $\mathbf{P}_1 \ll \mathbf{P}_0$. All tests $\widehat{H}$ lie above the error curve. Moreover, every threshold test $\widehat{H}_\eta$ lies on the error curve.

Proof. Fix any $\eta \in [0, \infty)$ and choose $\pi_0 \in [0,1)$ to satisfy $\pi_0/(1-\pi_0) = \eta$. Let $\widehat{H}_{\mathrm{MAP}}$ denote the MAP test for prior probabilities π_0 and $\pi_1 := 1 - \pi_0$. Then, for any test $\widehat{H}$, the minimum-error property of the MAP estimator ensures

$$\pi_1 \left(\mathbf{P}_1\Big(\widehat{H} = H_0\Big) + \eta \, \mathbf{P}_0\Big(\widehat{H} = H_1\Big)\right) \geq \pi_1 \left(\mathbf{P}_1\Big(\widehat{H}_{\mathrm{MAP}} = H_0\Big) + \eta \, \mathbf{P}_0\Big(\widehat{H}_{\mathrm{MAP}} = H_1\Big)\right). \tag{11.1.20}$$

Dividing through by π_1 (which is positive by construction), we have

$$\mathbf{P}_1\Big(\widehat{H} = H_0\Big) \geq \mathbf{P}_1\Big(\widehat{H}_{\mathrm{MAP}} = H_0\Big) + \eta \left(\mathbf{P}_0\Big(\widehat{H}_{\mathrm{MAP}} = H_1\Big) - \mathbf{P}_0\Big(\widehat{H} = H_1\Big)\right). \tag{11.1.21}$$

Since $\widehat{H}_{\mathrm{MAP}} = \widehat{H}_\eta$ and $\eta \in \mathbb{R}_{\geq 0}$ is arbitrary, we conclude

$$\mathbf{P}_1\Big(\widehat{H} = H_0\Big) \geq \mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big) + \eta \left(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big) - \mathbf{P}_0\Big(\widehat{H} = H_1\Big)\right), \quad \eta \in \mathbb{R}_{\geq 0}. \tag{11.1.22}$$

Taking supremum over $\eta \in \mathbb{R}_{\geq 0}$ shows

$$\mathbf{P}_1\Big(\widehat{H} = H_0\Big) \geq u\Big(\mathbf{P}_0\Big(\widehat{H} = H_1\Big)\Big) \tag{11.1.23}$$

by definition of the error curve u, so that $\widehat{H}$ lies above the error curve.

Now, it follows again by definition of the error curve that

$$u\Big(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big)\Big) \geq \mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big) \quad \text{for every } \eta \geq 0, \tag{11.1.24}$$

so that all threshold tests lie below the error curve. However, we just saw that any test lies above the error curve, so threshold tests must lie on it. □

The Neyman-Pearson lemma ensures that threshold tests are optimal in the sense that they lie on the error curve, and any other test lies above. However, it does not guarantee that every point on the error curve can be achieved by a threshold test, as the following example illustrates.

Example 11.1.8. Consider $\Omega = \{0,1\}$, $\mathcal{F} = 2^{\{0,1\}}$. Now let $\mathbf{P}_0 = \text{Bernoulli}(1/3)$ and $\mathbf{P}_1 = \text{Bernoulli}(2/3)$. Direct computation gives

$$\Lambda(\omega) = \begin{cases} 2 & \omega = 1 \\ 1/2 & \omega = 0 \end{cases}. \tag{11.1.25}$$

For any $\eta \in \mathbb{R}_{\geq 0}$, the threshold test $\widehat{H}_\eta$ can be expressed as

$$\widehat{H}_\eta(\omega) = \begin{cases} H_1 & \eta \leq 1/2 \\ H_\omega & 1/2 < \eta \leq 2\,, \\ H_0 & 2 < \eta \end{cases} \quad \omega \in \{0,1\}\,. \tag{11.1.26}$$

This allows us to compute the Type I and Type II error probabilities as

$$\Big(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big), \mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big)\Big) = \begin{cases} (1,0) & \eta \leq 1/2 \\ (1/3,1/3) & 1/2 < \eta \leq 2\,. \\ (0,1) & 2 < \eta \end{cases} \tag{11.1.27}$$

Evidently, the mapping $\eta \mapsto \Big(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big), \mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big)\Big)$ swept over all $\eta \in \mathbb{R}_{\geq 0}$ just produces 3 points in the plane, whereas the error curve is the piecewise-linear function that joins these points.

The above example, while negative, suggests that the error curve is the lower convex hull of the set of points $\Big(\mathbf{P}_0\Big(\widehat{H}_\eta = H_1\Big), \mathbf{P}_1\Big(\widehat{H}_\eta = H_0\Big)\Big)_{\eta \in \mathbb{R}_{\geq 0}}$. This is indeed the case, and can be verified using standard convexity arguments. The main point to be made here is that any point on the error curve can be expressed as a convex combination of points achieved by threshold tests. This leads to the definition of a **randomized threshold test**.

Definition 11.1.9. Fix parameters $\eta_0, \eta_1 \in \mathbb{R}_{\geq 0}$ and $p \in [0,1]$. The corresponding (randomized) threshold test $\widehat{H}$ is defined by taking $R \sim \text{Bernoulli}(p)$, independent of 1_A for all $A \in \mathcal{F}$, and putting

$$\widehat{H}(\omega) = \widehat{H}_{\eta_R}(\omega). \tag{11.1.28}$$

The randomized test $\widehat{H}$ defined above is a random variable that, averaged over R, achieves Type I error

$$\mathbf{P}_0\Big(\widehat{H} = H_1\Big) = (1-p)\,\mathbf{P}_0\Big(\widehat{H}_{\eta_0} = H_1\Big) + p\,\mathbf{P}_0\Big(\widehat{H}_{\eta_1} = H_1\Big) \tag{11.1.29}$$

and Type II error

$$\begin{aligned} \mathbf{P}_1\Big(\widehat{H} = H_0\Big) &= (1-p)\,\mathbf{P}_1\Big(\widehat{H}_{\eta_0} = H_0\Big) + p\,\mathbf{P}_1\Big(\widehat{H}_{\eta_1} = H_0\Big) & (11.1.30)\\ &= (1-p)\,u\Big(\mathbf{P}_0\Big(\widehat{H}_{\eta_0} = H_1\Big)\Big) + pu\Big(\mathbf{P}_0\Big(\widehat{H}_{\eta_1} = H_1\Big)\Big). & (11.1.31) \end{aligned}$$

As a result, by varying the parameters $\eta_0, \eta_1 \in \mathbb{R}_{\geq 0}$ and $p \in [0,1]$, any point on the error curve is achievable by a (possibly randomized) threshold test. For a fixed $\theta \in [0,1]$, the **Neyman-Pearson rule** is defined to be the (possibly randomized) threshold test that achieves Type II error probability $u(\theta)$ subject to the constraint that Type I error probability does not exceed θ. In other words, subject to a Type I error constraint, the Neyman-Pearson rule is the most powerful test.

11.1.4 Sufficient Statistics

A consequence of the Neyman-Pearson lemma and the subsequent discussion on randomized tests demonstrates that optimal hypothesis testing procedures can be implemented with knowledge of the likelihood $\Lambda(\omega)$ alone, without further information about the particular sample ω. In other words, $\Lambda(\omega)$ is sufficient for testing the hypothesis. This motivates the following definition.

Definition 11.1.10. In the context of the binary hypothesis testing problem, a mapping $T\colon \omega \mapsto T(\omega)$ is said to be a **sufficient statistic** if there exists a measurable function ν such that $\Lambda(\omega) = (\nu \circ T)(\omega)$ for all $\omega \in \Omega$.

11.2 Sequential Analysis

Normally, we cannot go lower than u(type I error rate) for a type II error rate. But usually this quantity is high. The answer is that we need more observations, which shifts the error curve since $\mathbf{P}_k\left(\widehat{H} = H_{1-k}\right)$ is lowered. However, observations are expensive! The problem of sequential analysis is to determine how many samples are required to make a decision with a given amount of confidence.

Formally, let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables; H_0 is the hypothesis that $X_i \sim \mathbf{P}_0$, and H_1 is the hypothesis that $X_i \sim \mathbf{P}_1$. Let $(\widehat{H}_n)_{n\in\mathbb{N}}$ be a sequence of tests adapted to the sequence $(X_n)_{n\in\mathbb{N}}$, T is a stopping time, and let the **sequential test** be $\widehat{H}_T$.

This test has type I and type II error rates:

$$\alpha := \mathbf{P}_0\left(\widehat{H}_T = H_1\right) \tag{11.2.1}$$

$$\beta := \mathbf{P}_1\left(\widehat{H}_T = H_0\right). \tag{11.2.2}$$

Remark 11.2.1. We can always assume $\alpha + \beta < 1$. This is because if $\alpha + \beta \geq 1$ then we can always achieve a test which is at least as powerful and uses 0 observations.

For this section, if $\mu \ll \nu$ are probability measures, we let $\Lambda = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ be the likelihood ratio. We also define the "relative entropy"

$$D(\mu \parallel \nu) = \begin{cases} \mathbf{E}_\mu(\log(\Lambda)) & \mu \ll \nu \\ +\infty & \mu \not\ll \nu \end{cases} = \begin{cases} \mathbf{E}_\nu(\Lambda\log(\Lambda)) & \mu \ll \nu \\ +\infty & \mu \not\ll \nu \end{cases}. \tag{11.2.3}$$

By Jensen's inequality,

$$\mathbf{E}_\nu(\Lambda\log(\Lambda)) \geq \mathbf{E}_\nu(\Lambda)\log(\mathbf{E}_\nu(\Lambda)) = 0 \tag{11.2.4}$$

with equality if and only if $\Lambda = 1$, which itself is true if and only if $\mu = \nu$. By an abuse of notation, if $p, q \in (0,1)$, we define

$$D(p \parallel q) = D(\text{Bernoulli}(p) \parallel \text{Bernoulli}(q)) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right). \tag{11.2.5}$$

Theorem 11.2.2. If $\widehat{H}_T$ is a sequential test, with $\alpha + \beta < 1$, then

$$\mathbf{E}_0(T) \geq \frac{D(\alpha \parallel 1-\beta)}{D(\mathbf{P}_0 \parallel \mathbf{P}_1)} \tag{11.2.6}$$

$$\mathbf{E}_1(T) \geq \frac{D(1-\beta \parallel \alpha)}{D(\mathbf{P}_1 \parallel \mathbf{P}_0)}. \tag{11.2.7}$$

Proof. First notice that the inequalities are completely symmetric in $\mathbf{P}_0$ and $\mathbf{P}_1$. So we only need to prove one of the inequalities; we show the second one.

Now there are two cases. If $\mathbf{E}_1(T) = \infty$ then the inequality is clearly true since $\infty >$ anything. So now suppose $\mathbf{E}_1(T) < \infty$.

Now there are two more cases. If $D(\mathbf{P}_1 \parallel \mathbf{P}_0) = \infty$ then the inequality is clearly true since $T \geq 0$. So now suppose $D(\mathbf{P}_1 \parallel \mathbf{P}_0) < \infty$. This implies that $\mathbf{P}_1 \ll \mathbf{P}_0$, which implies that we may define the likelihood ratio $\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0}$.

The same holds for the assumptions $\mathbf{E}_0(T) < \infty$ and $\mathbf{P}_0 \ll \mathbf{P}_1$.

Hence

$$\begin{aligned}
\mathbf{E}_1(|\log(\Lambda(X_1))|) &= \mathbf{E}_1(|\log(\Lambda)|) \qquad\qquad = \mathbf{E}_0(|\Lambda \log(\Lambda)|) && (11.2.8)\\
&\leq \mathbf{E}_0\left(\Lambda \log(\Lambda) + \frac{1}{\mathrm{e}}\right) + \frac{1}{\mathrm{e}} && (11.2.9)\\
&= D(\mathbf{P}_1 \parallel \mathbf{P}_0) + \frac{2}{\mathrm{e}} && (11.2.10)\\
&< \infty. && (11.2.11)
\end{aligned}$$

Then using Wald's identity,

$$\begin{aligned}
\mathbf{E}_1\left(\sum_{i=1}^{T} \log(\Lambda(X_i))\right) &= \mathbf{E}_1(T)\,\mathbf{E}_1(\log(\Lambda(X_1))) && (11.2.12)\\
&= \mathbf{E}_1(T) D(\mathbf{P}_1 \parallel \mathbf{P}_0). && (11.2.13)
\end{aligned}$$

Define $L_n = \prod_{i=1}^{n} \Lambda(X_i)$. We need to show that $\mathbf{E}_1(\log(L_T)) \geq D(1-\beta \parallel \alpha)$, then we are done.

First, we want to find the likelihood ratio $\frac{\mathrm{d}\mathbf{P}_1^n(\cdot|T=n)}{\mathrm{d}\mathbf{P}_0^n(\cdot|T=n)}$. Fix $B \subseteq \mathbb{R}^n$ such that $(x_1, \ldots, x_n) \in B$ implies $T(x_1, \ldots, x_n) = n$. Then

$$\begin{aligned}
\mathbf{P}_1((X_1, \ldots, X_n) \in B \mid T = n)\,\mathbf{P}_1(T = n) &= \mathbf{P}_1((X_1, \ldots, X_n) \in B, T = n) && (11.2.14)\\
&= \mathbf{P}_1((X_1, \ldots, X_n) \in B). && (11.2.15)
\end{aligned}$$

Hence

$$\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0}(x_1, \ldots, x_n \mid T = n) = \frac{\mathrm{d}\,\mathbf{P}_0(T = n)}{\mathrm{d}\,\mathbf{P}_1(T = n)} L_n(x_1, \ldots, x_n). \tag{11.2.16}$$

Fix some randomized threshold η. By Jensen's inequality,

$$\mathbf{E}_1(\log(L_n) \mid L_n \geq \eta, T = n) \geq -\log\big(\mathbf{E}_1\big(L_n^{-1} \bigm| L_n \geq \eta, T = n\big)\big). \tag{11.2.17}$$

Taking a detour, we want to calculate the right-hand side.

$$\begin{aligned}
\mathbf{E}_1\big(L_n^{-1} 1_{\{L_n \geq \eta\}} \bigm| T = n\big) &= \mathbf{P}_1(L_n \geq n \mid T = n)\,\mathbf{E}_1\big(1_{\{L_n \geq \eta\}} L_n^{-1} \bigm| L_n \geq \eta, T = n\big) && (11.2.18)\\
&\quad + \mathbf{P}_1(L_n < n \mid T = n)\,\underbrace{\mathbf{E}_1\big(1_{\{L_n \geq \eta\}} L_n^{-1} \bigm| L_n < \eta, T = n\big)}_{=0} && (11.2.19)\\
&= \mathbf{P}_1(L_n \geq n \mid T = n)\,\mathbf{E}_1\big(1_{\{L_n \geq \eta\}} L_n^{-1} \bigm| L_n \geq \eta, T = n\big) && (11.2.20)\\
&= \mathbf{P}_1(L_n \geq n \mid T = n)\,\mathbf{E}_1\big(L_n^{-1} \bigm| L_n \geq \eta, T = n\big). && (11.2.21)
\end{aligned}$$

Going back to the original computation,

$$\begin{aligned}
&\mathbf{E}_1(\log(L_n) \mid L_n \geq \eta, T = n) && (11.2.22)\\
&\geq -\log\big(\mathbf{E}_1\big(L_n^{-1} \bigm| L_n \geq \eta, T = n\big)\big) && (11.2.23)\\
&= -\log\left(\frac{\mathbf{E}_1\big(1_{\{L_n \geq \eta\}} L_n^{-1} \bigm| T = n\big)\,\mathbf{P}_1(T = n)}{\mathbf{P}_1(L_n \geq \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)} \cdot \frac{\mathbf{P}_0(T = n)}{\mathbf{P}_1(T = n)}\right) && (11.2.24)\\
&= -\log\left(\frac{\mathbf{P}_0(L_n \geq \eta \mid T = n)}{\mathbf{P}_1(L_n \geq \eta \mid T = n)} \cdot \frac{\mathbf{P}_0(T = n)}{\mathbf{P}_1(T = n)}\right) && (11.2.25)\\
&= \log\left(\frac{\mathbf{P}_1(L_n \geq \eta \mid T = n)}{\mathbf{P}_0(L_n \geq \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)}\right). && (11.2.26)
\end{aligned}$$

The same holds for $L_n < \eta$. Thus we can calculate

$$\begin{aligned}
&\mathbf{E}(\log(L_n) \mid T = n) && (11.2.27)\\
&\geq \mathbf{P}_1(L_n \geq \eta \mid T = n) \log\left(\frac{\mathbf{P}_1(L_n \geq \eta \mid T = n)}{\mathbf{P}_0(L_n \geq \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)}\right) && (11.2.28)\\
&\quad + \mathbf{P}_1(L_n < \eta \mid T = n) \log\left(\frac{\mathbf{P}_1(L_n < \eta \mid T = n)}{\mathbf{P}_0(L_n < \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)}\right). && (11.2.29)
\end{aligned}$$

Define $\alpha' = \mathbf{P}_0\left(\widehat{H}_n = H_1 \,\middle|\, T = n\right)$ and $\beta' = \mathbf{P}_1\left(\widehat{H}_n = H_0 \,\middle|\, T = n\right)$. Then the map $t \mapsto D(1 - t\beta' \parallel t\alpha')$ is decreasing in t. Thus by setting $t = 1$ and using Theorem 11.1.7 and the Neyman-Pearson rule,

$$\begin{aligned}
&\mathbf{E}(\log(L_n) \mid T = n) && (11.2.30)\\
&\geq \mathbf{P}_1\left(\widehat{H}_n = H_1 \,\middle|\, T = n\right) \log\left(\frac{\mathbf{P}_1(L_n \geq \eta \mid T = n)}{\mathbf{P}_0(L_n \geq \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)}\right) && (11.2.31)\\
&\quad + \mathbf{P}_1\left(\widehat{H}_n = H_0 \,\middle|\, T = n\right) \log\left(\frac{\mathbf{P}_1(L_n < \eta \mid T = n)}{\mathbf{P}_0(L_n < \eta \mid T = n)} \cdot \frac{\mathbf{P}_1(T = n)}{\mathbf{P}_0(T = n)}\right). && (11.2.32)
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbf{E}_1(\log(L_T)) &= \sum_{n \in \mathbb{N}_0} \mathbf{P}_1(T = n)\, \mathbf{E}_1(\log(L_n) \mid T = n) && (11.2.33)\\
&\geq \mathbf{P}_1\left(\widehat{H}_T = H_1\right) \log\left(\frac{\mathbf{P}_1\left(\widehat{H}_T = H_1\right)}{\mathbf{P}_0\left(\widehat{H}_T = H_1\right)}\right) && (11.2.34)\\
&\quad + \mathbf{P}_1\left(\widehat{H}_T = H_0\right) \log\left(\frac{\mathbf{P}_1\left(\widehat{H}_T = H_0\right)}{\mathbf{P}_0\left(\widehat{H}_T = H_0\right)}\right) && (11.2.35)\\
&= D(1 - \beta \parallel \alpha). && (11.2.36)
\end{aligned}$$

□

The next question to ask is whether we can achieve the lower bound via a threshold test. This motivates the sequential probability ratio test (SPRT). We fix thresholds $0 \leq \eta_0 \leq \eta_1$ and let $T = \inf\{n \in \mathbb{N} \colon L_n \notin (\eta_0, \eta_1)\}$. The SPRT is

$$\widehat{H}_T = \begin{cases} 1 & L_T \geq \eta_1 \\ H_0 & L_T \leq \eta_0 \end{cases}. \tag{11.2.37}$$

Proposition 11.2.3. For $0 \leq \eta_0 < \eta_1$, the SPRT achieves α (type I error rate), β (type II error rate) satisfying

$$\frac{\alpha}{1 - \beta} \leq \frac{1}{\eta_1}, \quad \frac{\beta}{1 - \alpha} \leq \eta_0. \tag{11.2.38}$$

In practice, we want $\alpha, \beta \ll 1$. We choose η_1 such that our desired $\alpha < \eta_1^{-1}$, and η_0 such that our desired $\beta < \eta_0$.

Theorem 11.2.4. Let SPRT achieve type I error rate α and type II error rate β. If $D(\mathbf{P}_0 \parallel \mathbf{P}_1) < \infty$ and $D(\mathbf{P}_1 \parallel \mathbf{P}_0) < \infty$, then

$$\mathbf{E}_0(T) = \frac{D(\alpha \parallel 1 - \beta)}{D(\mathbf{P}_0 \parallel \mathbf{P}_1)}, \quad \mathbf{E}_1(T) = \frac{D(1 - \beta \parallel \alpha)}{D(\mathbf{P}_1 \parallel \mathbf{P}_0)}. \tag{11.2.39}$$