Lecture Notes on Stochastic Processes

Srikanth B. Pai

Madras School of Economics, Chennai

Preface

These notes accompany a one-semester course on stochastic processes for students of economics. The first four chapters form the core of the course: probability spaces and measure (Chapter 1), conditional expectation as a projection, the theory of stochastic processes (Chapter 3), including stationarity, ergodicity, martingales, and the fundamental continuous-time processes, and the theory of Markov chains (Chapter 4).

Contents

CHAPTER 1

Probability Spaces

A **probability model** starts with a random experiment — a procedure whose outcome is uncertain in advance. We record the outcome as a point in a set Ω, called the **sample space**; individual outcomes are its elements. A single coin toss gives $\Omega = \{H, T\}$; three successive tosses give $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

An **event** is a subset $A \subseteq \Omega$: a property of the outcome that is either true or false once the experiment is run. "The first toss lands heads" corresponds to $\{HHH, HHT, HTH, HTT\}$ in the three-toss space. To do probability, we need a rule Pr that assigns to each event a number in $[0, 1]$ representing its likelihood.

The subtlety is that on large sample spaces — such as $\Omega = [0, 1]$ representing a uniformly random real number — it is *impossible* to assign probabilities consistently to every subset of Ω. We therefore single out a collection $\mathcal{F}$ of *admissible* events, called a σ-algebra, that is closed under the natural set operations and large enough to contain all events of practical interest. The triple $(\Omega, \mathcal{F}, \Pr)$ — sample space, σ-algebra, probability measure — is a **probability space**, the central object of this course. This chapter constructs each ingredient in turn.

1. The σ-Algebra

We begin by making precise the collection of sets to which we will assign probabilities.

DEFINITION 1.1 (σ-algebra). Let Ω be a non-empty set called the **sample space**. A collection $\mathcal{F}$ of subsets of Ω is a **σ-algebra** (or **σ-field**) on Ω if the following three axioms hold:

S1. $\Omega \in \mathcal{F}$.

S2. If $A \in \mathcal{F}$ then $A^c := \Omega \setminus A \in \mathcal{F}$.

S3. If $A_1, A_2, \ldots \in \mathcal{F}$ then $\displaystyle\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.

The axioms are deliberately minimal. The next three propositions derive further closure properties, each proof citing only what has already been established.

Proposition 1.2 (Finite union)**.** If $A_1, A_2, \ldots, A_n \in \mathcal{F}$ for some $n \in \mathbb{N}$, then $\displaystyle\bigcup_{k=1}^{n} A_k \in \mathcal{F}$.

PROOF. By Axiom S1, $\Omega \in \mathcal{F}$, and by Axiom S2, $\emptyset = \Omega^c \in \mathcal{F}$. Define the sequence $B_k = A_k$ for $k = 1, \ldots, n$ and $B_k = \emptyset$ for $k > n$. Each $B_k \in \mathcal{F}$, and by Axiom S3, $\bigcup_{k=1}^{\infty} B_k \in \mathcal{F}$. Since $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{n} A_k$, the conclusion follows. □

Proposition 1.3 (Finite intersection)**.** If $A_1, A_2, \ldots, A_n \in \mathcal{F}$ for some $n \in \mathbb{N}$, then $\displaystyle\bigcap_{k=1}^{n} A_k \in \mathcal{F}$.

PROOF. By Axiom S2, $A_k^c \in \mathcal{F}$ for each $k = 1, \dots, n$. By Proposition 1.2, $\bigcup_{k=1}^n A_k^c \in \mathcal{F}$. Applying Axiom S2 once more and De Morgan's law:

$$\bigcap_{k=1}^{n} A_k \;=\; \Big(\bigcup_{k=1}^{n} A_k^c\Big)^c \;\in\; \mathcal{F}. \qquad \square$$

Proposition 1.4 (Set difference)**.** If $A, B \in \mathcal{F}$ then $A \setminus B \in \mathcal{F}$.

PROOF. By Axiom S2, $B^c \in \mathcal{F}$. By Proposition 1.3 (with $n = 2$), $A \cap B^c \in \mathcal{F}$. Since $A \setminus B = A \cap B^c$, the result follows. $\square$

Before pushing the theory further, we pause on concrete examples — the first things to reach for whenever a σ-algebra feels abstract.

Example 1.5 (The power set)**.** The largest possible σ-algebra on any Ω is the **power set** 2^Ω, the collection of *all* subsets. It satisfies S1–S3 trivially: whatever set the axioms demand is already a member. This is the σ-algebra you used — without ever naming it — in an introductory statistics course. Whenever Ω was finite or countable and you spoke of "the probability of *any* event," you were implicitly taking $\mathcal{F} = 2^\Omega$, so that no subset was ever off-limits. Section 1 will show why this convenient default becomes untenable the moment Ω is uncountable.

Example 1.6 (An agent who sees only the first toss)**.** Toss a coin twice, $\Omega = \{HH, HT, TH, TT\}$, but imagine an observer who is shown *only* the first toss. Which events can this agent decide? Exactly those settled by the first toss alone:

$$\mathcal{F}_1 \;=\; \big\{\, \emptyset,\; \{HH, HT\},\; \{TH, TT\},\; \Omega \,\big\}.$$

The agent can answer "was the first toss heads?" ($\{HH, HT\}$) but *cannot* answer "was the second toss heads?", because the set $\{HH, TH\}$ is not on the list.

We verify $\mathcal{F}_1$ is a σ-algebra. **S1:** $\Omega \in \mathcal{F}_1$. **S2:** the complement of each member is again a member ($\emptyset \leftrightarrow \Omega$ and $\{HH, HT\} \leftrightarrow \{TH, TT\}$). **S3:** every union of members is again a member — any union with Ω gives Ω, unions with $\emptyset$ change nothing, and $\{HH, HT\} \cup \{TH, TT\} = \Omega$. All three axioms hold, so $\mathcal{F}_1$ is a genuine σ-algebra with only four elements — far coarser than the $2^4 = 16$ of the power set. A σ-algebra, then, is not just a bookkeeping device: its *size* encodes how much an observer can resolve.

DEFINITION 1.7 (Generated σ-algebra). Let $\mathcal{C} \subseteq 2^\Omega$ be any collection of subsets of Ω. The **σ-algebra generated by $\mathcal{C}$** is

$$\sigma(\mathcal{C}) \;:=\; \bigcap \big\{\, \mathcal{G} \;:\; \mathcal{G} \text{ is a } \sigma\text{-algebra on } \Omega \text{ and } \mathcal{C} \subseteq \mathcal{G} \,\big\}.$$

This is the smallest σ-algebra on Ω containing $\mathcal{C}$.

Example 1.8 ($\mathcal{F}_1$ is generated by a single event)**.** Return to $\Omega = \{HH, HT, TH, TT\}$ and the single event $E = \{\text{first toss is } H\} = \{HH, HT\}$. The four-element $\mathcal{F}_1$ of Example 1.6 did not fall from the sky; it is forced by E alone. We claim

$$\sigma(\{E\}) \;=\; \mathcal{F}_1 \;=\; \big\{\emptyset,\, \{HH, HT\},\, \{TH, TT\},\, \Omega\big\}.$$

$\sigma(\{E\}) \subseteq \mathcal{F}_1$. $\mathcal{F}_1$ is a σ-algebra (Example 1.6) and it contains E. Since $\sigma(\{E\})$ is the *intersection* of *all* σ-algebras containing E, and $\mathcal{F}_1$ is one of the sets being intersected, the intersection is contained in $\mathcal{F}_1$.
$\mathcal{F}_1 \subseteq \sigma(\{E\})$. Any σ-algebra $\mathcal{G}$ containing E must also contain, by S1–S2, the sets Ω, $\emptyset = \Omega^c$, and $E^c = \{TH, TT\}$ — that is, all four elements of $\mathcal{F}_1$. Since *every* $\mathcal{G}$ in the intersection contains $\mathcal{F}_1$, so does their intersection $\sigma(\{E\})$.
Both inclusions give $\sigma(\{E\}) = \mathcal{F}_1$. The moral is worth keeping: naming one event and closing up under the axioms reproduces *exactly* the information structure of the agent who sees only the first toss — no more, no less.

Example 1.9 (Generated σ-algebra on three coin tosses)**.** Take the three-toss sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ and let

$$A = \{\text{first toss is H}\} = \{HHH, HHT, HTH, HTT\},$$
$$B = \{\text{third toss is H}\} = \{HHH, HTH, THH, TTH\}.$$

What is $\sigma(\{A, B\})$?

The key observation is that A and B jointly partition Ω into four pairwise disjoint *building blocks*:

$$P_1 = A \cap B = \{HHH, HTH\}, \qquad P_2 = A \cap B^c = \{HHT, HTT\},$$
$$P_3 = A^c \cap B = \{THH, TTH\}, \qquad P_4 = A^c \cap B^c = \{THT, TTT\}.$$

Any σ-algebra containing A and B must contain their complements and intersections, hence must contain each P_i. Conversely, $\{P_1, P_2, P_3, P_4, \emptyset, \Omega\}$ together with all pairwise and triple unions of the P_i already forms a σ-algebra (it is closed under complements and finite unions by construction, and all unions are finite so S3 is automatic on a finite space). Therefore

$$\sigma(\{A, B\}) = \big\{ \text{all unions of subsets of } \{P_1, P_2, P_3, P_4\} \big\},$$

which has $2^4 = 16$ elements. The familiar events $A = P_1 \cup P_2$, $B = P_1 \cup P_3$, $A^c = P_3 \cup P_4$, $B^c = P_2 \cup P_4$, and $A \cup B = P_1 \cup P_2 \cup P_3$ are all among them.

Remark (Borel sets and non-measurability). Take $\Omega = [0, 1]$ and let $\mathcal{C}$ be the collection of all closed sub-intervals of $[0, 1]$. The generated σ-algebra $\mathcal{B}([0, 1]) := \sigma(\mathcal{C})$ is the **Borel σ-algebra** on $[0, 1]$.

A natural question: does $\mathcal{B}([0, 1])$ contain *every* subset of $[0, 1]$? The answer is **no**.

For students who have taken Real Analysis. The non-trivial fact is that $|\mathcal{B}([0, 1])| = \mathfrak{c} := |\mathbb{R}|$. But $|\mathcal{P}([0, 1])| = 2^{\mathfrak{c}} > \mathfrak{c}$ by Cantor's theorem, so non-Borel subsets must exist.

Remark (What the Borel σ-algebra actually contains). The generating family — closed intervals — looks tame, but closing under the axioms sweeps in far more than intervals. A few consequences worth internalising:

- **Singletons.** A point $\{a\} = [a, a]$ is a degenerate closed interval, hence Borel.
- **Open and half-open intervals.** Starting from a closed interval and removing one or both endpoints — each a singleton — uses only complements and finite intersections, so the result stays Borel.

- **Countable sets.** Any countable set $\{a_1, a_2, \ldots\}$ is a countable union of singletons, hence Borel by Axiom S3. In particular the rationals $\mathbb{Q} \cap [0,1]$ are Borel — even though they contain no interval at all.

So a great many "interesting" sets are Borel. The striking thing, reinforcing Section 1, is how *hard* it is to point to a set that is not: no explicit formula exists, and the only constructions invoke the Axiom of Choice.

REMARK (Why stop at Borel? The extension theorem). If a general Borel set can be too intricate to picture, how could we ever assign it a probability? This is the crux — and, unknowingly, it is exactly the position you were in during a first statistics course.
The resolution is a theorem we will use but not prove (Carathéodory's extension theorem): a probability specified consistently on the *generating* intervals extends *uniquely* to every Borel set. You never assign probabilities to pathological sets by hand; you fix them on intervals — by declaring a length, a density, or a CDF — and the theorem manufactures the rest. This is precisely why, in practice, a continuous distribution is pinned down by its behaviour on intervals and nothing more is required. "Uniform on $[0,1]$" is just the length measure on intervals, extended.

REMARK (Lebesgue completion — a footnote for later). One blemish survives. It can happen that a Borel set N has probability 0 yet possesses a subset $S \subseteq N$ that is *not* itself Borel. Intuition insists S should be negligible too, but strictly speaking $\Pr(S)$ is then undefined. The remedy is to enlarge $\mathcal{B}$ by declaring every subset of a null set measurable (with probability 0) and re-closing under the axioms; the result is the **Lebesgue σ-algebra** — the Borel sets plus these harmless additions. We will not need this refinement, and mention it only so that the phrase "Lebesgue-measurable" is not mysterious when you meet it elsewhere.

2. Limit Superior and Measurability

Two concepts appear repeatedly throughout the course: the *limit superior* of a sequence of sets ("the event that occurs infinitely often") and the question of whether such a set is measurable. We develop both here.

Solved Exercise 1.10 (Tail sets and limit superior)**.** Let $(A_n)_{n\geq 1}$ be a sequence of subsets of Ω. For each $n \geq 1$ define the **tail sets**

$$C_n := \bigcup_{k=n}^{\infty} A_k, \qquad D_n := \bigcap_{k=n}^{\infty} A_k,$$

and the **limit superior** and **limit inferior**:

$$\limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} C_n, \qquad \liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty} D_n.$$

Claim. The sequence (C_n) is decreasing and (D_n) is increasing. Moreover,

$$\omega \in \limsup_{n} A_n \iff \omega \in A_n \text{ for infinitely many } n.$$

PROOF. **Step 1: Monotonicity of (C_n) and (D_n).**
For (C_n): suppose $\omega \in C_{n+1} = \bigcup_{k\geq n+1} A_k$. Then $\omega \in A_k$ for some index k satisfying $k \geq n+1$. Since $k \geq n+1 > n$, that same k satisfies $k \geq n$, so $\omega \in \bigcup_{k\geq n} A_k = C_n$. Hence $C_{n+1} \subseteq C_n$, and the sequence (C_n) is *decreasing*.

For (D_n): suppose $\omega \in D_n = \bigcap_{k\geq n} A_k$. Then $\omega \in A_k$ for *every* $k \geq n$. In particular, $\omega \in A_k$ for every $k \geq n+1$, so $\omega \in \bigcap_{k\geq n+1} A_k = D_{n+1}$. Hence $D_n \subseteq D_{n+1}$, and (D_n) is *increasing*.

Step 2: Characterisation of $\limsup$ **— forward direction** $(\Rightarrow)$**.**

Goal. Suppose $\omega \in \bigcap_{n=1}^{\infty} C_n$. We must show $\omega \in A_n$ for infinitely many n.

Let $S := \{n \geq 1 : \omega \in A_n\}$ be the set of all indices where ω appears. We argue by contradiction: suppose S is *finite*.

- If $S = \emptyset$, set $N_0 = 1$.
- If $S \neq \emptyset$, set $N_0 = (\max S) + 1$.

In both cases, $\omega \notin A_k$ for every $k \geq N_0$ (in the first case because S is empty; in the second because every element of S is at most $N_0 - 1$). Therefore ω is not in any of the sets $A_{N_0}, A_{N_0+1}, \ldots$, so

$$\omega \notin \bigcup_{k=N_0}^{\infty} A_k = C_{N_0}.$$

But $\omega \in \bigcap_{n=1}^{\infty} C_n$ means ω belongs to *every* C_n; taking $n = N_0$ gives $\omega \in C_{N_0}$. This is a contradiction.

Therefore the assumption that S is finite is false: S is infinite, meaning $\omega \in A_n$ for infinitely many n.

Step 3: Characterisation of $\limsup$ **— backward direction** $(\Leftarrow)$**.**

Goal. Suppose $\omega \in A_n$ for infinitely many n. We must show $\omega \in \bigcap_{k=1}^{\infty} C_k$.

Since infinitely many indices n satisfy $\omega \in A_n$, we may list them in strictly increasing order:

$$n_1 < n_2 < n_3 < \cdots, \qquad \omega \in A_{n_j} \text{ for every } j \geq 1.$$

A strictly increasing sequence of positive integers satisfies $n_j \geq j$ for all $j \geq 1$. (Proof: $n_1 \geq 1$; if $n_j \geq j$ then $n_{j+1} \geq n_j + 1 \geq j + 1$.)

Fix any $k \geq 1$. Since $n_k \geq k$, the set A_{n_k} appears among the sets in the union $C_k = \bigcup_{l=k}^{\infty} A_l$. Since $\omega \in A_{n_k}$, we conclude $\omega \in C_k$.

Since $k \geq 1$ was arbitrary, $\omega \in C_k$ for every k, giving $\omega \in \bigcap_{k=1}^{\infty} C_k = \limsup_n A_n$. □

Corollary 1.11 (Measurability of $\limsup$)**.** Let $(A_n)_{n\geq 1}$ be a sequence in $\mathcal{F}$. Then $\limsup_n A_n \in \mathcal{F}$.

Proof. By Solved Exercise 1.10, $\limsup_n A_n = \bigcap_{n=1}^{\infty} C_n$ where $C_n = \bigcup_{k=n}^{\infty} A_k \in \mathcal{F}$ by Axiom S3. By De Morgan's law, $\bigcap_{n=1}^{\infty} C_n = \left(\bigcup_{n=1}^{\infty} C_n^c\right)^c$, and each $C_n^c \in \mathcal{F}$ (Axiom S2), so $\bigcup_n C_n^c \in \mathcal{F}$ (Axiom S3), hence $\limsup_n A_n \in \mathcal{F}$ (Axiom S2). □

Exercise 1.12 (Limit inferior and containment)**.** Using the notation of Solved Exercise 1.10, show:

(a) $\omega \in \liminf_n A_n$ if and only if $\omega \in A_n$ for *all but finitely many* n.

(b) $\liminf_n A_n \subseteq \limsup_n A_n$.

Hint for (a): $\omega \in \bigcup_n D_n$ iff $\omega \in D_N$ for some N, i.e. $\omega \in A_k$ for all $k \geq N$ — so ω misses A_k only for $k < N$, finitely many.

Hint for (b): "All but finitely many" implies "infinitely many" when the index set is infinite.

[Answer: see §8.]

3. Probability Measures

With a σ-algebra in place, the third ingredient of the probability space is a function that assigns likelihoods to events.

DEFINITION 1.13 (Probability measure and probability space). Let $(\Omega, \mathcal{F})$ be a measurable space (see Definition 1.1). A function $\Pr : \mathcal{F} \to [0,1]$ is a **probability measure** if it satisfies:

P1. $\Pr(\Omega) = 1$.

P2. For every sequence of *pairwise disjoint* sets $A_1, A_2, \ldots \in \mathcal{F}$ (meaning $A_i \cap A_j = \emptyset$ for all $i \neq j$):

$$\Pr\left(\bigsqcup_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} \sum_{k=1}^{n} \Pr(A_k).$$

The triple $(\Omega, \mathcal{F}, \Pr)$ is a **probability space**.

Proposition 1.14 (Probability of the empty set)**.** $\Pr(\emptyset) = 0$.

PROOF. Set $A_1 = \Omega$ and $A_k = \emptyset$ for all $k \geq 2$. These sets are pairwise disjoint and $\bigsqcup_{k=1}^{\infty} A_k = \Omega$. By P2:

$$\Pr(\Omega) = \lim_{n\to\infty} \sum_{k=1}^{n} \Pr(A_k) = \Pr(\Omega) + \lim_{n\to\infty} (n-1)\,\Pr(\emptyset).$$

By P1, $\Pr(\Omega) = 1$, so $(n-1)\Pr(\emptyset) \to 0$ as $n \to \infty$, which forces $\Pr(\emptyset) = 0$. □

Proposition 1.15 (Finite additivity)**.** For pairwise disjoint $A_1, \ldots, A_n \in \mathcal{F}$:

$$\Pr\left(\bigsqcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \Pr(A_k).$$

PROOF. Set $B_k = A_k$ for $k = 1, \ldots, n$ and $B_k = \emptyset$ for $k > n$. The sets $B_1, B_2, \ldots$ are pairwise disjoint and $\bigsqcup_{k=1}^{\infty} B_k = \bigsqcup_{k=1}^{n} A_k$. By P2:

$$\Pr\left(\bigsqcup_{k=1}^{n} A_k\right) = \lim_{m\to\infty} \sum_{k=1}^{m} \Pr(B_k) = \sum_{k=1}^{n} \Pr(A_k) + \lim_{m\to\infty} \sum_{k=n+1}^{m} \Pr(\emptyset).$$

By Proposition 1.14, $\Pr(\emptyset) = 0$, so the tail vanishes. □

Proposition 1.16 (Inclusion–exclusion for a pair)**.** For any $A, B \in \mathcal{F}$:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

PROOF. Write $A \cup B$ as a disjoint union: $A \cup B = A \sqcup (B \setminus A)$. By Proposition 1.4, $B \setminus A \in \mathcal{F}$. Applying Proposition 1.15 (with $n = 2$):

$$(*) \qquad \Pr(A \cup B) = \Pr(A) + \Pr(B \setminus A).$$

Similarly, $B = (A \cap B) \sqcup (B \setminus A)$ is a disjoint union; note $A \cap B \in \mathcal{F}$ by Proposition 1.3. Applying Proposition 1.15 again:

$$(**) \qquad \Pr(B) = \Pr(A \cap B) + \Pr(B \setminus A), \quad \text{so} \quad \Pr(B \setminus A) = \Pr(B) - \Pr(A \cap B).$$

Substituting $(**)$ into $(*)$ gives the result. □

Corollary 1.17 (Monotonicity and subtraction)**.** If $B \subseteq A$ and $A, B \in \mathcal{F}$, then $\Pr(B) \leq \Pr(A)$. More precisely, $\Pr(A \setminus B) = \Pr(A) - \Pr(B)$.

PROOF. Since $B \subseteq A$, we have $A = B \sqcup (A \setminus B)$. By Proposition 1.4, $A \setminus B \in \mathcal{F}$. Applying Proposition 1.15 (with $n = 2$):

$$\Pr(A) = \Pr(B) + \Pr(A \setminus B),$$

so $\Pr(A \setminus B) = \Pr(A) - \Pr(B)$. Since $\Pr(A \setminus B) \geq 0$, we also obtain $\Pr(B) \leq \Pr(A)$. □

4. Exercises on Probability Measures

Exercise 1.18 (Union bound)**.** Prove that for any $A_1, \ldots, A_n \in \mathcal{F}$:

$$\Pr\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \Pr(A_k).$$

Hint: Use induction on n. The base case $n = 1$ is an equality. For the inductive step, write $\bigcup_{k=1}^{n+1} A_k = \left(\bigcup_{k=1}^{n} A_k\right) \cup A_{n+1}$ and apply Proposition 1.16; then use $\Pr(\,\cdot\,) \geq 0$ to drop the intersection term.

[Answer: see §8.]

5. Continuity of Probability

Solved Exercise 1.19 (Continuity for expanding sets)**.** Let $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ be an increasing sequence of sets in $\mathcal{F}$. Then

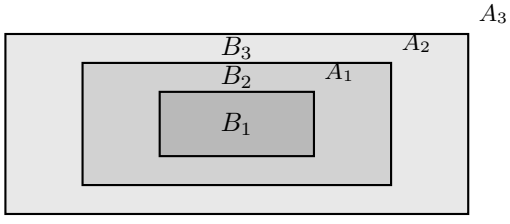
$$\Pr\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k \to \infty} \Pr(A_k).$$

Strategy. Since $A_1 \subseteq A_2 \subseteq \cdots$, each set contains the previous one, so the finite union up to step n is just the largest set: $\bigcup_{k=1}^{n} A_k = A_n$. As $n \to \infty$, A_n keeps growing and fills up $\bigcup_{k=1}^{\infty} A_k$. The result therefore says

$$\Pr\left(\lim_{n \to \infty} A_n\right) = \lim_{n \to \infty} \Pr(A_n),$$

i.e., we may exchange the probability sign and the limit — this is why the result is called *continuity of probability from below*.

To prove it, we need to use countable additivity (P2), which applies only to *disjoint* unions. The idea is to peel off the new part added at each step: set $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n \geq 2$. The sets B_n are pairwise disjoint and together cover exactly the same points as the A_n.



PROOF. **Step 1: Disjointify.** Define $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n \geq 2$. By Proposition 1.4, each $B_n \in \mathcal{F}$.

Step 2: Pairwise disjointness. Take $m < n$. Since $B_n = A_n \setminus A_{n-1}$, we have $B_n \subseteq A_{n-1}^c$. Since $m \leq n - 1$, we have $B_m \subseteq A_m \subseteq A_{n-1}$, so $B_m \cap B_n \subseteq A_{n-1} \cap A_{n-1}^c = \emptyset$.

Step 3: Same union. We claim $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$. ($\subseteq$): Each $B_k \subseteq A_k$, so $\bigcup B_k \subseteq \bigcup A_k$. ($\supseteq$): Let $\omega \in \bigcup_{k=1}^{\infty} A_k$ and let n be the *smallest* index with $\omega \in A_n$. If $n = 1$, then $\omega \in B_1$. If $n \geq 2$, then $\omega \notin A_{n-1}$, so $\omega \in A_n \setminus A_{n-1} = B_n$. Either way, $\omega \in \bigcup_{k=1}^{\infty} B_k$.

Step 4: Apply countable additivity. The sets $B_1, B_2, \ldots$ are pairwise disjoint (Step 2) and $\bigsqcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$ (Step 3). By P2:

$$\Pr\left(\bigcup_{k=1}^{\infty} A_k\right) = \Pr\left(\bigsqcup_{k=1}^{\infty} B_k\right) = \lim_{n\to\infty} \sum_{k=1}^{n} \Pr(B_k).$$

Step 5: Telescope the partial sum. For $k \geq 2$, since $A_{k-1} \subseteq A_k$, Corollary 1.17 gives $\Pr(B_k) = \Pr(A_k \setminus A_{k-1}) = \Pr(A_k) - \Pr(A_{k-1})$. Therefore:

$$\sum_{k=1}^{n} \Pr(B_k) = \Pr(A_1) + \sum_{k=2}^{n} \big[\Pr(A_k) - \Pr(A_{k-1})\big] = \Pr(A_n).$$

Conclusion. Combining Steps 4 and 5: $\Pr\big(\bigcup_{k=1}^{\infty} A_k\big) = \lim_{n\to\infty} \Pr(A_n)$. □

Exercise 1.20 (Continuity for contracting sets)**.** Let $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ be a decreasing sequence of sets in $\mathcal{F}$. Show that

$$\Pr\left(\bigcap_{k=1}^{\infty} A_k\right) \;=\; \lim_{k\to\infty} \Pr(A_k).$$

Hint: Take complements. Since the A_n are decreasing, the complements A_n^c are increasing (why?). Apply Solved Exercise 1.19 to (A_n^c), then use De Morgan's law and Corollary 1.17 to convert back to the A_n.

[Answer: see §8.]

Theorem 1.21 (First Borel–Cantelli Lemma)**.** Let $(A_n)_{n\geq 1}$ be a sequence in $\mathcal{F}$ such that $\sum_{n=1}^{\infty} \Pr(A_n) < \infty$. Then

$$\Pr\big(A_n \text{ i.o.}\big) \;:=\; \Pr\left(\limsup_{n\to\infty} A_n\right) = 0.$$

Proof. By Solved Exercise 1.10, $\limsup_n A_n = \bigcap_{k=1}^{\infty} C_k$ where $C_k := \bigcup_{l=k}^{\infty} A_l \in \mathcal{F}$ (Axiom S3), and $\bigcap_k C_k \in \mathcal{F}$ by Corollary 1.11. Since $\bigcap_{k=1}^{\infty} C_k \subseteq C_k$, by Corollary 1.17 and subadditivity (P2):

$$0 \;\leq\; \Pr\left(\limsup_{n} A_n\right) = \Pr\left(\bigcap_{k=1}^{\infty} C_k\right) \leq \Pr(C_k) \leq \sum_{l=k}^{\infty} \Pr(A_l) \qquad \text{for every } k \geq 1.$$

Since $\sum_{l=k}^{\infty} \Pr(A_l) = \sum_{l=1}^{\infty} \Pr(A_l) - \sum_{l=1}^{k-1} \Pr(A_l) \to 0$ as $k \to \infty$, we get $\Pr\big(\limsup_n A_n\big) \leq 0$, hence $\Pr\big(\limsup_n A_n\big) = 0$. □

6. Random Variables

Abstract outcomes $\omega \in \Omega$ are often non-numerical: sequences of coin tosses, hands of cards, paths of a particle. To compute with them we need to extract numbers from outcomes — "how many heads in ten tosses?", "what is the payout?" — via functions $X : \Omega \to \mathbb{R}$. The only requirement is that X be *compatible with* $\mathcal{F}$: for each real threshold b, the event "X does not exceed b" must be one of the admissible events, so that we may speak of its probability. This single, concrete demand is the whole definition.

DEFINITION 1.22 (Random variable). Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is a **random variable** (equivalently, an **$\mathcal{F}$-measurable function**) if

$$\{X \le b\} \;:=\; \{\omega \in \Omega : X(\omega) \le b\} \;\in\; \mathcal{F} \qquad \text{for every } b \in \mathbb{R}.$$

That is, every threshold event is admissible. This — and only this — is what we ever check.

REMARK. The set $\{\omega \in \Omega : X(\omega) \in B\}$ is standardly abbreviated to $\{X \in B\}$, which is precisely the preimage $X^{-1}(B)$. Similarly, $\{X = x\}$ means $X^{-1}(\{x\})$ and $\{X \le x\}$ means $X^{-1}((-\infty, x])$. This shorthand lets us write $\Pr(X \le x)$ without naming ω explicitly.

Thresholds are all we check — but do they pin down the probability of *every* question one might ask about X, such as "does X land in this set B of values?" They do, and the reason is one elementary fact about pre-images.

REMARK (Pre-images respect every set operation). For any function $X : \Omega \to \mathbb{R}$, write $X^{-1}(B) := \{\omega : X(\omega) \in B\}$. Then for arbitrary subsets $B, B_i \subseteq \mathbb{R}$,

$$X^{-1}(B^c) = \big(X^{-1}(B)\big)^c, \quad X^{-1}\Big(\bigcup_i B_i\Big) = \bigcup_i X^{-1}(B_i), \quad X^{-1}\Big(\bigcap_i B_i\Big) = \bigcap_i X^{-1}(B_i).$$

In words: the pre-image commutes with complements, unions, and intersections (each is a one-line chase of the definition, left as a school exercise). Note X^{-1} is *not* an inverse function — it is just notation for the set of points mapped into B. This single fact is what makes measurability easy to verify.

Proposition 1.23 (Thresholds determine every Borel event)**.** Let $X : \Omega \to \mathbb{R}$, and let

$$\mathcal{B}(\mathbb{R}) \;=\; \sigma\big(\{(-\infty, b] : b \in \mathbb{R}\}\big)$$

be the **Borel σ-algebra**, the smallest σ-algebra on $\mathbb{R}$ containing every half-line $(-\infty, b]$. Then X is a random variable in the sense of Definition 1.22 if and only if

$$X^{-1}(B) \;:=\; \{\omega : X(\omega) \in B\} \;\in\; \mathcal{F} \qquad \text{for } \textit{every} \text{ Borel set } B \in \mathcal{B}(\mathbb{R}).$$

PROOF. If $X^{-1}(B) \in \mathcal{F}$ for every Borel B, then in particular each half-line $(-\infty, b]$ is Borel, so $\{X \le b\} = X^{-1}((-\infty, b]) \in \mathcal{F}$: the threshold condition of Definition 1.22 holds.

Conversely, suppose $\{X \le b\} \in \mathcal{F}$ for every b, and set

$$\mathcal{G} \;:=\; \big\{\, B \subseteq \mathbb{R} : X^{-1}(B) \in \mathcal{F} \,\big\}.$$

By Section 6 the pre-image turns complements into complements and countable unions into countable unions *inside* $\mathcal{F}$; since $\mathcal{F}$ is a σ-algebra, $\mathcal{G}$ is closed under both operations and contains $\mathbb{R}$. Hence $\mathcal{G}$ is itself a σ-algebra. By hypothesis it contains every half-line $(-\infty, b]$. But $\mathcal{B}(\mathbb{R})$ is the *smallest* σ-algebra containing those half-lines, so $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{G}$. That is precisely $X^{-1}(B) \in \mathcal{F}$ for every Borel B. □

REMARK (You may forget the Borel σ-algebra). Advanced texts *define* a random variable by the condition on the right — $X^{-1}(B) \in \mathcal{F}$ for every Borel B — for which the Borel σ-algebra on the range is unavoidable. Proposition 1.23 shows this is equivalent to our threshold definition, so for this course you need not carry the Borel machinery in your head: confirming the events $\{X \le b\}$ is always enough. One special case is worth recording. If X takes only countably

many values $\{x_1, x_2, \ldots\}$, it suffices to check $\{X = x_k\} \in \mathcal{F}$ for each k, because every threshold event $\{X \le b\} = \bigcup_{k:\, x_k \le b} \{X = x_k\}$ is then a countable union of these.

So the working picture is simple: to test whether a function is a random variable, run through the threshold events $\{X \le b\}$ and confirm each lies in $\mathcal{F}$. Nothing more is ever required.

Example 1.24 (Measurable and non-measurable functions)**.** Toss a fair coin twice so that $\Omega = \{HH, HT, TH, TT\}$. Let $\mathcal{F}_1 = \sigma(\{HH, HT\})$ be the first-toss σ-algebra of Examples 1.6 and 1.8:

$$\mathcal{F}_1 = \big\{\emptyset,\ \{HH, HT\},\ \{TH, TT\},\ \Omega\big\}.$$

Think of $\mathcal{F}_1$ as the information available after seeing *only* the first toss.
Define two indicator functions:

$$X(\omega) = \begin{cases} 1 & \text{first toss is H,} \\ 0 & \text{first toss is T,} \end{cases}$$

$$Y(\omega) = \begin{cases} 1 & \text{second toss is H,} \\ 0 & \text{second toss is T.} \end{cases}$$

X is $\mathcal{F}_1$-measurable. The preimages of the two values are:

$$\{X = 1\} = \{HH, HT\} \in \mathcal{F}_1, \qquad \{X = 0\} = \{TH, TT\} \in \mathcal{F}_1.$$

Both lie in $\mathcal{F}_1$, so X is a random variable.
Y is *not* $\mathcal{F}_1$-measurable. We have $\{Y = 1\} = \{HH, TH\}$, but $\{HH, TH\} \notin \mathcal{F}_1$. The measurability condition fails for the value 1.
The distinction is informational: $\mathcal{F}_1$ encodes knowledge of the first toss only. X, which depends solely on the first toss, is visible to $\mathcal{F}_1$. Y, which depends on the second toss, is not.

The same failure shows up through the threshold test of Definition 1.22. The *total* head-count $N(\omega) = X(\omega) + Y(\omega)$ is not $\mathcal{F}_1$-measurable: taking $b = 0$ gives $\{N \le 0\} = \{TT\}$, a singleton outside $\mathcal{F}_1$. One failing value of b is enough — and it fails for the honest reason that counting *both* tosses requires seeing the second, which the agent cannot. Meanwhile X itself is exactly the indicator of "first toss heads," and up to relabelling it is the *only* non-constant random variable this coarse σ-algebra admits.

Example 1.25 (Continuous functions are random variables*)**.** Now let the sample space carry a Borel structure of its own: $\Omega = \mathbb{R}$ (or any interval) with $\mathcal{F} = \mathcal{B}(\mathbb{R})$. Then *every continuous* $X : \mathbb{R} \to \mathbb{R}$ is a random variable.
The argument is one line once continuity is stated topologically. $\mathcal{B}(\mathbb{R})$ is generated by the open intervals, so it contains every open set. A function is continuous exactly when the pre-image of every open set is open — and open sets are Borel. Hence $X^{-1}(\text{open}) \in \mathcal{B}(\mathbb{R})$; the same generated-$\sigma$-algebra reasoning as in Proposition 1.23, now with open sets as the generating family, upgrades this to $X^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for *all* Borel B.
This is why elementary courses never fuss over measurability: the functions one writes down — polynomials, exponentials, anything you can graph — are continuous, hence automatically random variables. Measurability only bites when the σ-algebra is deliberately coarse, as with $\mathcal{F}_1$ above. (The topological fact used here belongs to a real-analysis course; we invoke it and move on.)

Example 1.26 (The indicator random variable*)**.** Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and $A \in \mathcal{F}$ an event. The **indicator** of A is

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

It records a single bit — did A occur? — and is the archetypal Bernoulli random variable (Example 1.32). It will recur relentlessly throughout the course, so the little verification below is worth doing once, slowly.

$\mathbf{1}_A$ is a random variable. Write $X = \mathbf{1}_A$ and run through the thresholds of Definition 1.22:

$$\{X \le b\} = \begin{cases} \emptyset, & b < 0, \\ A^c, & 0 \le b < 1, \\ \Omega, & b \ge 1. \end{cases}$$

Each set lies in $\mathcal{F}$ (A^c by Axiom S2), so $\mathbf{1}_A$ is measurable. Notice what carried the proof: the measurability of the *function* is nothing but the measurability of the *event* A it is built from.

Exercise 1.27 (Algebra of indicators)**.** For events $A, B \in \mathcal{F}$, prove:

(a) $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A \cap B}$: the product of indicators is the indicator of the intersection.

(b) If $A \subseteq B$ then $\mathbf{1}_A \le \mathbf{1}_B$ pointwise on Ω.

These little identities are the reason indicators can be slipped into any calculation and manipulated like numbers; we will lean on them constantly, including when we build conditional expectation.

Hint: every function in sight takes only the values 0 *and* 1*, so check both sides agree at each* ω *by splitting on whether* ω *lies in* A*, in* B*, in both, or in neither.*

[Answer: see §8.]

Exercise 1.28 (Indicator of a partition)**.** Let $A \in \mathcal{F}$, and let $A_1, \dots, A_n \in \mathcal{F}$ be a **partition** of A: they are pairwise disjoint, $A_i \cap A_j = \emptyset$ for $i \ne j$, and $\bigcup_{i=1}^n A_i = A$. Prove that

$$\mathbf{1}_A \;=\; \sum_{i=1}^{n} \mathbf{1}_{A_i}$$

as functions on Ω. (The same holds for a countable partition $A = \bigsqcup_{i \ge 1} A_i$.)

Hint: fix ω*. If* $\omega \notin A$ *both sides are* 0*; if* $\omega \in A$*, it lies in exactly one* A_i*.*

[Answer: see §8.]

Exercise 1.29 (Killing the level set)**.** Let X be a random variable and $c \in \mathbb{R}$. Prove that

$$X\,\mathbf{1}_{\{X=c\}} \;=\; c\,\mathbf{1}_{\{X=c\}};$$

in particular $X\,\mathbf{1}_{\{X=0\}} = 0$, the zero function on Ω.

Hint: fix ω *and split on whether* $X(\omega) = c$*.*

[Answer: see §8.]

Measurability ensures $\Pr(X \in B)$ is well-defined for every Borel set B. For a discrete random variable the entire probability law is captured by listing the probabilities of individual values.

DEFINITION 1.30 (Discrete random variable and pmf). A random variable X is **discrete** if it takes values in a countable set $S = \{x_1, x_2, \ldots\} \subseteq \mathbb{R}$. Its **probability mass function (pmf)** is

$$p_X(x_k) := \Pr(X = x_k), \qquad k = 1, 2, \ldots$$

The events $\{X = x_k\}$ are pairwise disjoint and cover Ω, so countable additivity (P2) gives the **normalization** $\sum_k p_X(x_k) = 1$. The pmf recovers all probabilities: $\Pr(X \in B) = \sum_{k:\, x_k \in B} p_X(x_k)$ for any $B \in \mathcal{B}(\mathbb{R})$.

Exercise 1.31 (A discrete random variable is a combination of indicators)**.** Let N be a discrete random variable taking values in a countable set $S \subseteq \mathbb{R}$. Prove that

$$N = \sum_{k \in S} k\, \mathbf{1}_{\{N=k\}}$$

as functions on Ω. Deduce that if N is measurable with respect to a σ-algebra $\mathcal{G}$ (so that $\{N = k\} \in \mathcal{G}$ for every $k \in S$), then N is a countable linear combination of $\mathcal{G}$-measurable indicators.

Hint: fix ω and set $k = N(\omega)$; use Exercise 1.29 for each term, and note only the term $k = N(\omega)$ survives.

[Answer: see §8.]

The four distributions below are the workhorses of discrete probability.

Example 1.32 (Bernoulli(p) distribution)**.** A single trial succeeds with probability $p \in [0, 1]$; set $X = 1$ for success and $X = 0$ for failure. The pmf on $S = \{0, 1\}$ is:

$$p_X(1) = p, \qquad p_X(0) = 1 - p.$$

Normalization: $p + (1 - p) = 1$.

Example 1.33 (Binomial(n, p) distribution)**.** Run n independent Bernoulli(p) trials; let X count the successes. Then X takes values in $S = \{0, 1, \ldots, n\}$ with

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Normalization: $\sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1$ by the Binomial Theorem.

Example 1.34 (Geometric(p) distribution)**.** Repeat independent Bernoulli(p) trials; let X be the index of the *first* success. Then X takes values in $S = \{1, 2, 3, \ldots\}$ with

$$p_X(k) = (1 - p)^{k-1} p, \qquad k = 1, 2, 3, \ldots$$

Normalization: $\sum_{k=1}^{\infty} (1 - p)^{k-1} p = p \cdot \dfrac{1}{1 - (1 - p)} = 1$ (geometric series).

Example 1.35 (Poisson(λ) distribution)**.** Let $\lambda > 0$. A discrete random variable X with values in $S = \{0, 1, 2, \ldots\}$ has the Poisson(λ) distribution if

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Normalization: $\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$. It arises as the limit of $\mathrm{Bin}(n, \lambda/n)$ as $n \to \infty$, modelling counts of rare events occurring at average rate λ.

DEFINITION 1.36 (Expectation). The **expectation** (or **mean**) of a discrete random variable X with pmf p_X on $\{x_1, x_2, \ldots\}$ is

$$\mathbb{E}[X] := \sum_k x_k\, p_X(x_k),$$

provided $\sum_k |x_k|\, p_X(x_k) < \infty$. It is the probability-weighted average of the values of X.
For the Bernoulli(p) case: $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$.

REMARK (Expectation of an indicator is probability). Apply the definition to the indicator $\mathbf{1}_A$ of Example 1.26, which takes the value 1 on A and 0 on A^c:

$$\mathbb{E}[\mathbf{1}_A] = 1 \cdot \Pr(A) + 0 \cdot \Pr(A^c) = \Pr(A).$$

The expectation of an indicator *is* the probability of its event. This tiny identity is the hinge that lets us convert probabilities into expectations and back at will — a move we will make on almost every page to come.

DEFINITION 1.37 (Variance). The **variance** of X measures spread around the mean $\mu = \mathbb{E}[X]$:

$$\mathrm{Var}(X) := \mathbb{E}\big[(X-\mu)^2\big] = \sum_k (x_k - \mu)^2\, p_X(x_k).$$

Expanding $(x_k - \mu)^2 = x_k^2 - 2\mu x_k + \mu^2$ and using $\sum_k p_X(x_k) = 1$ gives the **computational formula**

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

For Bernoulli(p): $\mathbb{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$, so $\mathrm{Var}(X) = p - p^2 = p(1-p)$.
The **standard deviation** $\sigma_X = \sqrt{\mathrm{Var}(X)}$ has the same units as X and is often more interpretable.

REMARK. Summary of means and variances for the four standard families (derivations via series manipulations or the Binomial Theorem):

Distribution	Support	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	$\{0, 1\}$	p	$p(1-p)$
Binomial(n, p)	$\{0, 1, \ldots, n\}$	np	$np(1-p)$
Geometric(p)	$\{1, 2, 3, \ldots\}$	$1/p$	$(1-p)/p^2$
Poisson(λ)	$\{0, 1, 2, \ldots\}$	λ	λ

7. Distribution Functions

The pmf describes a discrete random variable by cataloguing its point masses, but it goes silent the moment X spreads continuously over an interval, where every single value has probability 0. One object copes with both cases at once — and, remarkably, reduces the entire theory of a random variable to ordinary calculus.

DEFINITION 1.38 (Cumulative distribution function). The **cumulative distribution function** (CDF) of a random variable X is

$$F_X(x) \;:=\; \Pr(X \le x) \;=\; \Pr\big(X^{-1}((-\infty, x])\big), \qquad x \in \mathbb{R}.$$

This is defined for *every* x precisely because X is a random variable: Definition 1.22 guarantees $\{X \le x\} \in \mathcal{F}$, so Pr may legitimately be applied. Measurability is not red tape here — it is the very thing that lets F_X exist. (Whenever you write $\Pr(\cdot)$, the object inside must be an event; $\{X \le x\}$ qualifies, and that is the whole point.)

Theorem 1.39 (Characterisation of CDFs)**.** A function $F : \mathbb{R} \to [0,1]$ is the CDF of some random variable if and only if it satisfies:

- **C1.** $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
- **C2.** F is non-decreasing;
- **C3.** F is right-continuous: $\lim_{y \downarrow x} F(y) = F(x)$ for every x.

The two directions of this theorem are wildly unequal in difficulty, and the gap is exactly where the abstract machinery of this chapter earns its keep.

The easy direction ($X \Rightarrow F$). Given a random variable, its CDF satisfies C1–C3. Monotonicity follows from $\{X \le x\} \subseteq \{X \le y\}$ when $x \le y$ together with Corollary 1.17; the limits C1 and the right-continuity C3 follow from the continuity of probability (Solved Exercise 1.19 and Exercise 1.20) applied to nested events such as $\{X \le x + 1/n\}$. A school student who knows how to take one-sided limits can even *state* these properties. This half is a genuine exercise (Exercise 1.40).

The hard direction ($F \Rightarrow X$). Handed only a function F with the three calculus properties, one must *exhibit* a probability space $(\Omega, \mathcal{F}, \Pr)$ and a random variable X on it whose CDF is exactly F. This is less mysterious than it first appears: it needs no new axiom, only the existence of **Lebesgue measure** on $[0,1]$ — itself nothing but the extension of length from intervals (Section 1). Take

$$(\Omega, \mathcal{F}, \Pr) \;=\; \big([0,1],\ \mathcal{B}([0,1]),\ \text{Lebesgue measure}\big)$$

and define the **quantile**

$$X(\omega) \;:=\; \inf\{\, x \in \mathbb{R} : F(x) \ge \omega \,\}, \qquad \omega \in (0,1).$$

Right-continuity (C3) and monotonicity (C2) of F give the key equivalence $X(\omega) \le b \iff \omega \le F(b)$, so $\{X \le b\} = \{\omega : \omega \le F(b)\}$, an interval of Lebesgue measure $F(b)$. Hence $\Pr(X \le b) = F(b)$: the CDF of X is F. So once Lebesgue measure exists — which Section 1 already grants — the "hard" direction is just the familiar inverse-transform trick, no miracle required. This is what quietly underwrote all of your earlier statistics: every CDF or density you computed with really does have a random variable behind it.

Exercise 1.40 (The easy direction of the CDF theorem)**.** Let X be a random variable with CDF F_X. Prove properties C1–C3.

Hint: For C2, apply Corollary 1.17 to $\{X \le x\} \subseteq \{X \le y\}$. For C1 and C3, express each limit as the probability of a monotone *sequence of events — for right-continuity use $\{X \le x\} = \bigcap_n \{X \le x + 1/n\}$ — and invoke continuity of probability (Solved Exercise 1.19 and Exercise 1.20).*

[Answer: see §8. The converse is the quantile construction sketched above.]

Definition 1.41 (Density and the general expectation). When F_X is differentiable, its derivative

$$f_X(x) \;:=\; F_X'(x)$$

is the **probability density** of X. If a density exists, the expectation of Definition 1.36 takes its familiar continuous form

$$\mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x,$$

the exact analogue of $\sum_k x_k\, p_X(x_k)$ with the sum replaced by an integral. The uniform distribution on $[0,1]$ — density $f(x)=1$ on $[0,1]$, CDF $F(x)=x$ there — is the running example: it is nothing but the length (Lebesgue) measure of Section 1 seen through the lens of a random variable.

8. The Strong Law of Large Numbers

Probability is trusted for a practical reason: the average of many independent measurements is expected to reveal the underlying mean. That expectation is a theorem, not a definition, and the Borel–Cantelli lemma already proved is strong enough to establish it. We first isolate the two tools the proof needs, a tail bound and a factorisation, and then assemble them.

Proposition 1.42 (Markov's inequality)**.** Let Z be a random variable, $a>0$, and $p>0$. Then

$$\Pr(|Z|\ge a) \le \frac{\mathbb{E}[|Z|^p]}{a^p}.$$

Proof. On the event $\{|Z|\ge a\}$ one has $|Z|^p \ge a^p$, and everywhere $|Z|^p \ge 0$, so pointwise $|Z|^p \ge a^p\,\mathbf{1}\{|Z|\ge a\}$. Taking expectations, using monotonicity and Section 6,

$$\mathbb{E}[|Z|^p] \ge a^p\,\mathbb{E}\big[\mathbf{1}\{|Z|\ge a\}\big] = a^p\,\Pr(|Z|\ge a).$$

Dividing by $a^p>0$ gives the claim. □

Definition 1.43 (Independent random variables). Random variables $X_1,\dots,X_n$ are **independent** if for all Borel sets $B_1,\dots,B_n\subseteq\mathbb{R}$,

$$\Pr(X_1\in B_1,\dots,X_n\in B_n) = \prod_{i=1}^{n}\Pr(X_i\in B_i).$$

An infinite family is independent if every finite subfamily is. The variables are **independent and identically distributed** (i.i.d.) if in addition they share one common distribution.

Proposition 1.44 (Expectation factorises over independence)**.** If $X_1,\dots,X_n$ are independent and $g_1,\dots,g_n$ are functions with $\mathbb{E}|g_i(X_i)|<\infty$, then

$$\mathbb{E}\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n}\mathbb{E}[g_i(X_i)].$$

Proof. In the discrete case the joint pmf factorises by Definition 1.43, and

$$\mathbb{E}\left[\prod_i g_i(X_i)\right] = \sum_{x_1,\dots,x_n}\prod_i g_i(x_i)\,\Pr(X_i=x_i) = \prod_i\sum_{x_i} g_i(x_i)\,\Pr(X_i=x_i) = \prod_i \mathbb{E}[g_i(X_i)].$$

In general the joint law is the product of the marginals, and the factorisation is the expectation of a product against a product measure; the full justification rests on the construction of product measure and Fubini's theorem, which we take as background. □

With a tail bound and a factorisation in hand, the law follows from Borel–Cantelli, once we grant one extra moment.

Theorem 1.45 (Strong law of large numbers)**.** Let $X_1, X_2, \dots$ be i.i.d. with mean $\mu = \mathbb{E}[X_1]$.

(1) If $\mathbb{E}|X_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$ almost surely.

(2) If the X_i are non-negative and $\mathbb{E}[X_1] = \infty$, then $\frac{1}{n}\sum_{i=1}^{n} X_i \to \infty$ almost surely.

Proof. We prove (i) under the extra assumption of a finite fourth moment, $\mathbb{E}[X_1^4] < \infty$, which keeps the argument elementary. The theorem is true as stated, needing only a first moment in (i) and none in (ii), but those proofs use more than we develop here (see, for example, Durrett, *Probability: Theory and Examples*).

Centre the summands, $Y_i := X_i - \mu$. The Y_i are i.i.d. with $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i^4] < \infty$, since expanding $(X_i - \mu)^4$ involves only moments of X_i up to the fourth, all finite because a finite fourth moment forces the lower moments to be finite. Put $S_n := \sum_{i=1}^{n} Y_i$ and expand the fourth moment,

$$\mathbb{E}[S_n^4] = \sum_{i,j,k,l=1}^{n} \mathbb{E}[Y_i Y_j Y_k Y_l].$$

By independence and $\mathbb{E}[Y_i] = 0$ (Proposition 1.44), a term vanishes as soon as one index differs from the other three, since the lone factor contributes $\mathbb{E}[Y_i] = 0$. Only two index patterns survive: all four equal, contributing n terms each equal to $\mathbb{E}[Y_1^4]$; and two distinct values each appearing twice, contributing $3n(n-1)$ terms each equal to $(\mathbb{E}[Y_1^2])^2$, the factor 3 counting the ways to pair four positions. Hence, using $n(n-1) \le n^2$,

$$\mathbb{E}[S_n^4] = n\,\mathbb{E}[Y_1^4] + 3n(n-1)\big(\mathbb{E}[Y_1^2]\big)^2 \le Cn^2, \qquad C := \mathbb{E}[Y_1^4] + 3\big(\mathbb{E}[Y_1^2]\big)^2.$$

Fix $\varepsilon > 0$. By Markov's inequality (Proposition 1.42) with $p = 4$ applied to S_n,

$$\Pr\left(\frac{|S_n|}{n} \ge \varepsilon\right) = \Pr(|S_n| \ge n\varepsilon) \le \frac{\mathbb{E}[S_n^4]}{(n\varepsilon)^4} \le \frac{Cn^2}{n^4\varepsilon^4} = \frac{C}{\varepsilon^4} \cdot \frac{1}{n^2}.$$

Therefore $\sum_{n=1}^{\infty} \Pr(|S_n|/n \ge \varepsilon) \le (C/\varepsilon^4)\sum_n n^{-2} < \infty$, and the First Borel–Cantelli Lemma (Theorem 1.21), applied to $A_n = \{|S_n|/n \ge \varepsilon\}$, gives $\Pr(|S_n|/n \ge \varepsilon \text{ i.o.}) = 0$. So for each fixed $\varepsilon > 0$, almost surely $|S_n|/n < \varepsilon$ for all large n. Intersecting these probability-one events over $\varepsilon = 1/m$, $m \ge 1$, a countable intersection, gives $S_n/n \to 0$ almost surely. Since $\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + S_n/n$, the sample mean converges to μ almost surely. □

Solutions to Exercises

Solution to Exercise 1.12.

(a) If $\omega \in \bigcup_n D_n$, then $\omega \in D_N$ for some N, meaning $\omega \in A_k$ for all $k \ge N$. So ω is absent from A_k only for $k < N$, finitely many indices. Conversely, if $\omega \in A_k$ for all but finitely many k, choose N so that $\omega \in A_k$ for all $k \ge N$; then $\omega \in D_N$.

(b) If $\omega \in \liminf_n A_n$, then by (a) $\omega \in A_k$ for all but finitely many k. Since the index set is infinite, ω still belongs to infinitely many A_k, so $\omega \in \limsup_n A_n$ by Solved Exercise 1.10. □

Solution to Exercise 1.18.

We proceed by induction on n.

Base case ($n = 1$): $\Pr(A_1) \le \Pr(A_1)$. ✓

Inductive step: Assume $\Pr\big(\bigcup_{k=1}^{n} A_k\big) \le \sum_{k=1}^{n} \Pr(A_k)$. Write $\bigcup_{k=1}^{n+1} A_k = \big(\bigcup_{k=1}^{n} A_k\big) \cup A_{n+1}$. By Proposition 1.16:

$$\Pr\left(\bigcup_{k=1}^{n+1} A_k\right) = \Pr\left(\bigcup_{k=1}^{n} A_k\right) + \Pr(A_{n+1}) - \Pr\left(\bigcup_{k=1}^{n} A_k \cap A_{n+1}\right).$$

Since $\Pr(\,\cdot\,) \ge 0$, dropping the subtracted term gives:

$$\Pr\left(\bigcup_{k=1}^{n+1} A_k\right) \le \Pr\left(\bigcup_{k=1}^{n} A_k\right) + \Pr(A_{n+1}) \le \sum_{k=1}^{n} \Pr(A_k) + \Pr(A_{n+1}) = \sum_{k=1}^{n+1} \Pr(A_k),$$

where the second inequality uses the inductive hypothesis. □

Solution to Exercise 1.20.

Since $A_1 \supseteq A_2 \supseteq \cdots$, taking complements reverses the inclusions: $A_1^c \subseteq A_2^c \subseteq \cdots$ (if $A_{n+1} \subseteq A_n$ then $A_n^c \subseteq A_{n+1}^c$ by Axiom S2). Apply Solved Exercise 1.19 to the increasing sequence (A_n^c):

$$\Pr\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \lim_{n\to\infty} \Pr(A_n^c).$$

By De Morgan, $\bigcup_{n=1}^{\infty} A_n^c = \big(\bigcap_{n=1}^{\infty} A_n\big)^c$. By Corollary 1.17 (with $B \subseteq \Omega$): $\Pr(A_n^c) = 1 - \Pr(A_n)$ and $\Pr\big((\bigcap A_n)^c\big) = 1 - \Pr\big(\bigcap A_n\big)$. Substituting:

$$1 - \Pr\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \big(1 - \Pr(A_n)\big),$$

which rearranges to $\Pr\big(\bigcap_{n=1}^{\infty} A_n\big) = \lim_{n\to\infty} \Pr(A_n)$. □

Solution to Exercise 1.27.

Both identities are checked pointwise; every function involved takes only the values 0 and 1, so it is enough to see that the two sides agree at each ω.

(a) Fix ω. If $\omega \in A \cap B$, then $\mathbf{1}_A(\omega) = \mathbf{1}_B(\omega) = 1$, so the product equals 1; and $\mathbf{1}_{A\cap B}(\omega) = 1$. If $\omega \notin A \cap B$, then ω misses at least one of A, B, so at least one factor is 0 and the product is 0; and $\mathbf{1}_{A\cap B}(\omega) = 0$. The two sides agree at every ω, hence $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A\cap B}$.

(b) Assume $A \subseteq B$ and fix ω. If $\mathbf{1}_A(\omega) = 1$ then $\omega \in A \subseteq B$, so $\mathbf{1}_B(\omega) = 1$ and the inequality $\mathbf{1}_A(\omega) \le \mathbf{1}_B(\omega)$ holds. If $\mathbf{1}_A(\omega) = 0$, it holds because $\mathbf{1}_B(\omega) \ge 0$. □

Solution to Exercise 1.28.

Fix $\omega \in \Omega$ and check the two sides agree. If $\omega \notin A$, then since each $A_i \subseteq A$ we have $\omega \notin A_i$ for every i, so $\mathbf{1}_A(\omega) = 0$ and every $\mathbf{1}_{A_i}(\omega) = 0$: both sides are 0. If $\omega \in A$, then because the A_i partition A, the point ω belongs to exactly one part, say A_j; so $\mathbf{1}_{A_j}(\omega) = 1$ and $\mathbf{1}_{A_i}(\omega) = 0$ for $i \ne j$, whence $\sum_i \mathbf{1}_{A_i}(\omega) = 1 = \mathbf{1}_A(\omega)$. Agreeing at every ω, the functions are equal. The countable case is identical, the single surviving term again being the unique part containing ω. □

Solution to Exercise 1.29.

Fix ω. If $X(\omega) = c$, then $\mathbf{1}_{\{X=c\}}(\omega) = 1$ and $X(\omega)\mathbf{1}_{\{X=c\}}(\omega) = c\cdot 1 = c\,\mathbf{1}_{\{X=c\}}(\omega)$. If $X(\omega) \ne c$, then $\mathbf{1}_{\{X=c\}}(\omega) = 0$ and both sides are 0. So $X\mathbf{1}_{\{X=c\}} = c\,\mathbf{1}_{\{X=c\}}$ everywhere; taking $c = 0$ gives $X\mathbf{1}_{\{X=0\}} = 0$. □

Solution to Exercise 1.31.

Fix ω and set $k^* = N(\omega) \in S$. For each $k \in S$, $\mathbf{1}_{\{N=k\}}(\omega) = 1$ if $k = k^*$ and 0 otherwise, so the sum collapses to its single nonzero term:

$$\sum_{k\in S} k\,\mathbf{1}_{\{N=k\}}(\omega) = k^* = N(\omega).$$

As ω was arbitrary, $N = \sum_{k\in S} k\,\mathbf{1}_{\{N=k\}}$. If N is $\mathcal{G}$-measurable then each level set $\{N = k\}$ is a $\mathcal{G}$-event, so each $\mathbf{1}_{\{N=k\}}$ is $\mathcal{G}$-measurable, and the display writes N as a countable linear combination of $\mathcal{G}$-measurable indicators. □

Solution to Exercise 1.40.

Every threshold event $\{X \le t\}$ lies in $\mathcal{F}$ by Definition 1.22, so each probability below is defined; write $F(t) = \Pr(X \le t)$.

C2 (monotone). If $x \le y$ then $\{X \le x\} \subseteq \{X \le y\}$, since $X(\omega) \le x$ forces $X(\omega) \le y$. Monotonicity of probability (Corollary 1.17) gives $F(x) \le F(y)$. Thus F is non-decreasing and bounded in $[0, 1]$, so each one-sided limit used below exists and may be evaluated along a single monotone sequence approaching the point.

C1 (limits at $\pm\infty$). Put $E_n := \{X \le n\}$ for $n = 1, 2, \ldots$. Since $E_n \subseteq E_{n+1}$, the sequence is increasing, and

$$\bigcup_{n=1}^{\infty} E_n = \Omega,$$

because each ω has a finite real value $X(\omega)$ and so lies in E_n for every integer $n \ge X(\omega)$. Continuity from below (Solved Exercise 1.19) gives $\lim_{n\to\infty} F(n) = \Pr(\Omega) = 1$; as F is non-decreasing, the full limit agrees: $\lim_{x\to\infty} F(x) = 1$.

For the lower limit put $E'_n := \{X \le -n\}$. These decrease, $E'_{n+1} \subseteq E'_n$, and

$$\bigcap_{n=1}^{\infty} E'_n = \emptyset,$$

since no real number satisfies $X(\omega) \le -n$ for all n. Continuity from above (Exercise 1.20) gives $\lim_{n\to\infty} F(-n) = \Pr(\emptyset) = 0$, and monotonicity upgrades this to $\lim_{x\to-\infty} F(x) = 0$.

C3 (right-continuity). Fix x and put $G_n := \{X \le x + \frac{1}{n}\}$. These decrease as n grows, and

$$\bigcap_{n=1}^{\infty} G_n = \{X \le x\} :$$

if $X(\omega) \le x + \frac{1}{n}$ for every n, then letting $n \to \infty$ forces $X(\omega) \le x$; the reverse inclusion is immediate. Continuity from above (Exercise 1.20) gives

$$\lim_{n\to\infty} F\big(x + \tfrac{1}{n}\big) = \Pr(X \le x) = F(x).$$

Because F is non-decreasing, the limit along the sequence $x + 1/n$ equals the full right-hand limit, so $\lim_{y\downarrow x} F(y) = F(x)$. □

CHAPTER 2

Conditional Expectation

In an introductory course, the expectation of a random variable was a single number $\mathbb{E}[X]$, the long-run average of its values. That number, useful as it is, throws away everything: it is the same whether or not you have watched the experiment unfold. Return to the two-toss agent of Example 1.6, who is shown *only* the first toss. Ask this agent for the value of a quantity X that depends on both tosses. The agent cannot report X itself; the second toss is hidden. But the agent is not helpless: on the strength of the first toss alone, they can report a best estimate, and that estimate is itself a random variable, one that varies only with what the agent can see.

We now have two extremes and nothing between them. The number $\mathbb{E}[X]$ ignores the first toss entirely; the pointwise value $X(\omega)$ uses information the agent does not have. The object we want, expectation conditioned on a σ-algebra, lives in between: it uses exactly the information encoded in the agent's coarse σ-algebra, and no more. Before we can define it, we need to understand the shape of the functions the agent is even able to report, the ones measurable with respect to a coarse σ-algebra. That question turns out to have a completely explicit answer on finite spaces, and the key to it is the notion of an *atom*.

1. Atoms and the Structure of Finite σ-Algebras

DEFINITION 2.1 (Atom). Let $(\Omega, \mathcal{F})$ be a measurable space. A set $A \in \mathcal{F}$ is an **atom** of $\mathcal{F}$ if

(i) $A \neq \emptyset$, and

(ii) the only $B \in \mathcal{F}$ with $B \subseteq A$ and $B \neq A$ is $B = \emptyset$.

In words, an atom is a non-empty measurable set with no *non-empty, measurable, proper* subset: it cannot be split any further from inside $\mathcal{F}$, and so is the smallest resolvable piece of $\mathcal{F}$.

The non-emptiness clause (i) is not cosmetic. Without it the empty set would vacuously satisfy (ii), and calling $\emptyset$ an atom would ruin the picture of atoms as indivisible *pieces* of Ω. We rule it out by hand.

Example 2.2 (The atoms of the first-toss σ-algebra)**.** Take $\Omega = \{HH, HT, TH, TT\}$ and the first-toss σ-algebra of Example 1.6,

$$\mathcal{F}_1 = \big\{\, \emptyset,\ \{HH, HT\},\ \{TH, TT\},\ \Omega \,\big\}.$$

We claim its atoms are exactly $\{HH, HT\}$ and $\{TH, TT\}$.

$\{HH, HT\}$ is an atom. Write $A = \{HH, HT\}$. Its subsets are $\emptyset$, $\{HH\}$, $\{HT\}$, A. Of these, only $\emptyset$ and A belong to $\mathcal{F}_1$ (the two singletons do not). So the only *measurable* proper subset of A is $\emptyset$, and (ii) holds; since $A \neq \emptyset$, (i) holds too. The same check applies verbatim to $\{TH, TT\}$.

Note that Ω is *not* an atom: it has $\{HH, HT\}$ as a non-empty measurable proper subset. The atoms are precisely the finest events the first-toss agent can resolve, and they partition Ω.

REMARK (The atoms of the Borel σ-algebra). It is worth asking what the atoms look like when $\mathcal{F}$ is large. On $\mathbb{R}$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, every singleton $\{x\}$ is an atom: it is Borel, non-empty, and its only proper subset is $\emptyset$. Moreover there are *no other* atoms. Any non-empty Borel set A with more than one point contains a singleton $\{x\} \subsetneq A$, and $\{x\}$ is a non-empty measurable proper subset, so A fails (ii). Hence the atoms of $\mathcal{B}(\mathbb{R})$ are exactly the singletons. This is the opposite regime from a finite σ-algebra: the atoms are as small as possible, and there are uncountably many of them.

The next proposition is the reason atoms matter for us. It uses only measurability and the atom property; no probability measure is involved.

Proposition 2.3 (A random variable is constant on an atom)**.** Let X be an $\mathcal{F}$-measurable random variable on Ω, and let $A \in \mathcal{F}$ be an atom of $\mathcal{F}$. Then X is constant on A: there is a real number c with $X(\omega) = c$ for every $\omega \in A$.

PROOF. Suppose, for contradiction, that X is not constant on A. Then there are points $\omega_1, \omega_2 \in A$ with $X(\omega_1) \neq X(\omega_2)$. Relabelling if necessary, assume $X(\omega_1) = c_1 > c_2 = X(\omega_2)$; this loses no generality, since we are free to call whichever point carries the larger value ω_1.

Consider the threshold set

$$B \;:=\; \{\omega \in \Omega : X(\omega) \leq c_2\}.$$

Since X is $\mathcal{F}$-measurable, $B \in \mathcal{F}$ by Definition 1.22. Now $\omega_2 \in B$, because $X(\omega_2) = c_2 \leq c_2$; but $\omega_1 \notin B$, because $X(\omega_1) = c_1 > c_2$.

Look at $A \cap B$. By Proposition 1.3, $A \cap B \in \mathcal{F}$. It is non-empty, since $\omega_2 \in A \cap B$. It is a *proper* subset of A, since $\omega_1 \in A$ but $\omega_1 \notin A \cap B$. Thus $A \cap B$ is a non-empty, measurable, proper subset of A, contradicting the atom property (ii) of Definition 2.1. Therefore no such ω_1, ω_2 exist, and X is constant on A. □

A single atom already accounts for the simplest random variables we know: an indicator $\mathbf{1}_A$ (Example 1.26) is constant on A, taking the value 1 there. Proposition 2.3 says every random variable behaves this way, atom by atom. When there are finitely many atoms and they exhaust Ω, this pins down the random variable completely.

Proposition 2.4 (Representation on a finite σ-algebra)**.** Suppose $A_1, A_2, \ldots, A_n$ are atoms of $\mathcal{F}$ that partition Ω, and $\mathcal{F} = \sigma(\{A_1, \ldots, A_n\})$ is the σ-algebra they generate. Then every $\mathcal{F}$-measurable random variable X has the form

$$X \;=\; \sum_{i=1}^{n} c_i \, \mathbf{1}_{A_i} \qquad \text{for some } c \in \mathbb{R}^n,$$

where c_i is the constant value of X on the atom A_i.

PROOF. By Proposition 2.3, X is constant on each atom; write c_i for its value on A_i, so $X(\omega) = c_i$ whenever $\omega \in A_i$. Because the atoms partition Ω, every $\omega \in \Omega$ lies in exactly one atom, say A_j. There $\mathbf{1}_{A_i}(\omega) = 0$ for $i \neq j$ and $\mathbf{1}_{A_j}(\omega) = 1$, so

$$\sum_{i=1}^{n} c_i \, \mathbf{1}_{A_i}(\omega) \;=\; c_j \;=\; X(\omega).$$

The two functions agree at every ω, hence are equal. □

REMARK (The vector-space picture). The $\mathcal{F}$-measurable random variables form a real vector space: sums and scalar multiples of measurable functions are again measurable. When $\mathcal{F}$ is generated by the atoms $A_1, \dots, A_n$, Proposition 2.4 says the indicators $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n}$ *span* this space. They are also linearly independent: evaluating $\sum_i c_i \mathbf{1}_{A_i}$ at a point of A_j returns c_j, so the combination is the zero function only if every $c_j = 0$. Hence $\{\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n}\}$ is a basis, and the space of $\mathcal{F}$-measurable random variables is n-dimensional, one dimension per atom.
This is the fact that makes the finite theory clean. "Measurable with respect to $\mathcal{F}$" now has three interchangeable meanings: constant on each atom; a linear combination of the atom indicators; a vector of n coordinates. Keep this basis in mind, because the definition of conditional expectation is about to lean on it.

2. The Information Generated by a Random Variable

Section 2.1 pinned down $\mathcal{G}$-measurability whenever $\mathcal{G}$ is generated by finitely many atoms: a function is $\mathcal{G}$-measurable exactly when it is constant on each atom, and the whole function is then captured by finitely many numbers, one per atom. That description leans on there being only finitely many atoms to list. It gives no purchase on the σ-algebra that a single real-valued random variable carries with it, where the generating information is a whole continuum of events rather than a short list of atoms. For that case we need a different description of measurability, and the Doob–Dynkin theorem supplies the one we will use for the rest of the course.

DEFINITION 2.5 (σ-algebra generated by a random variable). For a function $X : \Omega \to \mathbb{R}$, the **$\sigma$-algebra generated by X** is

$$\sigma(X) \;:=\; \sigma\big(\{\, X^{-1}(B) : B \in \mathcal{B}(\mathbb{R}) \,\}\big),$$

the σ-algebra generated by the preimages of all Borel sets. It is the smallest σ-algebra on Ω with respect to which X is measurable.

Read $\sigma(X)$ as *the information revealed by observing* X: an event lies in $\sigma(X)$ precisely when knowing the value of X settles whether that event occurred.

Proposition 2.6 (X is measurable with respect to its own σ-algebra)**.** For any $X : \Omega \to \mathbb{R}$, the map X is $\sigma(X)$-measurable.

PROOF. By Definition 1.22 it suffices to show $X^{-1}(B) \in \sigma(X)$ for every Borel set B. But $\sigma(X)$ is, by Definition 2.5, the σ-algebra generated by the collection $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$, and a generated σ-algebra contains every set in its generating collection (Definition 1.7). Hence $X^{-1}(B) \in \sigma(X)$. □

The proof is deliberately a tautology: $\sigma(X)$ was built to be just large enough to make X measurable, and no larger. The real question is the converse direction of information flow: which *other* random variables are $\sigma(X)$-measurable? Unwinding the definition, a $\sigma(X)$-measurable Y is one whose every preimage $Y^{-1}(B)$ can be assembled, through countable unions, intersections, and complements, out of preimages of the form $X^{-1}(B')$. That is a statement about sets, and it is awkward both to check and to use. The Doob–Dynkin theorem trades it for a statement about functions.

DEFINITION 2.7 (Borel measurable function). A function $f : \mathbb{R} \to \mathbb{R}$ is **Borel measurable** if $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for every Borel set B.

Continuous functions, monotone functions, and indeed every function met in this course are Borel measurable, so in practice this condition is a formality.

Theorem 2.8 (Doob–Dynkin factorization)**.** Let $X, Y : \Omega \to \mathbb{R}$. Then Y is $\sigma(X)$-measurable if and only if there is a Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ such that

$$Y \;=\; f(X), \qquad \text{that is,} \qquad Y(\omega) = f\big(X(\omega)\big) \text{ for every } \omega \in \Omega.$$

We do not prove the substantive direction. Manufacturing the function f from nothing but the assumption that Y is $\sigma(X)$-measurable requires measure-theoretic machinery beyond this course. The converse direction is elementary, and you will prove it in Exercise 2.9.

Exercise 2.9 (The easy direction of Doob–Dynkin)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be Borel measurable and set $Y = f(X)$. Prove that Y is $\sigma(X)$-measurable, completing the converse half of Theorem 2.8.

[Answer: see §5.]

Exercise 2.10 (Computing $\sigma(X)$ on four coin tosses)**.** Four fair coins are tossed, so Ω is the set of the 16 length-four strings over $\{H, T\}$, with $\mathcal{F} = 2^{\Omega}$. Let $X(\omega)$ be the number of times two consecutive heads occur in ω; for instance $X(HHHH) = 3$, $X(HHTH) = 1$, and $X(HTHT) = 0$. Determine $\sigma(X)$: list its atoms explicitly, and state how many sets it contains.

[Answer: see §5.]

3. Conditional Expectation

Fix a probability space $(\Omega, \mathcal{F}, \Pr)$ and a random variable X that is **integrable**, meaning $\mathbb{E}[|X|] < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a **sub-σ-algebra**: a σ-algebra on Ω whose members all belong to $\mathcal{F}$.

The asymmetry between $\mathcal{F}$ and $\mathcal{G}$ is the whole point. If X is $\mathcal{F}$-measurable it need *not* be $\mathcal{G}$-measurable: the threshold sets $\{X \le b\}$ land in $\mathcal{F}$, but nothing forces them into the smaller collection $\mathcal{G}$. Conversely, any $\mathcal{G}$-measurable random variable is automatically $\mathcal{F}$-measurable (Exercise 2.12). So the $\mathcal{G}$-measurable random variables sit *inside* the $\mathcal{F}$-measurable ones as a subspace. For the first-toss agent, $\mathcal{F}$ is full information and $\mathcal{G} = \mathcal{F}_1$ is the first toss alone; a quantity like "total number of heads" is $\mathcal{F}$-measurable but not $\mathcal{G}$-measurable, so the agent cannot report it. What the agent *can* report is some $\mathcal{G}$-measurable function. Which one best represents X?

Definition 2.11 (Conditional expectation). Let X be integrable on $(\Omega, \mathcal{F}, \Pr)$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. A random variable Z is called a **conditional expectation of X given $\mathcal{G}$**, written $Z = \mathbb{E}[X \mid \mathcal{G}]$, if it satisfies both:

G1. Z is $\mathcal{G}$-measurable, and

G2. $\mathbb{E}[X\,\mathbf{1}_G] = \mathbb{E}[Z\,\mathbf{1}_G]$ for every event $G \in \mathcal{G}$.

Axiom G1 says the answer respects the information in $\mathcal{G}$: it is a function the agent can actually report. Axiom G2 says that, tested against any event the agent can see, Z carries the same average weight as X. Taken bare, the definition looks unmotivated. The following remark explains where it comes from; we will then set the motivation aside and compute only from G1 and G2.

REMARK (Conditional expectation as orthogonal projection). Equip the (square-integrable) random variables with the inner product

$$\langle U, V\rangle \;:=\; \mathbb{E}[UV].$$

This is bilinear, symmetric, and satisfies $\langle U, U\rangle = \mathbb{E}[U^2] \geq 0$, so it is a genuine inner product, with norm $\|U\| = \sqrt{\mathbb{E}[U^2]}$. The $\mathcal{F}$-measurable random variables form an inner-product space; the $\mathcal{G}$-measurable ones form a subspace, call it $L^2(\mathcal{G})$. A random variable X may point out of this subspace.

The nearest point of $L^2(\mathcal{G})$ to X, in the norm above, is the orthogonal projection $\widehat{X}$, characterised by two properties: $\widehat{X} \in L^2(\mathcal{G})$, and the residual $\varepsilon := X - \widehat{X}$ is orthogonal to *every* element of $L^2(\mathcal{G})$, that is, $\mathbb{E}[\varepsilon\, V] = 0$ for all $\mathcal{G}$-measurable V. Take $V = \mathbf{1}_G$ for an event $G \in \mathcal{G}$ (indicators of $\mathcal{G}$-events are $\mathcal{G}$-measurable):

$$\mathbb{E}\big[(X - \widehat{X})\,\mathbf{1}_G\big] = 0 \quad\Longleftrightarrow\quad \mathbb{E}[X\,\mathbf{1}_G] = \mathbb{E}[\widehat{X}\,\mathbf{1}_G].$$

That is exactly G2 with $Z = \widehat{X}$, and G1 holds because $\widehat{X} \in L^2(\mathcal{G})$. So the two axioms are not arbitrary: they say $\mathbb{E}[X \mid \mathcal{G}]$ is the orthogonal projection of X onto the information $\mathcal{G}$, the mean-square best $\mathcal{G}$-measurable approximation to X.

An econometrician will recognise the decomposition $X = \widehat{X} + \varepsilon$ as a population regression: $\widehat{X}$ is the part explained by the available information, and the residual ε is orthogonal to everything the analyst can measure, hence uncorrelated with every regressor. We will not use this geometry again; from here we work only from G1 and G2, which are easier to compute with. But when we say "Z is the best estimate of X given $\mathcal{G}$," this projection is the picture behind the words.

The definition is easiest to feel through its extreme cases and its basic structural properties. The next three exercises work them out.

Exercise 2.12 (Coarser information is inherited)**.** Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Prove that every $\mathcal{G}$-measurable random variable is $\mathcal{F}$-measurable. Then exhibit a probability space, a sub-σ-algebra $\mathcal{G}$, and an $\mathcal{F}$-measurable random variable that is *not* $\mathcal{G}$-measurable, so the converse fails.

[Answer: see §5.]

Exercise 2.13 (The trivial σ-algebra returns the mean)**.** The **trivial σ-algebra** on Ω is $\mathcal{T} = \{\emptyset, \Omega\}$. Prove that for every integrable X,

$$\mathbb{E}[X \mid \mathcal{T}] \;=\; \mathbb{E}[X],$$

the constant random variable equal to the number $\mathbb{E}[X]$.

[Answer: see §5.]

In an introductory statistics course you wrote $\mathbb{E}[Y \mid X]$, the conditional expectation of Y "given X," and treated it as a function of X. We can now say exactly what that phrase means: $\mathbb{E}[Y \mid X]$ is *defined* to be $\mathbb{E}[Y \mid \sigma(X)]$, the conditional expectation with respect to the information generated by X (Definition 2.5). The next exercise confirms that this abstract object really is a function of X, so the two notations agree.

Exercise 2.14 (Conditional expectation given X is a function of X)**.** Let Y be integrable and let $X : \Omega \to \mathbb{R}$. Show that there is a Borel measurable function $g : \mathbb{R} \to \mathbb{R}$ with

$$\mathbb{E}[Y \mid \sigma(X)] \;=\; g(X).$$

Hint: what does axiom G1 say about $\mathbb{E}[Y \mid \sigma(X)]$? Feed that into the Doob–Dynkin theorem (Theorem 2.8).

[Answer: see §5.]

4. A Formula on Finite σ-Algebras

The definition is abstract, so we now make it computable in the finite case, where atoms give us complete control. The result is a single formula we will reuse throughout the course.

Proposition 2.15 (Conditional expectation via atoms)**.** Let $C_1, C_2, \dots, C_r$ partition Ω into atoms with $\Pr(C_i) > 0$ for each i, and let $\mathcal{G} = \sigma(\{C_1, \dots, C_r\})$. Then for every integrable X,

$$\mathbb{E}[X \mid \mathcal{G}] \;=\; \sum_{i=1}^{r} \frac{\mathbb{E}[X\,\mathbf{1}_{C_i}]}{\Pr(C_i)}\,\mathbf{1}_{C_i}.$$

Proof. Let Z denote the right-hand side. We verify that Z satisfies the two axioms of Definition 2.11.

G1: Z is $\mathcal{G}$-measurable. Z is a linear combination of the indicators $\mathbf{1}_{C_i}$. Each $C_i \in \mathcal{G}$, so each $\mathbf{1}_{C_i}$ is $\mathcal{G}$-measurable (Example 1.26), and a linear combination of $\mathcal{G}$-measurable functions is $\mathcal{G}$-measurable.

G2: the averaging identity. Fix any event $G \in \mathcal{G}$. Since $\mathcal{G}$ is generated by the partition, G is a union of some subcollection of the atoms: there is an index set $J \subseteq \{1, \dots, r\}$ with

$$G \;=\; \bigsqcup_{j \in J} C_j.$$

(Every element of a σ-algebra generated by a partition is a union of atoms: complements and intersections of unions of atoms are again unions of atoms, so nothing else can arise.) Because the C_j are pairwise disjoint, the indicator of G splits as $\mathbf{1}_G = \sum_{j \in J} \mathbf{1}_{C_j}$.

Now compute $\mathbb{E}[Z\,\mathbf{1}_G]$. For each i, the product $\mathbf{1}_{C_i}\mathbf{1}_G = \mathbf{1}_{C_i \cap G}$ by Exercise 1.27, and since the atoms are disjoint,

$$C_i \cap G = \begin{cases} C_i, & i \in J, \\ \emptyset, & i \notin J. \end{cases}$$

Therefore

$$\mathbb{E}[Z\,\mathbf{1}_G] = \sum_{i=1}^{r} \frac{\mathbb{E}[X\,\mathbf{1}_{C_i}]}{\Pr(C_i)}\,\mathbb{E}[\mathbf{1}_{C_i}\mathbf{1}_G] = \sum_{i \in J} \frac{\mathbb{E}[X\,\mathbf{1}_{C_i}]}{\Pr(C_i)}\,\mathbb{E}[\mathbf{1}_{C_i}],$$

using linearity of expectation. By Section 6, $\mathbb{E}[\mathbf{1}_{C_i}] = \Pr(C_i)$, which cancels the denominator (this is where $\Pr(C_i) > 0$ is needed):

$$\mathbb{E}[Z\,\mathbf{1}_G] = \sum_{i \in J} \mathbb{E}[X\,\mathbf{1}_{C_i}] = \mathbb{E}\left[X \sum_{i \in J} \mathbf{1}_{C_i}\right] = \mathbb{E}[X\,\mathbf{1}_G],$$

again by linearity and $\mathbf{1}_G = \sum_{j \in J} \mathbf{1}_{C_j}$. So G2 holds for every $G \in \mathcal{G}$.

Since Z satisfies G1 and G2, it is the conditional expectation: $Z = \mathbb{E}[X \mid \mathcal{G}]$. □

REMARK (The coefficients are elementary conditional averages). The coefficient of $\mathbf{1}_{C_i}$ in Proposition 2.15 is

$$\frac{\mathbb{E}[X\,\mathbf{1}_{C_i}]}{\Pr(C_i)} \;=\; \mathbb{E}[X \mid C_i],$$

the elementary conditional expectation of X given the single event C_i that you met in an introductory course. So the σ-algebra conditional expectation is nothing exotic: on each atom C_i it simply equals the old event-conditional average $\mathbb{E}[X \mid C_i]$, and $\mathbb{E}[X \mid \mathcal{G}]$ packages one such number per atom into a single random variable. The abstract definition and the elementary formula agree; the projection of Section 3 is what unifies them.

5. Worked Examples

Solved Exercise 2.16 (Total heads given the first toss)**.** Toss a fair coin three times independently, so $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ with each outcome of probability $\frac{1}{8}$. Let N be the total number of heads. Let

$$C_H = \{\text{first toss is } H\} = \{HHH, HHT, HTH, HTT\}, \qquad C_T = C_H^c,$$

and let $\mathcal{F}_1 = \sigma(\{C_H, C_T\})$. Compute $\mathbb{E}[N \mid \mathcal{F}_1]$.

SOLUTION. The atoms of $\mathcal{F}_1$ are C_H, C_T, each of probability $\frac{1}{2}$. By Proposition 2.15,

$$\frac{\mathbb{E}[N\,\mathbf{1}_{C_H}]}{\Pr(C_H)} = \frac{\frac{1}{8}(3+2+2+1)}{\frac{1}{2}} = \frac{1}{\frac{1}{2}} = 2, \qquad \frac{\mathbb{E}[N\,\mathbf{1}_{C_T}]}{\Pr(C_T)} = \frac{\frac{1}{8}(2+1+1+0)}{\frac{1}{2}} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1,$$

so $\mathbb{E}[N \mid \mathcal{F}_1] = 2\,\mathbf{1}_{C_H} + \mathbf{1}_{C_T}$. □

Exercise 2.17 (Total heads given the first two tosses)**.** Keep the three-toss experiment of Solved Exercise 2.16. Let $\mathcal{F}_2$ be the σ-algebra generated by the partition into the four atoms

$$C_{HH} = \{HHH, HHT\}, \quad C_{HT} = \{HTH, HTT\},$$
$$C_{TH} = \{THH, THT\}, \quad C_{TT} = \{TTH, TTT\}$$

determined by the first two tosses. Compute $\mathbb{E}[N \mid \mathcal{F}_2]$, and check that your answer equals (number of heads in the first two tosses) $+ \frac{1}{2}$.

[Answer: see §5.]

Exercise 2.18 (Conditional expectation of an indicator)**.** Let $A, B \in \mathcal{F}$ be events with $0 < \Pr(B) < 1$. Compute $\mathbb{E}[\mathbf{1}_A \mid \sigma(B)]$, and express your answer through the elementary conditional probabilities $\Pr(A \mid B)$ and $\Pr(A \mid B^c)$.

[Answer: see §5.]

The examples so far have lived on finite sample spaces. The formula is not confined to them: it applies whenever $\mathcal{G}$ is generated by finitely many atoms of positive probability, even on a continuous space. We set up the standard continuous model and work an example on it.

DEFINITION 2.19 (The uniform probability space on $[0,1]$). The **uniform probability space** is $(\Omega, \mathcal{F}, \Pr)$ with $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$ the Borel σ-algebra, and $\Pr$ the unique probability measure with $\Pr([a,b]) = b - a$ for all $0 \le a \le b \le 1$ (Section 1). The expectation of an

integrable random variable $Z : [0,1] \to \mathbb{R}$ is the integral

$$\mathbb{E}[Z] \;=\; \int_0^1 Z(u)\,\mathrm{d}u.$$

Solved Exercise 2.20 (Conditional expectation on the unit interval)**.** On the uniform probability space of Definition 2.19, let $X(u) = 2u^2$ and

$$Y(u) = \begin{cases} 0, & u \in [0, \tfrac{1}{3}], \\ 1, & u \in (\tfrac{1}{3}, \tfrac{2}{3}], \\ 2, & u \in (\tfrac{2}{3}, 1]. \end{cases}$$

Compute $\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \sigma(Y)]$.

SOLUTION. $\sigma(Y)$ has atoms $C_1 = [0, \tfrac{1}{3}]$, $C_2 = (\tfrac{1}{3}, \tfrac{2}{3}]$, $C_3 = (\tfrac{2}{3}, 1]$, each of probability $\tfrac{1}{3}$. By Proposition 2.15, with $\mathbb{E}[X\,\mathbf{1}_{C_i}] = \int_{C_i} 2u^2\,\mathrm{d}u$,

$$\frac{\int_0^{1/3} 2u^2\,\mathrm{d}u}{1/3} = \frac{2/81}{1/3} = \tfrac{2}{27}, \qquad \frac{\int_{1/3}^{2/3} 2u^2\,\mathrm{d}u}{1/3} = \frac{14/81}{1/3} = \tfrac{14}{27}, \qquad \frac{\int_{2/3}^{1} 2u^2\,\mathrm{d}u}{1/3} = \frac{38/81}{1/3} = \tfrac{38}{27},$$

so $\mathbb{E}[X \mid Y] = \tfrac{2}{27}\,\mathbf{1}_{C_1} + \tfrac{14}{27}\,\mathbf{1}_{C_2} + \tfrac{38}{27}\,\mathbf{1}_{C_3}$. □

Exercise 2.21 (When the formula does not apply: work from the definition)**.** On the uniform probability space of Definition 2.19, let $X(u) = 2u^2$ and

$$Y(u) = \begin{cases} 2, & u \in [0, \tfrac{1}{2}), \\ u, & u \in [\tfrac{1}{2}, 1]. \end{cases}$$

Here Y takes infinitely many values, so the finite-partition formula of Proposition 2.15 does not apply. Working directly from axioms G1 and G2 of Definition 2.11, compute $\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \sigma(Y)]$.

[Answer: see §5.]

Solutions to Exercises

Solution to Exercise 2.9.

Let B be any Borel set. Chase a point $\omega \in \Omega$ through the definitions:

$$\omega \in Y^{-1}(B) \iff f\big(X(\omega)\big) \in B \iff X(\omega) \in f^{-1}(B) \iff \omega \in X^{-1}\big(f^{-1}(B)\big).$$

Hence $Y^{-1}(B) = X^{-1}\big(f^{-1}(B)\big)$. Since f is Borel measurable, $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ (Definition 2.7), so $X^{-1}\big(f^{-1}(B)\big)$ is one of the generators $\{X^{-1}(B') : B' \in \mathcal{B}(\mathbb{R})\}$ of $\sigma(X)$ and therefore belongs to $\sigma(X)$ (Definition 1.7). As B was an arbitrary Borel set, every preimage $Y^{-1}(B)$ lies in $\sigma(X)$, so Y is $\sigma(X)$-measurable by Definition 1.22. □

Solution to Exercise 2.10.

Evaluate X on all 16 strings and collect them by value. Counting adjacent HH pairs in positions $(1,2), (2,3), (3,4)$ gives the level sets

$$A_3 = \{X = 3\} = \{HHHH\},$$
$$A_2 = \{X = 2\} = \{HHHT,\ THHH\},$$
$$A_1 = \{X = 1\} = \{HHTH,\ HHTT,\ HTHH,\ THHT,\ TTHH\},$$
$$A_0 = \{X = 0\} = \{HTHT,\ HTTH,\ HTTT,\ THTH,\ THTT,\ TTHT,\ TTTH,\ TTTT\},$$

of sizes $1, 2, 5, 8$, which sum to 16. Each level set is $A_k = X^{-1}(\{k\})$, so all four belong to $\sigma(X)$. Conversely, for any Borel B the preimage $X^{-1}(B) = \bigcup_{k \in B} A_k$ is a union of these sets, so $\sigma(X)$ contains nothing finer: the atoms of $\sigma(X)$ are exactly A_0, A_1, A_2, A_3. Hence $\sigma(X)$ consists of all unions of these four atoms, and $|\sigma(X)| = 2^4 = 16$. □

Solution to Exercise 2.12.

Forward direction. Let Y be $\mathcal{G}$-measurable and let B be any Borel set. Then $Y^{-1}(B) \in \mathcal{G}$ by Definition 1.22, and since $\mathcal{G} \subseteq \mathcal{F}$ we get $Y^{-1}(B) \in \mathcal{F}$. As B was arbitrary, Y is $\mathcal{F}$-measurable.

The converse fails. Take $\Omega = \{HH, HT, TH, TT\}$ with $\mathcal{F} = 2^\Omega$ the power set, and let $\mathcal{G} = \mathcal{F}_1 = \{\emptyset, \{HH, HT\}, \{TH, TT\}, \Omega\}$ be the first-toss σ-algebra of Example 1.6. Let $Y = \mathbf{1}_{\{HH,TH\}}$, the indicator of "the second toss is heads." Since $\mathcal{F}$ is the power set, Y is $\mathcal{F}$-measurable. But $Y^{-1}(\{1\}) = \{HH, TH\} \notin \mathcal{G}$, so Y is not $\mathcal{G}$-measurable. Thus an $\mathcal{F}$-measurable random variable need not be $\mathcal{G}$-measurable. □

Solution to Exercise 2.13.

Let Z be the constant random variable $Z \equiv \mathbb{E}[X]$. We check the two axioms of Definition 2.11 with $\mathcal{G} = \mathcal{T} = \{\emptyset, \Omega\}$.

G1. For any Borel set B, $Z^{-1}(B) = \Omega$ if $\mathbb{E}[X] \in B$ and $Z^{-1}(B) = \emptyset$ otherwise; both lie in $\mathcal{T}$. So Z is $\mathcal{T}$-measurable.

G2. The only events in $\mathcal{T}$ are $\emptyset$ and Ω. For $G = \Omega$: $\mathbb{E}[X\,\mathbf{1}_\Omega] = \mathbb{E}[X]$, and $\mathbb{E}[Z\,\mathbf{1}_\Omega] = \mathbb{E}[Z] = \mathbb{E}[X]$, since the expectation of the constant $\mathbb{E}[X]$ is $\mathbb{E}[X]$. For $G = \emptyset$: both $\mathbf{1}_\emptyset = 0$, so $\mathbb{E}[X\,\mathbf{1}_\emptyset] = 0 = \mathbb{E}[Z\,\mathbf{1}_\emptyset]$. Both cases hold.

Since Z satisfies G1 and G2, $\mathbb{E}[X \mid \mathcal{T}] = \mathbb{E}[X]$. □

Solution to Exercise 2.14.

Let $Z = \mathbb{E}[Y \mid \sigma(X)]$. By axiom G1 of Definition 2.11, Z is $\sigma(X)$-measurable. By the Doob–Dynkin theorem (Theorem 2.8), every $\sigma(X)$-measurable random variable equals $g(X)$ for some Borel measurable $g : \mathbb{R} \to \mathbb{R}$. Hence $\mathbb{E}[Y \mid \sigma(X)] = g(X)$, which is exactly what licenses the classical notation $\mathbb{E}[Y \mid X] = g(X)$. □

Solution to Exercise 2.17.

The atoms $C_{HH}, C_{HT}, C_{TH}, C_{TT}$ each have probability $\frac{1}{4}$. By Proposition 2.15,

$$\frac{\mathbb{E}[N\,\mathbf{1}_{C_{HH}}]}{\Pr(C_{HH})} = \frac{\frac{1}{8}(3+2)}{\frac{1}{4}} = \frac{5}{2}, \qquad \frac{\mathbb{E}[N\,\mathbf{1}_{C_{TH}}]}{\Pr(C_{TH})} = \frac{\frac{1}{8}(2+1)}{\frac{1}{4}} = \frac{3}{2},$$

$$\frac{\mathbb{E}[N\,\mathbf{1}_{C_{HT}}]}{\Pr(C_{HT})} = \frac{\frac{1}{8}(2+1)}{\frac{1}{4}} = \frac{3}{2}, \qquad \frac{\mathbb{E}[N\,\mathbf{1}_{C_{TT}}]}{\Pr(C_{TT})} = \frac{\frac{1}{8}(1+0)}{\frac{1}{4}} = \frac{1}{2},$$

so

$$\mathbb{E}[N \mid \mathcal{F}_2] = \tfrac{5}{2}\,\mathbf{1}_{C_{HH}} + \tfrac{3}{2}\,\mathbf{1}_{C_{HT}} + \tfrac{3}{2}\,\mathbf{1}_{C_{TH}} + \tfrac{1}{2}\,\mathbf{1}_{C_{TT}},$$

which equals (number of heads in the first two tosses) $+ \frac{1}{2}$ on each atom. □

Solution to Exercise 2.18.

Since $0 < \Pr(B) < 1$, the σ-algebra $\sigma(B) = \{\emptyset, B, B^c, \Omega\}$ has the two atoms B and B^c, both of positive probability. By Proposition 2.15,

$$\mathbb{E}[\mathbf{1}_A \mid \sigma(B)] = \frac{\mathbb{E}[\mathbf{1}_A\,\mathbf{1}_B]}{\Pr(B)}\,\mathbf{1}_B + \frac{\mathbb{E}[\mathbf{1}_A\,\mathbf{1}_{B^c}]}{\Pr(B^c)}\,\mathbf{1}_{B^c}.$$

By Exercise 1.27 and Section 6, $\mathbb{E}[\mathbf{1}_A\,\mathbf{1}_B] = \mathbb{E}[\mathbf{1}_{A\cap B}] = \Pr(A\cap B)$, and likewise $\mathbb{E}[\mathbf{1}_A\,\mathbf{1}_{B^c}] = \Pr(A\cap B^c)$. The coefficients are therefore the elementary conditional probabilities $\Pr(A\cap B)/\Pr(B) = \Pr(A\mid B)$ and $\Pr(A\cap B^c)/\Pr(B^c) = \Pr(A\mid B^c)$, giving

$$\mathbb{E}[\mathbf{1}_A \mid \sigma(B)] = \Pr(A\mid B)\,\mathbf{1}_B + \Pr(A\mid B^c)\,\mathbf{1}_{B^c}.$$

□

Solution to Exercise 2.21.

Form of the answer. By G1, $Z = \mathbb{E}[X \mid \sigma(Y)]$ is $\sigma(Y)$-measurable, so by Doob–Dynkin (Theorem 2.8) $Z = g(Y)$ for some Borel $g:\mathbb{R}\to\mathbb{R}$. Using $Y\equiv 2$ on $[0,\frac12)$ and $Y(u)=u$ on $[\frac12,1]$,

$$Z(u) = g(Y(u)) = \begin{cases} g(2), & u\in[0,\frac12),\\ g(u), & u\in[\frac12,1]. \end{cases}$$

It remains to find $g(2)$ and $g(y)$ for $y\in[\frac12,1]$; we read them off from G2, testing against the events $G = Y^{-1}(B)$ for well-chosen Borel B.

The value $g(2)$. Take $B=\{2\}$. Since $Y=2$ exactly on $[0,\frac12)$, $G = Y^{-1}(\{2\}) = [0,\frac12)$, and G2 gives

$$\underbrace{\int_0^{1/2} 2u^2\,\mathrm{d}u}_{\mathbb{E}[X\,\mathbf{1}_G]} = \tfrac{1}{12} = \underbrace{\int_0^{1/2} g(2)\,\mathrm{d}u}_{\mathbb{E}[Z\,\mathbf{1}_G]} = \tfrac12\,g(2),$$

so $g(2)=\frac16$.

The function g on $[\frac12,1]$. For $t\in(\frac12,1)$ take $B=[\frac12,t)$. Since $Y<\frac12$ never holds and $Y=2\notin B$, we have $G = Y^{-1}([\frac12,t)) = [\frac12,t)$, and there $Z=g(u)$, so G2 reads

$$\int_{1/2}^{t} 2u^2\,\mathrm{d}u = \mathbb{E}[X\,\mathbf{1}_G] = \mathbb{E}[Z\,\mathbf{1}_G] = \int_{1/2}^{t} g(u)\,\mathrm{d}u.$$

This holds for every $t\in(\frac12,1)$; differentiating both sides in t (fundamental theorem of calculus) gives $g(t)=2t^2$ on $[\frac12,1]$.

Conclusion.

$$\mathbb{E}[X\mid Y](u) = \begin{cases} \frac16, & u\in[0,\frac12),\\ 2u^2, & u\in[\frac12,1]. \end{cases}$$

□

CHAPTER 3

Stochastic Processes

A stock price, or a country's quarterly output, is recorded as one number, then another, then another: a sequence of values unfolding in time. The natural first description, that each period's value is a random variable, is correct, but it leaves two facts implicit, and both organise this chapter.

The first is that these random variables are defined on a single common probability space. One realisation of the world does not deliver one number today and, from a separate experiment, another tomorrow; it fixes the entire sequence at once. A single outcome ω therefore determines a whole function $t \mapsto X_t(\omega)$, the trajectory over time, which is called a sample path. The observed data is one such path.

The second is that every question one asks of such a process, whether the mean is constant over time, whether successive values are correlated, whether the past helps predict the future, concerns the joint behaviour of *finitely many* of the variables at once, never a single X_t in isolation. This suggests specifying the process entirely through the joint distributions of its finite subfamilies, without constructing the underlying $(\Omega, \mathcal{F}, \Pr)$ at all. That this is legitimate is the central technical result of the chapter, due to Kolmogorov, and it underlies the standard practice of time-series analysis.

1. Processes, Sample Paths, and Finite-Dimensional Distributions

DEFINITION 3.1 (Stochastic process). Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and I an index set. A **stochastic process** is a collection

$$\{X_t\}_{t \in I}$$

of random variables $X_t \colon \Omega \to \mathbb{R}$, all defined on the same space $(\Omega, \mathcal{F}, \Pr)$. The set I is the **index set**; we read it as time. For each fixed outcome $\omega \in \Omega$, the map

$$t \longmapsto X_t(\omega), \qquad t \in I,$$

is the **sample path** (or **realisation**) of the process at ω.

The definition admits two readings. With t fixed and ω varying, X_t is a random variable, the state of the process at time t. With ω fixed and t varying, $t \mapsto X_t(\omega)$ is a sample path, one complete history. The observed data is always a single sample path; the randomness lies in which ω is realised.

Since every such question concerns finitely many of the variables at once, their joint laws are named as follows.

DEFINITION 3.2 (Finite-dimensional distributions). Let $\{X_t\}_{t \in I}$ be a stochastic process. For each finite tuple of distinct indices $t_1, \dots, t_k \in I$, the **finite-dimensional distribution** (FDD) at $(t_1, \dots, t_k)$ is the joint law of $(X_{t_1}, \dots, X_{t_k})$, recorded by its joint CDF

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) \;:=\; \Pr(X_{t_1} \le x_1, \dots, X_{t_k} \le x_k), \qquad (x_1, \dots, x_k) \in \mathbb{R}^k.$$

The collection of all such CDFs, over all finite tuples, is the **family of finite-dimensional distributions** of the process.

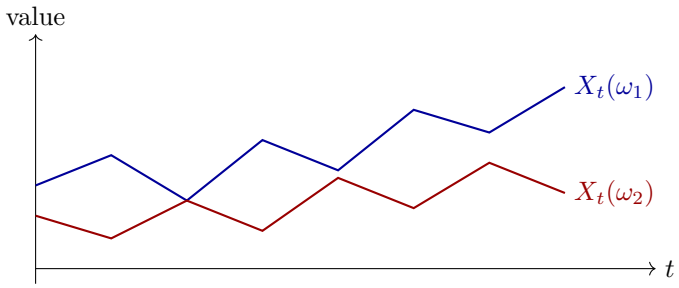


FIGURE 1. Two sample paths of one stochastic process. With ω fixed, the map $t \mapsto X_t(\omega)$ traces a single complete history; the observed data is one such path, and the randomness lies in which ω is realised.

The FDDs are not an arbitrary list of functions. Since they are read off a single process, they satisfy two natural conditions. First, the indices are labels: reordering them together with their arguments cannot change the probability. Second, ignoring a coordinate coincides with never having sampled it, so sending one argument to $+\infty$ recovers the FDD of the remaining indices. A family of candidate FDDs obeying these two conditions is called **consistent**; the conditions are stated precisely in the appendix. Consistency is exactly what is needed to reconstruct a process from its FDDs, so that specifying the FDDs *is* specifying the process.

2. A Gallery of Examples

The following examples show how much the finite-dimensional distributions already determine. The first two reclassify familiar objects as stochastic processes; the third is the central example of the chapter.

Example 3.3 (A single random variable)**.** Take $I = \{1\}$. A process indexed by a one-point set is nothing but a single random variable X_1. Everything studied in an introductory statistics course is therefore the smallest possible stochastic process. The one FDD is just the CDF of X_1.

Example 3.4 (A deterministic function is a process)**.** Let $I = \mathbb{R}$ and let $f\colon \mathbb{R} \to \mathbb{R}$ be an ordinary function, say $f(t) = t^2$. Set $X_t(\omega) = f(t)$ for every ω. This is a legitimate stochastic process in which nothing is random: at each time t the variable X_t takes the single value $f(t)$ with probability one, so its law is the point mass $\delta_{f(t)}$. The FDD at $(t_1, \dots, t_k)$ is degenerate,

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) = \prod_{i=1}^{k} \mathbf{1}\{\, f(t_i) \le x_i \,\},$$

a product of indicators. A deterministic function is thus a stochastic process in which the randomness is absent at every time. The point is not the triviality itself but that the finite-dimensional description remains valid in this degenerate case, returning point masses.

REMARK. The point mass δ_c has no density in the ordinary sense: all of its mass sits at the single point c. Physicists write it as the "Dirac delta" and manipulate it as if it were a function that is zero everywhere except at c, yet integrates to one; making that rigorous requires Schwartz's theory of distributions. We will not need any of it. For us δ_c is simply the law "equal to c with probability one," and the indicator formula above is all we ever compute.

The examples so far reclassify familiar objects. The next is genuinely a process of the kind the theory is built for.

Example 3.5 (Tossing a coin forever)**.** Toss a fair coin once per period, forever. An outcome is an entire infinite sequence of heads and tails, so

$$\Omega = \{H, T\}^{\mathbb{N}} = \{\, \omega = (\omega_1, \omega_2, \dots) : \omega_k \in \{H, T\} \,\},$$

the set of all functions from $\mathbb{N}$ to $\{H, T\}$. Define the process

$$X_k(\omega) = \begin{cases} 1, & \omega_k = H, \\ 0, & \omega_k = T, \end{cases} \qquad k \in \mathbb{N},$$

so X_k records the k-th toss. This process serves as the principal example in what follows.
The instructive difficulty is the size of the sample space. Writing each sequence of 0s and 1s after a binary point identifies Ω with the interval $[0, 1]$, which is uncountable by Cantor's diagonal argument. One cannot then take $\mathcal{F} = 2^{\Omega}$ and assign positive probability to individual paths: each path has probability zero, and there are uncountably many of them. A direct construction of Pr on Ω is therefore not available.
The FDDs, by contrast, are immediate. For distinct indices $k_1, \dots, k_n$ and values $a_1, \dots, a_n \in \{0, 1\}$, independence and fairness give

$$\Pr(X_{k_1} = a_1, \dots, X_{k_n} = a_n) = \frac{1}{2^n}.$$

This single formula specifies every FDD of the process. Nothing has been said about constructing Pr on the uncountable Ω, and the next section shows that this is unnecessary.

The coin is the simplest instance of a sequence of independent, identically distributed draws. The fair Bernoulli law may be replaced by any distribution.

Definition 3.6 (IID process). Let F be a CDF. A process $\{X_t\}_{t \in I}$ is **independent and identically distributed** (IID) with marginal F if, for every finite tuple $t_1, \dots, t_k$,

$$F_{t_1, \dots, t_k}(x_1, \dots, x_k) = \prod_{i=1}^{k} F(x_i).$$

The product records independence; using the *same* F in every factor records identical distribution.

The coin of Example 3.5 is the IID process with F the fair Bernoulli CDF. Definition 3.6 constructs a full family of FDDs from a single CDF F. That this family may be treated as an actual process on some $(\Omega, \mathcal{F}, \Pr)$ is established in the next section.

3. Specifying a Process by Its Finite-Dimensional Distributions

In Example 3.5 every FDD was written down without constructing Pr on the uncountable path space, and in Definition 3.6 a family of FDDs was built from a single CDF with no probability space specified. In each case the computations proceed as if a process existed. That it does is guaranteed by the following theorem of Kolmogorov, which underlies time-series analysis.

Theorem 3.7 (Kolmogorov consistency, informal statement)**.** Let there be given a family of candidate finite-dimensional distributions, one joint CDF $F_{t_1, \dots, t_k}$ for each finite tuple of indices, and suppose this family is *consistent*: reordering indices reorders arguments the same way, and integrating out a coordinate (sending its argument to $+\infty$) returns the FDD of the remaining

indices. Then there exists a probability space $(\Omega, \mathcal{F}, \Pr)$ and a stochastic process $\{X_t\}_{t\in I}$ on it whose finite-dimensional distributions are exactly the prescribed $F_{t_1,\dots,t_k}$.

The consequence is that defining a stochastic process requires no explicit construction of Ω, $\mathcal{F}$, or Pr. It suffices to specify a consistent family of FDDs; the theorem supplies the rest, and the construction is unique in the relevant sense, that any two assign the same probabilities to the same finite-sample events. Accordingly, a process is described from here on by its FDDs alone, with the σ-algebra left implicit.

The rigorous statement, including the precise consistency conditions, the cylinder sets on which Pr is first defined, and the way Carathéodory extension (Section 1) carries it to the full σ-algebra, is in the appendix to this chapter. It is more than a first undergraduate course needs, but the body of the chapter will feel free to appeal to it.

Example 3.8 (The deterministic process, re-read)**.** The deterministic process $f(t) = t^2$ of Example 3.4 illustrates the theorem. Consider $\Pr(X_2 \le 3,\, X_5 \le 26,\, X_7 \le 40)$. Since $X_t = t^2$ surely, the event is $\{4 \le 3\} \cap \{25 \le 26\} \cap \{49 \le 40\}$; the first and third conditions fail, so the probability is 0. Every finite-sample probability is determined in this way, the FDDs are consistent, and the theorem yields the process without any description of its sample space.

4. Stationarity

In full generality the law of a stochastic process at one time bears no relation to its law at another. Under such freedom nothing is predictable, and, for an economist, nothing is *testable*: if the mechanism generating the data in 2010 differs from the one generating it in 2025, a hypothesis confirmed on the first decade need not hold on the second, and the empirical finding cannot be reproduced by sampling at another time. Reproducibility across time is a precondition for treating a claim about the data as scientific.

Stationarity is the property that supplies it. It is not a class of process, like Bernoulli or Gaussian, but a property that a given process either has or lacks; since a process is determined by its FDDs, the property is stated on them.

We assume from now on that the index set I is closed under addition, so that a shift $t \mapsto t + j$ keeps us inside I: the integers $\mathbb{Z}$, the naturals $\mathbb{N}$, or the reals $\mathbb{R}$ all qualify. Without this the definition below cannot even be written.

Definition 3.9 (Strict stationarity). A process $\{X_t\}_{t\in I}$ is **strictly stationary** if its finite-dimensional distributions are invariant under time shifts: for every finite tuple $t_1, \dots, t_k \in I$, every shift j with $t_1 + j, \dots, t_k + j \in I$, and all $x_1, \dots, x_k$,

$$F_{t_1+j,\,\dots,\,t_k+j}(x_1, \dots, x_k) = F_{t_1,\dots,t_k}(x_1, \dots, x_k).$$

Equivalently, $(X_{t_1}, \dots, X_{t_k})$ and $(X_{t_1+j}, \dots, X_{t_k+j})$ have the same joint distribution: sliding the whole window along the time axis changes nothing.

Two consequences follow immediately; they are the properties used throughout.

Proposition 3.10 (First two moments under strict stationarity)**.** Let $\{X_t\}$ be strictly stationary with $\mathbb{E}[X_t^2] < \infty$. Then:

(i) the mean is constant, $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ for all t; and

(ii) the covariance depends only on the lag, $\operatorname{Cov}(X_t, X_s) = \operatorname{Cov}(X_{t-s}, X_0)$, a function of $t - s$ alone.

PROOF. (i) Take $k = 1$ and $t_1 = t$. Stationarity with $j = -t$ gives that X_t and X_0 share a CDF, hence $\mathbb{E}[X_t] = \int x \, \mathrm{d}F_t(x) = \int x \, \mathrm{d}F_0(x) = \mathbb{E}[X_0]$.

(ii) By (i) the mean is a constant μ. Take $k = 2$ and shift by $j = -s$: (X_t, X_s) and (X_{t-s}, X_0) have the same joint law, so

$$\mathrm{Cov}(X_t, X_s) = \mathbb{E}[(X_t - \mu)(X_s - \mu)] = \mathbb{E}[(X_{t-s} - \mu)(X_0 - \mu)] = \mathrm{Cov}(X_{t-s}, X_0),$$

which depends on t and s only through $t - s$. □

Proposition 3.10 has a direct economic reading. When a structural relation is asserted to hold, the assertion is not indexed to a particular decade; a relation valid in one era but not the next would not constitute a theory. Stationarity is the formal content of this permanence: the same sampling procedure applied today and twenty years hence draws from one distribution, so that a valid theory can be tested repeatedly with agreeing results. For this reason applied econometrics tests for stationarity *first*, transforming or differencing the data until the test passes before drawing any structural conclusion.

Exercise 3.11 (IID processes are stationary)**.** Show that every IID process (Definition 3.6) is strictly stationary.

Exercise 3.12 (Constant processes are stationary)**.** Fix $c \in \mathbb{R}$ and set $X_t = c$ for all $t \in I$. Show that $\{X_t\}$ is strictly stationary. Show that the same holds if a single random variable Z is repeated, $X_t \equiv Z$.

In practice we can never verify the infinitely many FDD identities that strict stationarity demands. Data let us estimate first and second moments and little else. So we isolate exactly the part of stationarity that Proposition 3.10 showed is testable, and promote it to a definition in its own right.

DEFINITION 3.13 (Covariance stationarity). A process (X_t) with $\mathbb{E}[X_t^2] < \infty$ is **covariance-stationary** if $\mathbb{E}[X_t]$ does not depend on t and the autocovariance $\mathrm{Cov}(X_t, X_{t+k}) = \gamma_k$ is a function of the lag k alone, not of t. (The names *weakly*, *wide-sense*, and *second-order stationary* are used elsewhere for the same notion; we say covariance-stationary throughout.)

By Proposition 3.10, strict stationarity plus a finite second moment implies covariance stationarity. The converse fails: covariance stationarity constrains only the first two moments and says nothing about the rest of the joint law. The two notions coincide for Gaussian processes, whose FDDs are determined by their means and covariances, but not in general. The canonical covariance-stationary building block is the following.

DEFINITION 3.14 (White noise). A **white-noise** sequence (u_t) has $\mathbb{E}[u_t] = 0$, $\mathrm{Var}(u_t) = \sigma^2 < \infty$ for all t, and $\mathrm{Cov}(u_s, u_t) = 0$ for $s \neq t$.

White noise is covariance-stationary by construction, yet need not be IID or even strictly stationary: uncorrelated is weaker than independent, and the higher moments are left free.

5. Ergodicity

Stationarity makes a process reproducible in principle but does not by itself permit computation, and here a genuine obstruction arises.

Consider the mean daily return of a stock. The mean is an average across outcomes ω, over the ensemble of all the ways the world might have gone: across a notional continuum of parallel realisations, each tracing a different sample path, $\mathbb{E}[X_t]$ averages the value at time t over realisations. Only one realisation is observed, however, and the sole average available from it is taken along the single sample path, over time. The time average is then used in place of the ensemble average, which is not observable.

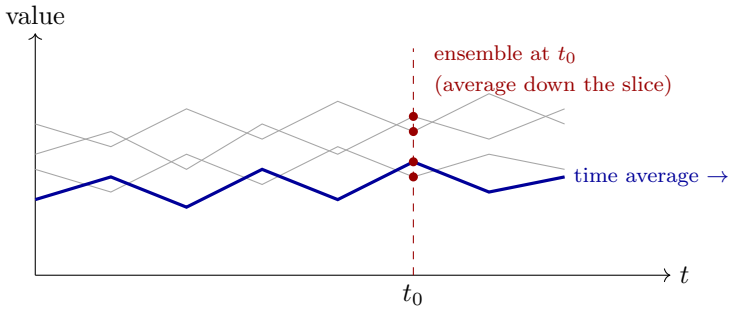


FIGURE 2. The two averages. The ensemble mean $\mathbb{E}[X_{t_0}]$ averages *down* the dashed slice, over the unobserved parallel realisations; the time average runs *along* the single bold path that is observed. Ergodicity is the statement that, for a stationary process, the second recovers the first.

Ergodicity is the property under which the time average recovers the ensemble average, so that one long sample path suffices in place of the unobservable ensemble.

DEFINITION 3.15 (Ergodicity in the mean). A covariance-stationary process (X_t) with mean μ is **ergodic in the mean** if its sample mean converges to μ in mean square,

$$\frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow{L^2} \mu \qquad (n \to \infty).$$

The name and the concept were borrowed from statistical mechanics, where the same question arose first: can the pressure of a gas, an average over an unimaginable number of molecules, be read off the time behaviour of the system? Economists inherited both the word and the predicament, since they too observe one history and must infer a distribution from it.

The simplest processes for which the time average provably recovers the ensemble mean are the independent, identically distributed sequences, where the law of large numbers does the work directly.

Exercise 3.16 (IID processes are ergodic in the mean)**.** Let $\{X_t\}$ be IID with $\mathbb{E}|X_1| < \infty$ and finite variance σ^2. Show that $\{X_t\}$ is ergodic in the mean.

6. Discrete Time, Continuous Time, Chains, and the Markov Property

The section closes by fixing the basic classification of processes and the associated terminology used in the remainder of the course.

DEFINITION 3.17 (Discrete- and continuous-time processes; chains). A stochastic process $\{X_t\}_{t\in I}$ is **discrete-time** if the index set I is countable, typically $I \subseteq \mathbb{Z}$ (the naturals or the integers), and **continuous-time** if I is an interval of $\mathbb{R}$. Independently of the index set, the process is a **chain** if its state space, the set of values each X_t can take, is at most countable.

So "discrete" and "continuous" refer to time, the index set; "chain" refers to the values. The coin of Example 3.5 is a discrete-time chain. A stock price sampled every second is discrete-time but not a chain; its idealised continuous-time limit is neither.

The final notion is the one that organises the whole theory of Markov chains: a process whose future, given its entire past, in fact depends on the past only through the present.

DEFINITION 3.18 (Markov property, discrete time). A discrete-time process $\{X_t\}_{t\geq 1}$ has the **Markov property** if, for every bounded measurable $g\colon \mathbb{R}\to\mathbb{R}$ and every $t\geq 2$,

$$\mathbb{E}[\,g(X_t)\mid X_1, X_2, \ldots, X_{t-1}\,] = \mathbb{E}[\,g(X_t)\mid X_{t-1}\,] \qquad \text{almost surely.}$$

Here the conditioning is on the σ-algebras generated by the indicated random variables (Definition 2.5). That is, given the present state X_{t-1}, the earlier history $X_1,\ldots,X_{t-2}$ carries no further information about the future.

An IID process satisfies the Markov property trivially, since X_t is independent of the entire past, so both sides equal the unconditional $\mathbb{E}[g(X_t)]$. The content of the property appears only under genuine dependence, all of which is then channelled through the present state: the AR(1) recursion $X_t = \varphi X_{t-1} + u_t$ is the leading example, and Markov chains are the systematic study of such processes.

7. Martingales

Stationarity concerns the invariance of the law in time, and the Markov property concerns the dependence of the future on the past. A third structural notion concerns neither the full law nor the dependence structure, but the conditional mean of the future: a process is a martingale when the best forecast of its next value, given all information to date, is its present value. This is the formal content of a fair game and of unpredictable increments, and it underlies the modelling of asset prices and the analysis of sequential procedures.

DEFINITION 3.19 (Martingale, submartingale, supermartingale). Let $(\mathcal{F}_t)_{t\geq 1}$ be a **filtration**, an increasing family of σ-algebras $\mathcal{F}_1\subseteq\mathcal{F}_2\subseteq\cdots$ representing the information available by time t, and let $(X_t)_{t\geq 1}$ be **adapted** (X_t measurable with respect to $\mathcal{F}_t$) and integrable ($\mathbb{E}|X_t|<\infty$). The process is a **martingale** if

$$\mathbb{E}[X_{t+1}\mid\mathcal{F}_t] = X_t \qquad \text{almost surely, for every } t,$$

a **submartingale** if $\mathbb{E}[X_{t+1}\mid\mathcal{F}_t]\geq X_t$, and a **supermartingale** if $\mathbb{E}[X_{t+1}\mid\mathcal{F}_t]\leq X_t$. Absent a stated filtration, $\mathcal{F}_t=\sigma(X_1,\ldots,X_t)$ (Definition 2.5) is understood, and the conditional expectation is that of Definition 2.11.

Taking unconditional expectations, a martingale has constant mean $\mathbb{E}[X_t]=\mathbb{E}[X_1]$, a submartingale a nondecreasing mean, and a supermartingale a nonincreasing one; (X_t) is a submartingale exactly when $(-X_t)$ is a supermartingale. The following are the standard sources of examples.

Proposition 3.20 (Standard constructions)**.** Adopt the natural filtration $\mathcal{F}_n=\sigma(\cdot)$ of the sequences below.

(i) *Sums of zero-mean terms.* If $\xi_1,\xi_2,\ldots$ are independent and integrable with $\mathbb{E}[\xi_k]=0$, then $S_n=\sum_{k=1}^{n}\xi_k$ is a martingale. If instead $\mathbb{E}[\xi_k]\geq 0$ for all k it is a submartingale, and if $\mathbb{E}[\xi_k]\leq 0$ a supermartingale.

(ii) *Products of unit-mean terms.* If $Y_1,Y_2,\ldots$ are independent, nonnegative, and integrable with $\mathbb{E}[Y_k]=1$, then $M_n=\prod_{k=1}^{n}Y_k$ is a nonnegative martingale.

(iii) *Exponential (Wald) martingale.* If $\xi_1, \xi_2, \dots$ are i.i.d. with moment generating function $m(\theta) = \mathbb{E}[e^{\theta \xi_1}] < \infty$, then for the random walk $S_n = \sum_{k=1}^{n} \xi_k$,

$$M_n = \frac{e^{\theta S_n}}{m(\theta)^n}$$

is a martingale.

PROOF. (i) As ξ_{n+1} is independent of $\mathcal{F}_n$, $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = S_n + \mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = S_n + \mathbb{E}[\xi_{n+1}]$, which is S_n, $\geq S_n$, or $\leq S_n$ according to the sign of $\mathbb{E}[\xi_{n+1}]$.

(ii) Since M_n is $\mathcal{F}_n$-measurable and Y_{n+1} is independent of $\mathcal{F}_n$, $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n \, \mathbb{E}[Y_{n+1}] = M_n$, and $M_n \geq 0$.

(iii) This is case (ii) with $Y_k = e^{\theta \xi_k}/m(\theta)$, which are independent, nonnegative, and of unit mean, $\mathbb{E}[Y_k] = m(\theta)/m(\theta) = 1$. □

A further construction underlies the rest: for any integrable X and any filtration, $M_t = \mathbb{E}[X \mid \mathcal{F}_t]$ is a martingale by the tower property of conditional expectation, the *Doob martingale* of X. In continuous time, Brownian motion and the compensated Poisson process $N_t - \lambda t$ (Definition 3.23, Definition 3.22) are martingales, the continuous and pure-jump prototypes.

Martingales formalise the absence of predictable change, and this places them at the centre of two literatures. In economics and finance, the hypothesis that asset prices are unpredictable is the statement that the discounted price is a martingale: under the efficient-markets hypothesis no trading rule based on current information earns an expected excess return, so the price is a fair game. The same structure recurs in Hall's permanent-income model, where the marginal utility of consumption is a martingale, and in no-arbitrage pricing, where discounted asset prices are martingales under a risk-neutral measure. In statistics and data science, the multiplicative martingale (ii) is the likelihood ratio, the basis of sequential testing (Wald's sequential probability ratio test) and of the anytime-valid inference built on nonnegative martingales and Ville's inequality; sums of martingale differences yield the concentration inequalities (Azuma–Hoeffding, Freedman) that bound dependent sums, and the convergence of stochastic-approximation schemes, including stochastic gradient descent and temporal-difference learning, is established by expressing their noise as martingale differences and invoking martingale convergence.

8. Lévy Processes: Poisson and Brownian Motion

The IID sequence is the archetype of a process without memory: each value independent of the rest, with the same law at every index. The continuous-time analogue requires a notion of independent, time-homogeneous randomness on $[0, \infty)$.

Independence of the values X_t across t is untenable: there are uncountably many instants, and a path built from an independent value at each is too irregular to define a usable function. The correct object of independence is not the values but the *increments* over disjoint intervals: the behaviour of the process on $[t_1, t_2]$ is independent of its behaviour on a later $[t_3, t_4]$, and the law of an increment depends only on the interval's length. This condition defines the fundamental continuous-time processes.

DEFINITION 3.21 (Lévy process). A continuous-time process $\{X_t\}_{t \geq 0}$ is a **Lévy process** if

(i) $X_0 = 0$ almost surely;

(ii) **independent increments**: for any $0 \leq t_0 < t_1 < \cdots < t_n$, the increments $X_{t_1} - X_{t_0}, \ \dots, \ X_{t_n} - X_{t_{n-1}}$ are independent;

(iii) **stationary increments**: the law of $X_{t+s} - X_t$ depends only on the length s, not on t;

(iv) its sample paths are right-continuous with left limits (*càdlàg*).

One caution relative to Definition 3.9: a Lévy process is generally not stationary in that sense. Its variance typically grows with t, so its one-dimensional law changes over time; what is stationary is the law of its *increments*. Brownian motion, below, is a standard example: it is not a stationary process, though its increments are. Independent increments also imply the **Markov property**, since the next increment is independent of the entire past and the future thus depends on the history only through the present position (Definition 3.18).

Two Lévy processes represent opposite extremes: one moves only by jumps and is otherwise constant, the other moves only continuously and never jumps. Every Lévy process is built from these two.

DEFINITION 3.22 (Poisson process). Fix a rate $\lambda > 0$. The **Poisson process** $\{N_t\}_{t\geq 0}$ is the Lévy process with $N_0 = 0$ whose increments are Poisson distributed: for every $t \geq 0$ and $s > 0$,

$$\Pr(N_{t+s} - N_t = k) = e^{-\lambda s}\,\frac{(\lambda s)^k}{k!}, \qquad k = 0, 1, 2, \dots$$

Everything else about the process is a consequence, not part of the definition. The increment over $[0, t]$ gives $N_t = N_t - N_0 \sim \text{Poisson}(\lambda t)$, so at once $\mathbb{E}[N_t] = \text{Var}(N_t) = \lambda t$. The path structure is deeper and is established when the process is studied in its own right: the sample paths may be taken as right-continuous nondecreasing step functions that rise by unit jumps, so that N_t counts the "events" in $[0, t]$ and $\{N_t\}$ is a continuous-time chain with state space $\{0, 1, 2, \dots\}$, and the waiting times between successive jumps are IID Exponential(λ).

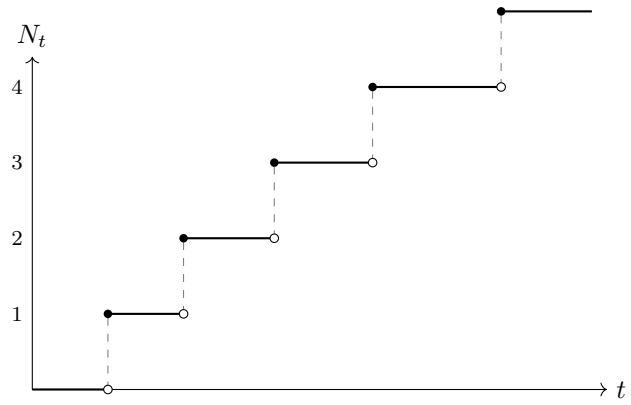


FIGURE 3. A sample path of a Poisson process: a right-continuous step function that starts at 0 and rises by unit jumps at the event times. Filled dots mark the value taken at each jump (right-continuity); open dots mark the value just before.

DEFINITION 3.23 (Brownian motion / Wiener process). A **(standard) Brownian motion**, or **Wiener process**, $\{B_t\}_{t\geq 0}$ is the Lévy process with $B_0 = 0$ and *continuous* sample paths whose increments are Gaussian:

$$B_{t+s} - B_t \sim \mathcal{N}(0, s), \qquad s > 0.$$

Again the definition carries only $B_0 = 0$, continuity, and the increment law; the familiar formulas are derived from it. The increment over $[0, t]$ gives $B_t \sim \mathcal{N}(0, t)$, whence $\mathbb{E}[B_t] = 0$ and $\text{Var}(B_t) = t$, while the covariance $\text{Cov}(B_s, B_t) = \min(s, t)$ is not assumed but computed in Exercise 3.24. That

the paths, though continuous, are differentiable at no point is a genuine theorem proved later; it is what makes Brownian motion the archetype of a purely continuous Lévy process.

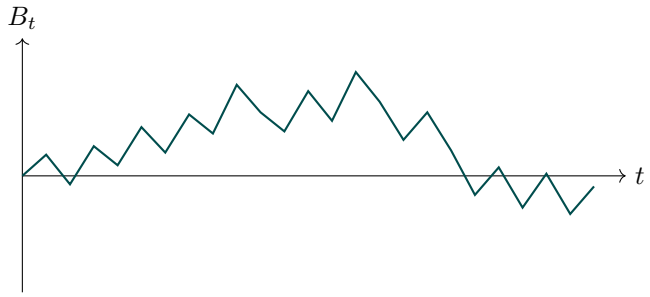


FIGURE 4. A sample path of Brownian motion: continuous everywhere, yet so jagged that it is differentiable nowhere. Over any interval of length s the increment is $\mathcal{N}(0, s)$, independent of the path's earlier behaviour.

Exercise 3.24 (Covariance of Brownian motion)**.** Let $\{B_t\}$ be a standard Brownian motion and $0 \le s \le t$. Using independent increments, show $\operatorname{Cov}(B_s, B_t) = s = \min(s, t)$.

These two processes are not merely convenient examples. The **Lévy–Itô decomposition** states that every Lévy process is the independent sum of a linear drift, a Brownian part, and a superposition of Poisson-type jumps. The Poisson process and Brownian motion are thus the two irreducible components, the pure-jump and the pure-continuous extremes, from which every continuous-time process with independent, stationary increments is assembled. The discrete-time IID sequence of the earlier sections is the analogue in discrete time: independent, identically distributed increments, here accumulating in continuous time.

Solutions to Exercises

Solution to Exercise 3.11.

For distinct $t_1, \dots, t_k$ and any shift j,

$$F_{t_1+j,\dots,t_k+j}(x_1, \dots, x_k) = \prod_{i=1}^{k} F(x_i) = F_{t_1,\dots,t_k}(x_1, \dots, x_k),$$

since each factor is the same marginal F regardless of the index. □

Solution to Exercise 3.12.

For $X_t = c$,

$$F_{t_1+j,\dots,t_k+j}(x_1, \dots, x_k) = \prod_{i=1}^{k} \mathbf{1}\{c \le x_i\} = F_{t_1,\dots,t_k}(x_1, \dots, x_k),$$

independent of the indices. For $X_t \equiv Z$,

$$F_{t_1+j,\dots,t_k+j}(x_1, \dots, x_k) = \Pr(Z \le x_1, \dots, Z \le x_k) = \Pr\Big(Z \le \min_i x_i\Big) = F_{t_1,\dots,t_k}(x_1, \dots, x_k).$$

□

Solution to Exercise 3.16.

The summands are independent with common variance σ^2, so

$$\operatorname{Var}\Big(\tfrac{1}{n}\sum_{t=1}^{n} X_t\Big) = \frac{1}{n^2}\sum_{t=1}^{n} \operatorname{Var}(X_t) = \frac{\sigma^2}{n} \longrightarrow 0,$$

so $\frac{1}{n}\sum_{t=1}^{n} X_t \to \mathbb{E}[X_1]$ in mean square, which is Definition 3.15. □

Solution to Exercise 3.24.

Write $B_t = B_s + (B_t - B_s)$. The increment $B_t - B_s$ is independent of $B_s = B_s - B_0$, so $\mathrm{Cov}(B_s, B_t - B_s) = 0$, and

$$\mathrm{Cov}(B_s, B_t) = \mathrm{Cov}(B_s, B_s) + \mathrm{Cov}(B_s, B_t - B_s) = \mathrm{Var}(B_s) + 0 = s = \min(s,t).$$

□

Appendix: The Kolmogorov Extension Theorem

This appendix states Theorem 3.7 rigorously and describes the construction it rests on: how a consistent family of finite-dimensional distributions builds a probability space and a process whose cylinder-set CDFs are exactly those distributions. It is beyond a first undergraduate course, and is recorded so the body may appeal to it precisely.

The canonical path space. Fix an index set I and take the sample space to be the set of all real-valued paths,

$$\Omega = \mathbb{R}^I = \{\, \omega = (\omega_t)_{t\in I} : \omega_t \in \mathbb{R} \,\}.$$

For $t \in I$ let $X_t \colon \Omega \to \mathbb{R}$ be the **coordinate projection** $X_t(\omega) = \omega_t$. These coordinate maps are taken as the process, so that a point of Ω *is* a sample path.

Cylinder sets and the product σ-algebra. A **cylinder set** constrains only finitely many coordinates. Given distinct $t_1, \dots, t_k \in I$ and a Borel set $B \in \mathcal{B}(\mathbb{R}^k)$,

$$C = \{\, \omega \in \Omega : (\omega_{t_1}, \dots, \omega_{t_k}) \in B \,\} = (X_{t_1}, \dots, X_{t_k})^{-1}(B).$$

The **product σ-algebra** $\mathcal{F} = \mathcal{B}(\mathbb{R})^{\otimes I}$ is the σ-algebra generated by all cylinder sets, equivalently the smallest σ-algebra making every coordinate map X_t measurable. Cylinder sets form an algebra (they are closed under complement and finite union), and this is the algebra on which the measure is first defined and then extended.

Consistency conditions. A family $\{\mu_{t_1,\dots,t_k}\}$ of Borel probability measures, one on $\mathbb{R}^k$ for each finite tuple of distinct indices (equivalently the joint CDFs $F_{t_1,\dots,t_k}$ of Definition 3.2), is **consistent** if it obeys:

K1. (Permutation invariance.) For every permutation π of $\{1,\dots,k\}$ and all Borel $B_1, \dots, B_k$,

$$\mu_{t_{\pi(1)},\dots,t_{\pi(k)}}(B_{\pi(1)} \times \cdots \times B_{\pi(k)}) = \mu_{t_1,\dots,t_k}(B_1 \times \cdots \times B_k).$$

In CDF form, $F_{t_{\pi(1)},\dots,t_{\pi(k)}}(x_{\pi(1)}, \dots, x_{\pi(k)}) = F_{t_1,\dots,t_k}(x_1, \dots, x_k)$.

K2. (Marginalisation.) Dropping the last index equals integrating out its coordinate:

$$\mu_{t_1,\dots,t_{k-1},t_k}(B_1 \times \cdots \times B_{k-1} \times \mathbb{R}) = \mu_{t_1,\dots,t_{k-1}}(B_1 \times \cdots \times B_{k-1}),$$

equivalently $F_{t_1,\dots,t_{k-1},t_k}(x_1, \dots, x_{k-1}, +\infty) = F_{t_1,\dots,t_{k-1}}(x_1, \dots, x_{k-1})$.

These are precisely the two conditions described after Definition 3.2: the indices are labels (K1), and ignoring a coordinate coincides with not having sampled it (K2). They are necessary, since any genuine process satisfies them; the theorem asserts that they are also sufficient.

Theorem 3.25 (Kolmogorov extension theorem)**.** Let I be an arbitrary index set and let $\{\mu_{t_1,\dots,t_k}\}$ be a family of Borel probability measures on the spaces $\mathbb{R}^k$, indexed by the finite tuples of distinct elements of I, satisfying the consistency conditions **K1** and **K2**. Then there exists a unique probability measure Pr on the canonical path space $(\Omega, \mathcal{F}) = (\mathbb{R}^I, \mathcal{B}(\mathbb{R})^{\otimes I})$ such that the coordinate process $X_t(\omega) = \omega_t$ has the prescribed family as its finite-dimensional

distributions: for all distinct $t_1, \dots, t_k$ and all Borel $B \subseteq \mathbb{R}^k$,

$$\Pr((X_{t_1}, \dots, X_{t_k}) \in B) = \mu_{t_1,\dots,t_k}(B).$$

In particular $\Pr(X_{t_1} \le x_1, \dots, X_{t_k} \le x_k) = F_{t_1,\dots,t_k}(x_1, \dots, x_k)$.

Sketch of the construction. The proof is a Carathéodory extension, exactly the mechanism used in Chapter 1 to build Lebesgue measure (Section 1), applied on the algebra of cylinder sets.

(1) **Define** Pr **on cylinders.** For a cylinder $C = (X_{t_1}, \dots, X_{t_k})^{-1}(B)$ set $\Pr(C) := \mu_{t_1,\dots,t_k}(B)$. Condition **K1** makes this independent of the order in which the constraining indices are listed, and **K2** makes it independent of whether extra, unconstrained coordinates are included in the description of C; together they guarantee Pr is *well defined* on cylinders and finitely additive on their algebra.

(2) **Countable additivity on the algebra.** One verifies that Pr is countably additive on the cylinder algebra. This is the one analytic step: it uses the regularity of Borel probability measures on $\mathbb{R}^k$ (inner approximation by compact sets), so that a would-be decreasing sequence of cylinders with probabilities bounded away from 0 has nonempty intersection, contradicting $\varnothing$ having measure 0.

(3) **Extend and conclude uniqueness.** The Carathéodory extension theorem then extends Pr uniquely from the cylinder algebra to the generated product σ-algebra $\mathcal{F}$. Uniqueness holds because the cylinder algebra is a π-system generating $\mathcal{F}$ on which the two measures agree. By construction the coordinate maps X_t have the prescribed finite-dimensional distributions.

Thus a consistent family of FDDs is not merely compatible with some process; it *is* a process, realised concretely on the space of its own sample paths. This is the precise sense in which, throughout the chapter, specifying the finite-dimensional distributions and checking consistency is all that defining a stochastic process ever requires.

CHAPTER 4

Markov Chains

A frog lives on two lily pads, east (e) and west (w). Each pad carries its own coin. Every morning the frog tosses the coin on the pad it currently occupies: on heads it jumps to the other pad, on tails it stays. The coin on the east pad shows heads with probability p, the coin on the west pad shows heads with probability q, and the tosses are independent across mornings. The sequence of occupied pads $X_0, X_1, X_2, \dots$, each in $\{e, w\}$, records the frog's whereabouts.

The one-morning rule is trivial to write down. The questions worth asking are not about one morning. After many mornings, how likely is the frog to be found on the east pad? Does that probability settle to a fixed number, and if so does the answer depend on where the frog began, or has the chain *forgotten* its start? A first intuition says the chain should settle; we will see that it need not, and that whether it does is decided by a single number read off the two coins. Answering these questions for the frog, by bare hands and before any general theory, is the business of this chapter's first section. Everything the later theory does for an arbitrary chain is visible here in miniature.

1. A Frog on Two Lily Pads

Write $\mu_n(e) := \Pr(X_n = e)$ and $\mu_n(w) := \Pr(X_n = w)$ for the chances of finding the frog on each pad on morning n: this is the marginal distribution of X_n, a function on the state space $\{e, w\}$. Abbreviate $e_n := \mu_n(e)$ and $w_n := \mu_n(w)$, and, with the states listed in the order (e, w), collect them in a *row* vector

$$\underline{\mu_n} := \big(e_n,\ w_n\big), \qquad e_n + w_n = 1.$$

The coin rule fixes the four conditional probabilities of tomorrow's pad given today's, grouped here by today's pad:

$$\Pr(X_{n+1} = e \mid X_n = e) = 1 - p, \qquad \Pr(X_{n+1} = w \mid X_n = e) = p,$$
$$\Pr(X_{n+1} = e \mid X_n = w) = q, \qquad \Pr(X_{n+1} = w \mid X_n = w) = 1 - q.$$

The first three days. Suppose the frog begins on the east pad, so on day 0

$$\underline{\mu_0} = (1,\ 0).$$

The law of total probability advances the distribution one day. Splitting on where the frog was the morning before,

$$e_n = \Pr(X_n = e) = \Pr(X_n = e \mid X_{n-1} = e)\, e_{n-1} + \Pr(X_n = e \mid X_{n-1} = w)\, w_{n-1} = (1-p)\, e_{n-1} + q\, w_{n-1},$$

and likewise $w_n = p\, e_{n-1} + (1 - q)\, w_{n-1}$. Applied to day 0, these give day 1,

$$\underline{\mu_1} = \big((1-p)\cdot 1 + q \cdot 0,\ \ p \cdot 1 + (1-q)\cdot 0\big) = (1-p,\ p),$$

so on the first morning the frog is on the east pad with probability $1 - p$ and on the west with probability p: it has moved exactly when its east coin came up heads. Applying the same two equations to $\underline{\mu_1}$ gives day 2,

$$\underline{\mu_2} = \big((1-p)(1-p) + q\,p,\ \ p(1-p) + (1-q)\,p\big) = \big((1-p)^2 + pq,\ \ p\,(2-p-q)\big),$$

whose two entries again sum to 1. The east-pad probability has already stopped being the naive $(1-p)$ of one step: a frog that jumped away on day 1 can return by day 2, and the term pq is exactly that return.

The total-probability step is a matrix equation. The two update equations, written together, are a single vector-matrix product, the row vector of today's probabilities multiplied *on the right* by a matrix:

$$(e_n,\ w_n) = (e_{n-1},\ w_{n-1}) \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \qquad \text{that is} \qquad \underline{\mu_n} = \underline{\mu_{n-1}}\, P, \quad P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

The matrix P is nothing more than the four conditional probabilities arranged so that each *row* is the distribution of tomorrow given a definite pad today: the first row is where the frog goes from east, the second from west. Its rows therefore sum to 1. The law of total probability is not merely *like* this vector-matrix product; for a two-state chain it *is* this product, one column of the multiplication for each destination pad.

Iterating the one-day step from day 0 collapses the whole history into a power of P:

$$\underline{\mu_1} = \underline{\mu_0} P, \quad \underline{\mu_2} = \underline{\mu_1} P = \underline{\mu_0} P^2, \quad \ldots, \quad \boxed{\underline{\mu_n} = \underline{\mu_0}\, P^n}$$

Where the frog is on any future morning is read off a single matrix raised to a power, not off a sum over the 2^n possible coin histories. The remaining question, how P^n behaves as n grows, is answered by diagonalising P.

Diagonalisation and the explicit solution. The eigenvalues of P solve $\lambda^2 - (\operatorname{tr} P)\lambda + \det P = 0$ with $\operatorname{tr} P = 2-p-q$ and $\det P = (1-p)(1-q) - pq = 1-p-q$. They are

$$\lambda_1 = 1, \qquad \lambda_2 = 1-p-q.$$

An eigenvalue of 1 is automatic: the rows of P sum to 1, so $P\mathbf{1} = \mathbf{1}$ and the all-ones column is a right eigenvector. But it is not the right eigenvectors that advance a distribution. The update $\underline{\mu_n} = \underline{\mu_{n-1}} P$ multiplies P on the *left* by a row vector, so the objects that govern the dynamics are the *left* eigenvectors, the rows u with $uP = \lambda u$. Solving $u(P - \lambda I) = 0$ gives

$$u_1 = (q,\ p) \text{ (for } \lambda_1 = 1), \qquad u_2 = (1,\ -1) \text{ (for } \lambda_2 = 1-p-q).$$

Write $s := p + q$, so $\lambda_2 = 1 - s$. Every probability row vector expands in this basis. Rescaling u_1 to a distribution gives $\underline{\pi} := \frac{1}{p+q}(q,\ p)$, and because u_2 has entries summing to zero while both $\underline{\mu_0}$ and $\underline{\pi}$ sum to one, the starting distribution is

$$\underline{\mu_0} = \underline{\pi} + c\, u_2, \qquad c = e_0 - \tfrac{q}{p+q},$$

the constant read off the east coordinate. Multiplying on the right by P^n scales each eigen-row by its eigenvalue to the power n: the $\underline{\pi}$ part is fixed for all n, while the u_2 part is multiplied by $(1-s)^n$. Hence $\underline{\mu_n} = \underline{\pi} + c\,(1-s)^n u_2$, and reading off the east coordinate gives the closed form, for any starting probability e_0,

$$e_n = \Pr(X_n = e) = \pi + (1-p-q)^{\,n}\,(e_0 - \pi), \qquad \pi := \frac{q}{p+q},$$

and $w_n = 1 - e_n$. For the frog started on the east pad, $e_0 = 1$ and

$$e_n = \frac{q}{p+q} + (1-p-q)^{\,n}\,\frac{p}{p+q}.$$

Everything about the long run is now in the single factor $(1-p-q)^n$. When both coins are genuinely random, $0 < p < 1$ and $0 < q < 1$, we have $|1-p-q| < 1$, so the second term decays and $e_n \to \pi$ no matter where the frog started: the chain forgets its origin. The *manner* of the approach depends on the sign of $1-p-q$. If $p+q < 1$ the factor is positive and e_n approaches π monotonically; if

$p+q>1$ it is negative and e_n oscillates about π while converging; in the extreme $p=q=1$ the factor is -1 and e_n never settles, alternating forever between the two pads.

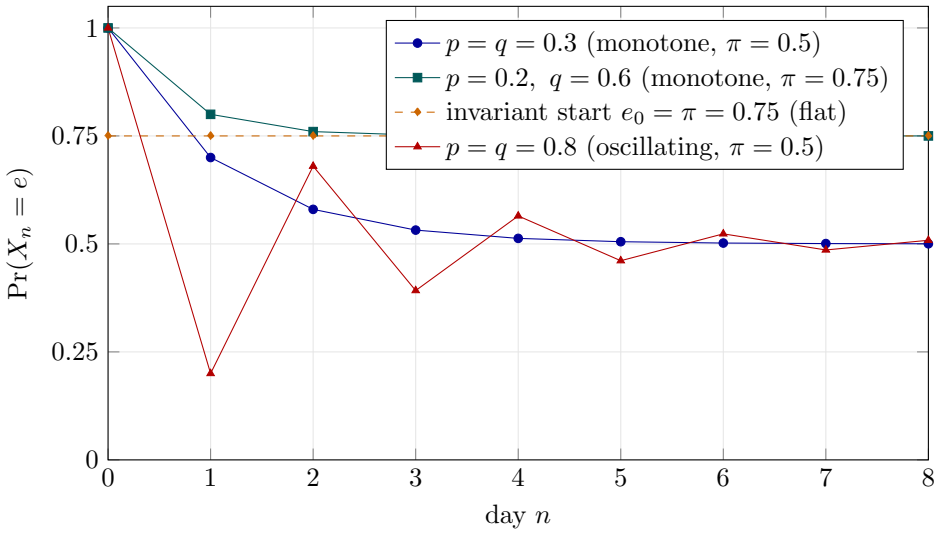


FIGURE 1. The east-pad probability $\Pr(X_n=e)=\pi+(1-p-q)^n(e_0-\pi)$ against the day n, for four choices of the coins. When $p+q<1$ the approach to $\pi=q/(p+q)$ is monotone; when $p+q>1$ it oscillates; and when the frog is started from the invariant distribution $e_0=\pi$ the probability never moves.

The invariant distribution. The limit $\pi=q/(p+q)$ is not an accident of the starting pad. The left eigenvector $u_1=(q,p)$ for $\lambda_1=1$, rescaled to a probability vector, is

$$\underline{\pi}=\frac{1}{p+q}(q,\ p)=\big(q/(p+q),\ p/(p+q)\big),\qquad\text{and}\qquad \underline{\pi}\,P=\underline{\pi}.$$

A distribution left unchanged by one day of the dynamics, $\underline{p}=\underline{p}\,P$, is called **invariant**. For the frog it is unique, and it is exactly the number the east-pad probability converges to. It is also the flat line in Figure 1: a frog whose starting pad is already drawn from $\underline{\pi}$ has $\underline{\mu_n}=\underline{\pi}P^n=\underline{\pi}$ for every n, so its east-pad probability sits at π from the very first morning and never moves.

Starting invariant makes the process stationary. Constant marginals are only the surface of the matter. If the frog begins in the invariant distribution, the whole process is *strictly stationary* (Definition 3.9): its behaviour over any block of days is unchanged by shifting the block in time.

Proposition 4.1 (Invariant start yields stationarity)**.** If $\underline{\mu_0}=\underline{\pi}$, then the frog process $\{X_n\}_{n\ge0}$ is strictly stationary.

PROOF. By repeated conditioning and the coin rule, the joint probability of any block of days factors into the starting probability and one transition factor per step: for states $i_0,\dots,i_k\in\{e,w\}$,

$$\Pr(X_0=i_0,X_1=i_1,\dots,X_k=i_k)=\Pr(X_0=i_0)\prod_{t=1}^{k}\Pr(X_t=i_t\mid X_{t-1}=i_{t-1}),$$

and each transition factor depends only on the pair (i_{t-1},i_t), not on the day t. Shifting the block forward by one day and conditioning the same way,

$$\Pr(X_1=i_0,X_2=i_1,\dots,X_{k+1}=i_k)=\Pr(X_1=i_0)\prod_{t=1}^{k}\Pr(X_t=i_t\mid X_{t-1}=i_{t-1}),$$

with the *same* transition factors. The two joint probabilities differ only in their first factor, $\Pr(X_0 = i_0)$ against $\Pr(X_1 = i_0)$. But when $\underline{\mu_0} = \underline{\pi}$ we have $\underline{\mu_1} = \underline{\mu_0}P = \underline{\pi}P = \underline{\pi} = \underline{\mu_0}$, so these agree for every state i_0; and by induction $\underline{\mu_t} = \underline{\pi}$ for all t, so the same holds for a shift by any number of days. Hence every finite-dimensional distribution is invariant under time shifts, which is strict stationarity. □

Is the stationary frog ergodic? Stationarity says the ensemble is reproducible; it does not by itself say that watching one frog for a long time recovers the ensemble. That is the ergodic question of Definition 3.15, and for the frog we can settle it by hand. Fix $0 < p < 1$ and $0 < q < 1$, start from the invariant distribution $\underline{\mu_0} = \underline{\pi}$, and track the indicator of the east pad,

$$Y_n := \mathbf{1}\{X_n = e\}, \qquad \mathbb{E}[Y_n] = \Pr(X_n = e) = \pi.$$

The long-run fraction of mornings the frog spends on the east pad is the time average $\frac{1}{N}\sum_{n=1}^{N} Y_n$. Ergodicity in the mean is the statement that this time average converges to the ensemble value π.

Proposition 4.2 (The east-pad indicator is ergodic in the mean)**.** For $0 < p < 1$, $0 < q < 1$ and $\underline{\mu_0} = \underline{\pi}$, the indicator $Y_n = \mathbf{1}\{X_n = e\}$ satisfies $\frac{1}{N}\sum_{n=1}^{N} Y_n \to \pi$ in mean square.

Proof. The process is stationary by Proposition 4.1, so Y_n has constant mean π and its autocovariances depend only on the lag. For $k \geq 0$,

$$\Pr(X_0 = e,\, X_k = e) = \Pr(X_0 = e)\,\Pr(X_k = e \mid X_0 = e) = \pi\big[\pi + (1-p-q)^k(1-\pi)\big],$$

using the explicit east-pad formula with the chain started at e. Hence

$$\gamma_k := \operatorname{Cov}(Y_0, Y_k) = \Pr(X_0 = e, X_k = e) - \pi^2 = \pi(1-\pi)\,(1-p-q)^k,$$

and by stationarity $\gamma_{-k} = \gamma_k$, so $\gamma_k = \pi(1-\pi)r^{|k|}$ with $r := 1-p-q$ and $|r| < 1$. The variance of the time average is a weighted sum of these autocovariances,

$$\operatorname{Var}\Big(\tfrac{1}{N}\sum_{n=1}^{N} Y_n\Big) = \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\gamma_{m-n} = \frac{1}{N}\sum_{k=-(N-1)}^{N-1}\Big(1 - \tfrac{|k|}{N}\Big)\gamma_k \leq \frac{1}{N}\sum_{k=-\infty}^{\infty}|\gamma_k|.$$

The final sum is a convergent geometric series,

$$\sum_{k=-\infty}^{\infty}|\gamma_k| = \pi(1-\pi)\sum_{k=-\infty}^{\infty}|r|^{|k|} = \pi(1-\pi)\,\frac{1+|r|}{1-|r|} < \infty,$$

finite precisely because $|r| < 1$, that is because both coins are genuinely random. Therefore $\operatorname{Var}(\frac{1}{N}\sum_{n=1}^{N} Y_n) \leq C/N \to 0$ for the constant $C = \pi(1-\pi)(1+|r|)/(1-|r|)$, and the time average converges to π in mean square. □

A single frog, watched long enough, reveals the ensemble fraction π: the proportion of mornings it is seen on the east pad converges to $q/(p+q)$. The correlation between distant days decays geometrically, at rate $|1-p-q|$, and that decay is what makes one long record substitute for the unobservable ensemble. All of this, stationarity and ergodicity included, was obtained for the frog by direct computation, with no general theory of Markov chains. The rest of the chapter builds that theory, so that the matrix P, its powers, and the identity $\underline{\mu_n} = \underline{\mu_0}P^n$ are available for any chain, not just this one.

2. Transition Probabilities

Throughout the rest of the chapter, $\{X_n\}_{n\geq 0}$ is a discrete-time chain (Definition 3.17) with an at most countable state space S, and it has the Markov property (Definition 3.18): given the present state, the future is independent of the past. We add one standing assumption, which holds for the frog and for the models in this course, and which is what lets the one-step numbers be tabulated once and for all.

DEFINITION 4.3 (Discrete-time Markov chain). A **discrete-time Markov chain** (DTMC) on a countable state space S is a process $\{X_n\}_{n\geq 0}$ taking values in S such that, for every $n \geq 0$ and all states $i_0, \ldots, i_{n-1}, i, j$ with $\Pr(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) > 0$,

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

and this common value does not depend on n. A chain with this last property is called **time-homogeneous**, and all chains in this chapter are assumed to be so.

The two clauses do distinct work. The Markov property says the past $i_0, \ldots, i_{n-1}$ may be discarded once the present i is known. Time-homogeneity says the transition mechanism is the same every morning, so a single table of numbers governs the chain forever. Together they reduce the chain's dynamics to one array, exactly the matrix P found for the frog.

DEFINITION 4.4 (One-step transition probabilities). For a time-homogeneous DTMC the **one-step transition probability** from state i to state j is

$$p_{ij} := \Pr(X_{n+1} = j \mid X_n = i),$$

which by homogeneity is independent of n. The **transition probability matrix** (tpm) is $P := (p_{ij})_{i,j\in S}$.

The index order is chosen so that a row distribution multiplies P on the right the way the law of total probability demands: the first index is the origin, the second the destination, and row i of P is the conditional distribution of tomorrow's state given that today's is i. Consequently

$$p_{ij} \geq 0 \quad \text{for all } i, j, \qquad \text{and} \qquad \sum_{j\in S} p_{ij} = 1 \quad \text{for every } i :$$

each *row* sums to one. A non-negative matrix whose rows sum to one is **stochastic**, and every tpm in this convention is stochastic. (Some texts index the other way, writing $\Pr(j \to i)$ and letting a column distribution multiply the transpose $P^\top$ on the left; the two conventions are transposes of one another, and we keep the row form, in which $\underline{\mu_n}$ multiplies P on the right, as it did for the frog.)

If $\underline{\mu_n} = (\Pr(X_n = i))_{i\in S}$ is the distribution written as a row vector, the law of total probability is $\underline{\mu_{n+1}} = \underline{\mu_n} P$, exactly as in Section 1. The one-step matrix answers a one-step question; the long-run questions concern many steps at once.

DEFINITION 4.5 (n-step transition probabilities). For $n \geq 0$ the **n-step transition probability** from i to j is

$$p_{ij}^{(n)} := \Pr(X_n = j \mid X_0 = i),$$

which by homogeneity equals $\Pr(X_{m+n} = j \mid X_m = i)$ for every $m \geq 0$. The **n-step transition matrix** is $P^{(n)} := \left(p_{ij}^{(n)}\right)_{i,j\in S}$.

Two boundary cases fix the notation. For $n = 0$ the chain has not moved, so $p_{ij}^{(0)} = \Pr(X_0 = j \mid X_0 = i)$ equals 1 when $i = j$ and 0 otherwise; thus $P^{(0)} = I$, the identity matrix. For $n = 1$ the definition reproduces the one-step probabilities, $p_{ij}^{(1)} = p_{ij}$, so

$$P^{(1)} = P.$$

We therefore write P, not $P^{(1)}$, for the one-step matrix, and reserve the parenthesised superscript for $n \geq 2$, where $P^{(n)}$ is so far a genuinely new object: nothing yet says it is computable from P.

3. The Chapman–Kolmogorov Equation

For the frog the n-step behaviour turned out to be a power of P. That this holds for every chain is the content of the Chapman–Kolmogorov equation. To pass from a state i to a state j in $m + n$ steps, the chain must stand *somewhere* after the first m; enumerate the possibilities for that midpoint and the multi-step probabilities factor.

Theorem 4.6 (Chapman–Kolmogorov equation)**.** For a time-homogeneous DTMC and all integers $m, n \geq 0$,

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)}\, p_{kj}^{(n)} \qquad \text{for all } i, j \in S.$$

Equivalently, in matrix form, $P^{(m+n)} = P^{(m)} P^{(n)} = P^{(n)} P^{(m)}$.

Proof. Fix states i, j and assume $\Pr(X_0 = i) > 0$, so the conditioning is defined. Condition on $X_0 = i$ throughout, and partition the event $\{X_{m+n} = j\}$ according to the state k occupied at the intermediate time m. The events $\{X_m = k\}$, over $k \in S$, are disjoint and exhaust the sample space, so by countable additivity

$$p_{ij}^{(m+n)} = \Pr(X_{m+n} = j \mid X_0 = i) = \sum_{k \in S} \Pr(X_{m+n} = j,\ X_m = k \mid X_0 = i).$$

Apply the multiplication rule to each term, splitting off the intermediate state (and dropping any k with $\Pr(X_m = k \mid X_0 = i) = 0$, whose term is zero):

$$\Pr(X_{m+n} = j,\ X_m = k \mid X_0 = i) = \Pr(X_m = k \mid X_0 = i)\ \Pr(X_{m+n} = j \mid X_m = k,\ X_0 = i).$$

The last factor simplifies by the Markov property. Given the present state $X_m = k$, the future value X_{m+n} is independent of the past value $X_0 = i$, so the condition $X_0 = i$ may be dropped:

$$\Pr(X_{m+n} = j \mid X_m = k,\ X_0 = i) = \Pr(X_{m+n} = j \mid X_m = k) = p_{kj}^{(n)},$$

the last equality by homogeneity (Definition 4.5). The first factor is $\Pr(X_m = k \mid X_0 = i) = p_{ik}^{(m)}$ by definition. Substituting,

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)}\, p_{kj}^{(n)},$$

which is the entry in row i, column j of the matrix product $P^{(m)} P^{(n)}$. Hence $P^{(m+n)} = P^{(m)} P^{(n)}$, and the two orders agree because powers of P commute. □

The one step in the argument that uses more than bookkeeping is the passage from $\Pr(X_{m+n} = i \mid X_m = k,\ X_0 = j)$ to $\Pr(X_{m+n} = i \mid X_m = k)$. This is the Markov property in the form "given the present, the future forgets the past." The defining property of Definition 4.3 is stated one step ahead and with the whole history in the conditioning; the version used here, a single past date X_0 and a future n steps beyond the present, is the same statement applied along the chain, and is developed as an exercise below.

Everything now follows by turning the additive index $m + n$ into a product of matrices.

Corollary 4.7 (The n-step matrix is the n-th power)**.** For every $n \geq 0$, $P^{(n)} = P^n$, the n-th matrix power of the one-step tpm P.

PROOF. Induct on n. For $n = 0$, $P^{(0)} = I = P^0$, and for $n = 1$, $P^{(1)} = P = P^1$. Suppose $P^{(n)} = P^n$ for some $n \geq 1$. Taking the single step $m = 1$ in Theorem 4.6,

$$P^{(n+1)} = P^{(1)} P^{(n)} = P\, P^n = P^{n+1}.$$

By induction the identity holds for all $n \geq 0$. □

Any multi-step question is now a matter of arithmetic: the probability that a chain in state i today is in state j after n steps is the (i, j) entry of P^n, computed by multiplying one small matrix by itself, not by summing over histories. The sum over paths has not disappeared; it is exactly what matrix multiplication performs, once, and reuses at every power. The distribution at any time follows in the same stroke.

Corollary 4.8 (Marginal distribution at time n)**.** Let $\underline{\mu_0} = \big(\Pr(X_0 = i)\big)_{i \in S}$ be the initial distribution, written as a row vector. Then the distribution of X_n is

$$\underline{\mu_n} = \underline{\mu_0}\, P^n, \qquad \text{that is} \qquad \Pr(X_n = j) = \sum_{i \in S} \Pr(X_0 = i)\, p_{ij}^{(n)}.$$

PROOF. By the law of total probability and Corollary 4.7,

$$\Pr(X_n = j) = \sum_{i \in S} \Pr(X_0 = i)\, \Pr(X_n = j \mid X_0 = i) = \sum_{i \in S} (\underline{\mu_0})_i\, p_{ij}^{(n)},$$

which is the j-th entry of the row vector $\underline{\mu_0}\, P^n$. □

This is the identity $\underline{\mu_n} = \underline{\mu_0} P^n$ of Section 1, now established for every time-homogeneous chain rather than derived by hand for two states.

Example 4.9 (Two steps for the frog)**.** Take the frog with $p = 0.2$ and $q = 0.6$, ordering the states (e, w), so

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix},$$

stochastic as it must be, its rows summing to one. The two-step probability of returning to the east pad reads, through Chapman–Kolmogorov, as the sum over the intermediate pad after one step,

$$p_{ee}^{(2)} = p_{ee}\, p_{ee} + p_{ew}\, p_{we} = (0.8)(0.8) + (0.2)(0.6) = 0.64 + 0.12 = 0.76,$$

the chance of staying east both mornings plus the chance of jumping west and returning. This is the (e, e) entry of P^2, and it matches the hand computation $e_2 = (1-p)^2 + pq = 0.76$ from Section 1.

Exercise 4.10 (The multi-step Markov property)**.** Using only Definition 4.3, show that for a time-homogeneous DTMC and any $m, n \geq 1$ and states i, k, j with positive conditioning probability,

$$\Pr(X_{m+n} = j \mid X_m = k,\ X_0 = i) = \Pr(X_{m+n} = j \mid X_m = k).$$

(This is the step invoked in the proof of Theorem 4.6.)

Exercise 4.11 (Three steps for the frog)**.** For the chain of Example 4.9, compute $p_{ew}^{(3)}$, first by the path sum over the two intermediate pads and then as the (e, w) entry of P^3, and check that the two agree.

4. Invariant Distributions

The frog had one distribution the dynamics never disturbed. Starting the frog there froze its marginal at $\underline{\pi}$ for every morning, and, more than that, made the whole process stationary. That distribution was found by hand for two states, as the rescaled left eigenvector of a 2×2 matrix. For an arbitrary chain we cannot diagonalise by inspection, but we do not need to: the property that singled it out is an equation, and the same equation makes sense on any state space.

Definition 4.12 (Invariant distribution). A probability mass function π on S, that is $\pi(i) \geq 0$ for all i and $\sum_{i \in S} \pi(i) = 1$, is **invariant** for the tpm P if

$$\pi = \pi P, \qquad \text{equivalently} \qquad \pi(j) = \sum_{i \in S} \pi(i)\, p_{ij} \quad \text{for every } j \in S.$$

When S is finite and ordered, π is written as a row vector, and the defining relation is the vector-matrix fixed-point equation $\pi = \pi P$. For the frog this π is the row vector $\underline{\pi}$ of Section 1.

The equation $\pi = \pi P$ is a left-eigenvector equation: an invariant distribution is a left eigenvector of P for the eigenvalue 1, normalised to a probability vector and required to be non-negative. That the eigenvalue 1 is present is guaranteed for a finite chain, since $P\mathbf{1} = \mathbf{1}$ forces 1 into the spectrum; a non-negative left eigenvector for it then exists as well. A finite chain therefore has at least one invariant distribution. Whether it is unique, and whether the marginals actually approach it, are separate questions decided by structural properties of the chain taken up later; for the frog the invariant distribution was unique and was the limit.

The word records exactly what the equation buys.

Lemma 4.13 (Invariance persists)**.** If the chain's marginal reaches an invariant distribution, $\underline{\mu_n} = \pi$ for some n, then $\underline{\mu_m} = \pi$ for all $m \geq n$.

Proof. If $\underline{\mu_m} = \pi$ then $\underline{\mu_{m+1}} = \underline{\mu_m} P = \pi P = \pi$; the claim follows by induction from $m = n$. □

Once the distribution of X_n equals π, it equals π at every later time: the distribution does not change, it is invariant. This is a statement about the marginal alone. It does not yet say the process is stationary, which is a property of all the finite-dimensional distributions (Definition 3.9), not of a single marginal, and is a strictly stronger requirement. That starting from an invariant π does make the whole process stationary is proved in Section 7.

5. Harmonic Functions

Advancing a distribution multiplies P on the left by a row vector, and the row vectors it fixes are the invariant distributions. The other multiplication is equally natural: let P act on the right, on a column vector, that is on a function of the state. Its fixed points are a different object, and one that turns questions about the chain into ordinary algebra later in the course.

DEFINITION 4.14 (Harmonic function). A function $h : S \to \mathbb{R}$, regarded when S is finite and ordered as a column vector $\underline{h} = (h(i))_{i\in S}^{\top}$, is **harmonic** for the tpm P if

$$P\underline{h} = \underline{h}, \qquad \text{equivalently} \qquad h(i) = \sum_{j\in S} p_{ij}\, h(j) \quad \text{for every } i \in S.$$

The right-hand sum is a conditional expectation. Since $p_{ij} = \Pr(X_{n+1} = j \mid X_n = i)$,

$$\sum_{j\in S} p_{ij}\, h(j) = \mathbb{E}\big[h(X_{n+1}) \mid X_n = i\big],$$

so h is harmonic exactly when its value at each state equals the expected value of h one step later:

$$h(i) = \mathbb{E}\big[h(X_{n+1}) \mid X_n = i\big].$$

A harmonic function is already equal to its own one-step average. Read along the chain, this is a martingale identity, which is the reason harmonic functions matter here.

Exercise 4.15 (Harmonic functions give martingales)**.** Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC and let $h : S \to \mathbb{R}$ be harmonic and bounded. Show that $M_n := h(X_n)$ is a martingale with respect to the natural filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$.

Remark 4.16 (Where the name comes from)**.** The term is borrowed from analysis, where the harmonic functions are the solutions of the Laplace equation $\nabla^2 v = 0$, characterised by the mean-value property: the value at a point equals the average over a surrounding sphere. The defining relation $h(i) = \sum_j p_{ij} h(j)$ is the discrete counterpart, the value at a state equalling a weighted average over the states one step away, and the two notions coincide for the random walks that discretise the Laplacian. The connection is not pursued here.

6. A Chain Is Fixed by Its Start and One Step

Kolmogorov's consistency theorem says a process is pinned down by its finite-dimensional distributions. A Markov chain, however, is handed to us through only two objects: the initial distribution $\underline{\mu_0}$ and the one-step matrix P. If every finite-dimensional distribution can be rebuilt from these two, then the entire law of the chain, and so every probabilistic question about it, is determined by the pair $(\underline{\mu_0}, P)$ alone. The next theorem shows it can.

Theorem 4.17 (Finite-dimensional distributions of a chain)**.** For a time-homogeneous DTMC, any times $0 \leq t_1 < t_2 < \cdots < t_k$, and any states $x_1, \ldots, x_k \in S$, writing $t_0 := 0$,

$$\Pr(X_{t_1} = x_1, \ldots, X_{t_k} = x_k) = \sum_{x_0\in S} \mu_0(x_0) \prod_{r=1}^{k} p_{x_{r-1}\,x_r}^{(t_r - t_{r-1})}.$$

PROOF. Condition on the starting state by the law of total probability,

$$\Pr(X_{t_1} = x_1, \ldots, X_{t_k} = x_k) = \sum_{x_0\in S} \Pr(X_0 = x_0)\, \Pr(X_{t_1} = x_1, \ldots, X_{t_k} = x_k \mid X_0 = x_0),$$

where $\Pr(X_0 = x_0) = \mu_0(x_0)$. Expand the conditional joint probability by the multiplication rule, taking the times in increasing order so that each new factor conditions on all earlier ones:

$$\Pr(X_{t_1} = x_1, \ldots, X_{t_k} = x_k \mid X_0 = x_0) = \prod_{r=1}^{k} \Pr\big(X_{t_r} = x_r \mid X_0 = x_0, X_{t_1} = x_1, \ldots, X_{t_{r-1}} = x_{r-1}\big).$$

By the multi-step Markov property (Exercise 4.10), given the most recent state $X_{t_{r-1}} = x_{r-1}$ the value X_{t_r} is independent of the earlier states, so each factor drops all conditioning but the latest:

$$\Pr\big(X_{t_r} = x_r \mid X_0 = x_0, \dots, X_{t_{r-1}} = x_{r-1}\big) = \Pr(X_{t_r} = x_r \mid X_{t_{r-1}} = x_{r-1}) = p_{x_{r-1}\,x_r}^{(t_r - t_{r-1})},$$

the last equality by homogeneity (Definition 4.5). Substituting gives the stated product. □

Every factor on the right is an entry of a power of P, since $p_{ij}^{(m)} = (P^m)_{ij}$ by Corollary 4.7. The whole finite-dimensional distribution is thus assembled from the numbers in $\underline{\mu_0}$ and the powers of the one matrix P.

Corollary 4.18 (The law is determined by $(\underline{\mu_0}, P)$)**.** The finite-dimensional distributions of a time-homogeneous DTMC are determined by the initial distribution $\underline{\mu_0}$ and the tpm P. By Kolmogorov's consistency theorem the law of $\{X_n\}_{n\ge 0}$ is therefore determined by the pair $(\underline{\mu_0}, P)$ alone.

The dependence is of the same shape as a first-order recursion: an initial condition, $\underline{\mu_0}$, together with a one-step rule, P, determines the entire trajectory of distributions and every joint law along it. This is what lets us answer any question posed about a chain once its start and its one-step matrix are given, as the following example does for a two-state chain.

Example 4.19 (Conditional and unconditional means)**.** Let $S = \{0, 1\}$, ordered $(0, 1)$, with

$$P = \begin{pmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{pmatrix}, \qquad \underline{\mu_0} = (2/3,\ 1/3).$$

Because $X_n \in \{0, 1\}$, its expectation is $\mathbb{E}[X_n] = \Pr(X_n = 1)$. From state 1 the next-step law is row 1 of P, so

$$\mathbb{E}[X_2 \mid X_1 = 1] = \sum_j j\, p_{1j} = p_{11} = \tfrac{1}{4}, \qquad \mathbb{E}[X_2 \mid X_1 = 0] = p_{01} = \tfrac{2}{3}.$$

As $\sigma(X_1)$ is generated by the two atoms $\{X_1 = 0\}$ and $\{X_1 = 1\}$, the conditional expectation is the simple function taking these two values,

$$\mathbb{E}[X_2 \mid \sigma(X_1)] = \tfrac{2}{3}\,\mathbf{1}\{X_1 = 0\} + \tfrac{1}{4}\,\mathbf{1}\{X_1 = 1\}.$$

For the unconditional mean, advance the distribution and read the second entry,

$$\underline{\mu_1} = \underline{\mu_0} P = \big(\tfrac{17}{36},\ \tfrac{19}{36}\big), \qquad \mathbb{E}[X_2] = \Pr(X_2 = 1) = \tfrac{2}{3}\cdot\tfrac{17}{36} + \tfrac{1}{4}\cdot\tfrac{19}{36} = \tfrac{193}{432}.$$

7. Invariant Distributions and Stationarity

We can now make the link between invariance and stationarity precise. A Markov chain is not stationary in general; its marginals move, as the frog's did before settling. The formula of Theorem 4.17 shows precisely which starting distribution removes that motion from the whole process, not merely from the marginals.

Theorem 4.20 (Invariant start yields stationarity)**.** Let $\{X_n\}_{n\ge 0}$ be a time-homogeneous DTMC with invariant distribution π. If $\underline{\mu_0} = \pi$, then the process is strictly stationary.

Proof. Fix times $0 \le t_1 < \cdots < t_k$, states $x_1, \dots, x_k$, and a shift $s \ge 0$. By Theorem 4.17 applied at the times $t_1, \dots, t_k$,

$$\Pr(X_{t_1} = x_1, \dots, X_{t_k} = x_k) = \sum_{x_0 \in S} \mu_0(x_0) \prod_{r=1}^{k} p_{x_{r-1}\,x_r}^{(t_r - t_{r-1})}, \qquad t_0 := 0.$$

Now shift every index by s. By homogeneity the sub-chain $(X_s, X_{s+1}, \dots)$ is again a DTMC with the same P and initial distribution $\underline{\mu_s}$, so the same theorem, applied at the times $t_1+s < \cdots < t_k+s$, gives

$$\Pr(X_{t_1+s} = x_1, \dots, X_{t_k+s} = x_k) = \sum_{x_0 \in S} \mu_s(x_0) \prod_{r=1}^{k} p_{x_{r-1}\,x_r}^{(t_r - t_{r-1})}.$$

The two products are *identical*: shifting every time by the same s leaves each gap $t_r - t_{r-1}$ unchanged, so every transition factor is the same. The expressions can differ only in the initial factor, $\mu_0(x_0)$ against $\mu_s(x_0)$. But $\underline{\mu_0} = \pi$ is invariant, so by Lemma 4.13 $\underline{\mu_s} = \pi = \underline{\mu_0}$, whence $\mu_s(x_0) = \mu_0(x_0)$ for every x_0. The two finite-dimensional distributions coincide. As this holds for all k, all times, and all shifts s, the process is strictly stationary. □

This is exactly Proposition 4.1, now for an arbitrary chain rather than the frog, and proved from the general finite-dimensional formula instead of by direct conditioning. It answers one of the recurring questions of the course. A Markov chain is not handed to us stationary; to make it stationary, start it from an invariant distribution. And since stationarity is a prerequisite for the time averages of a single realisation to mean anything, finding invariant distributions is not bookkeeping but the first step toward using a chain empirically.

8. Computing Invariant Distributions

Finding where to start a chain so that it runs stationary is, by Definition 4.12, solving the linear system $\pi = \pi P$ together with the normalisation $\sum_i \pi(i) = 1$. One caution carries the entire risk of error: the multiplication is on the *left*, by a row, so it is the *columns* of P that appear in the equations. A chain written in the opposite convention has P replaced by its transpose and every equation flipped. A structured example makes the computation concrete.

Example 4.21 (Random walk on a graph)**.** Let G have vertices $V = \{1, 2, 3, 4\}$ and edges $E = \big\{\{1,2\}, \{1,3\}, \{2,3\}, \{3,4\}\big\}$. A walker at a vertex steps next to a uniformly chosen neighbour, so $p_{vw} = 1/\deg(v)$ when $\{v, w\} \in E$ and 0 otherwise. With the degrees $\deg = (2, 2, 3, 1)$,

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

each row a uniform distribution over the walker's neighbours. Writing $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$, the equation $\pi = \pi P$ read column by column is

$$\pi_1 = \tfrac{1}{2}\pi_2 + \tfrac{1}{3}\pi_3, \quad \pi_2 = \tfrac{1}{2}\pi_1 + \tfrac{1}{3}\pi_3, \quad \pi_3 = \tfrac{1}{2}\pi_1 + \tfrac{1}{2}\pi_2 + \pi_4, \quad \pi_4 = \tfrac{1}{3}\pi_3.$$

The first two give $\pi_1 = \pi_2$, and then $\pi_1 = \frac{1}{2}\pi_1 + \frac{1}{3}\pi_3$ gives $\pi_1 = \frac{2}{3}\pi_3$; also $\pi_4 = \frac{1}{3}\pi_3$. (The third equation is redundant, as one always is: the columns of $\pi P - \pi$ sum to zero.) Normalising,

$$\big(\tfrac{2}{3} + \tfrac{2}{3} + 1 + \tfrac{1}{3}\big)\pi_3 = \tfrac{8}{3}\,\pi_3 = 1, \qquad \pi_3 = \tfrac{3}{8},$$

so

$$\pi = \big(\tfrac{1}{4},\ \tfrac{1}{4},\ \tfrac{3}{8},\ \tfrac{1}{8}\big).$$

Started from this π, the random walk is stationary. Notice that $\pi_v = \deg(v)/8$ and that $8 = 2|E|$: the stationary weight of a vertex is proportional to its degree.

That last observation is not special to this graph.

Proposition 4.22 (Stationary distribution of a random walk on a graph)**.** Let $G = (V, E)$ be a finite graph in which every vertex has at least one edge, and let $\{X_n\}$ be the random walk with $p_{vw} = 1/\deg(v)$ for $\{v, w\} \in E$ and 0 otherwise. Then

$$\pi(v) = \frac{\deg(v)}{2|E|}$$

is an invariant distribution.

PROOF. It is a distribution: by the handshake identity $\sum_v \deg(v) = 2|E|$, so $\sum_v \pi(v) = 1$, and each $\pi(v) \geq 0$. For invariance, fix a vertex w and sum over its neighbours, the only vertices from which a step can land on w:

$$(\pi P)(w) = \sum_{v \in V} \pi(v)\, p_{vw} = \sum_{v \sim w} \frac{\deg(v)}{2|E|} \cdot \frac{1}{\deg(v)} = \sum_{v \sim w} \frac{1}{2|E|} = \frac{\deg(w)}{2|E|} = \pi(w),$$

since w has exactly $\deg(w)$ neighbours. Hence $\pi = \pi P$. □

Exercise 4.23 (Stationary distribution of the two-state chain)**.** For the chain of Example 4.19, solve $\pi = \pi P$ with $\pi(0) + \pi(1) = 1$ to find its invariant distribution, and state the initial distribution $\underline{\mu_0}$ that makes the process stationary.

Exercise 4.24 (Degree formula on the path)**.** Apply Proposition 4.22 to the path graph on vertices $1 - 2 - 3 - 4 - 5$ (edges between consecutive vertices) to write down the stationary distribution of the random walk on it, and verify $\pi = \pi P$ directly for the interior vertex 3.

With stationarity now reducible to solving $\pi = \pi P$, the remaining empirical question is ergodicity. It, and the finer classification of a chain, rest on structural features: how its states communicate, how often the chain returns to a state, how long a return takes. To study returns we must condition on the chain at the moment it first reaches a state, and that moment is not a fixed clock time but is read off the trajectory as the chain runs. It is random. Whether the machinery built so far survives conditioning at such a moment is the question the rest of the chapter answers.

9. Random Times and Stopping Times

Every use of the Markov property so far has fixed the clock in advance: given X_n at a chosen time n, the future forgets the past. But the questions that make a chain worth studying rarely name a morning ahead of time. When does the frog first reach the west pad? When does the walker first return home? The time in each is not chosen; it is whatever the trajectory delivers, and so it is a random variable. If we mean to condition on the chain's state at such a moment, we must ask whether the Markov property still holds when the present is a random time. The answer will be no in general, and the reason for the failure will tell us exactly which random times to allow.

DEFINITION 4.25 (Random time). Let $(\Omega, \mathcal{F}, \Pr)$ be the probability space on which the DTMC $\{X_n\}_{n \geq 0}$ is defined. A **random time** is a measurable function $T : \Omega \to \{0, 1, 2, \dots\}$, that is a random variable valued in the index set of the process.

The values of T are times, not states: $T(\omega)$ names a morning, and the range is the index set $\{0, 1, 2, \dots\}$ rather than the state space S. The example to keep in mind is the first time the chain reaches a prescribed state.

Example 4.26 (First hitting time)**.** Fix a state $j \in S$. Define $T_j : \Omega \to \{0, 1, 2, \dots\}$ by

$$T_j(\omega) = n \iff X_0(\omega) \neq j,\ \dots,\ X_{n-1}(\omega) \neq j,\ X_n(\omega) = j,$$

the first morning on which the chain occupies state j. This prescription assigns a whole number to each ω, so it is a function; to call it a random time we must check it is $\mathcal{F}$-measurable. Since T_j takes discrete values, it suffices that each level set $\{T_j = n\}$ lie in $\mathcal{F}$, and

$$\{T_j = n\} = \{X_0 \neq j\} \cap \{X_1 \neq j\} \cap \cdots \cap \{X_{n-1} \neq j\} \cap \{X_n = j\}.$$

Each X_k is a random variable, so $\{X_k = j\} = X_k^{-1}(\{j\})$ and its complement $\{X_k \neq j\}$ are in $\mathcal{F}$; a finite intersection of $\mathcal{F}$-measurable events is $\mathcal{F}$-measurable. Hence $\{T_j = n\} \in \mathcal{F}$ for every n, and T_j is a random time.

Given a random time we can sample the chain at it. For a random time T, write

$$X_T(\omega) := X_{T(\omega)}(\omega) :$$

fixing the outcome ω fixes both the whole trajectory and the value $T(\omega)$, so X_T is a well-defined random variable, the state occupied at the random morning T.

Example 4.27 (Sampling at the first hitting time)**.** Take $S = \{0, 1, 2\}$ and one outcome ω whose trajectory begins

$$X_0(\omega), X_1(\omega), X_2(\omega), \cdots = 0, 1, 1, 2, 0, \dots$$

Reading off the first time state 2 appears, $T_2(\omega) = 3$, and then $X_{T_2}(\omega) = X_3(\omega) = 2$. This is no accident of the trajectory: by the very definition of T_j, the chain stands at j at time T_j, so

$$X_{T_j} = j \qquad \text{always,}$$

a random time but a deterministic sampled value. For other random times the sampled value X_T need not simplify.

Does the Markov property survive a random present? The Markov property, and the Chapman–Kolmogorov identity built on it, said that from a state i the chain moves on as if freshly started: for a fixed time n, $\Pr(X_{n+s} = j \mid X_0 = i_0, \dots, X_n = i) = p_{ij}^{(s)}$. The natural hope is that the same holds when the fixed present n is replaced by a random time T:

$$\Pr\big(X_{T+s} = j \mid X_0 = i_0,\ \dots,\ X_T = i\big) \stackrel{?}{=} p_{ij}^{(s)}.$$

It does not hold for every random time. One anticipating time breaks it.

Example 4.28 (An anticipating time that breaks the Markov property)**.** Fix a state j and let U_j be the *second* time the chain visits j, a random time by the argument of Example 4.26. Set

$$T := U_j - 1.$$

Then $T + 1 = U_j$, so $X_{T+1} = X_{U_j} = j$ with certainty, exactly as in Example 4.27. The one-step version of the display above, with $s = 1$ and destination k, therefore reads

$$\Pr\big(X_{T+1} = k \mid X_0 = i_0,\ \dots,\ X_T = i\big) = \Pr(j = k) = \mathbf{1}\{k = j\},$$

which is 1 when $k = j$ and 0 otherwise. But p_{ik} is a genuine transition probability, in general strictly between 0 and 1; it is easy to build a chain in which $p_{ij} \notin \{0, 1\}$. So the two sides disagree, and the Markov property fails at T.

The failure has a cause worth naming. The time $T = U_j - 1$ is defined by looking one step *ahead*: to know that T has occurred we must already know that the next visit to j is about to

happen. Such a time anticipates the future, and it is precisely the anticipation that lets X_{T+1} be predicted with certainty. We want the opposite: times whose occurrence can be declared from the past alone, without peeking ahead. That requirement, made precise, is a stopping time.

DEFINITION 4.29 (Stopping time). Let $\mathcal{F}_n := \sigma(X_0, X_1, \dots, X_n)$ be the natural filtration of the process. A random time T is a **stopping time** if

$$\{T = n\} \in \mathcal{F}_n \qquad \text{for every } n \geq 0.$$

The condition says that whether T has occurred by time n can be decided from $X_0, \dots, X_n$ alone; no coordinate beyond time n enters. The definition is purely measure-theoretic. It mentions no transition probabilities and no Markov property, so it applies to any discrete-time process, not only to chains; the Markov structure entered only as the motivation for wanting such times.

Because the $\mathcal{F}_n$ form a filtration, $\mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, the level-set condition can be traded for a cumulative one.

Lemma 4.30 (Equivalent form of the stopping condition)**.** A random time T is a stopping time if and only if $\{T \leq n\} \in \mathcal{F}_n$ for every $n \geq 0$.

PROOF. If T is a stopping time then $\{T \leq n\} = \bigcup_{m=0}^{n} \{T = m\}$, and each $\{T = m\} \in \mathcal{F}_m \subseteq \mathcal{F}_n$, so the union lies in $\mathcal{F}_n$. Conversely, if $\{T \leq n\} \in \mathcal{F}_n$ for all n, then $\{T = n\} = \{T \leq n\} \setminus \{T \leq n-1\}$, and $\{T \leq n-1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, so $\{T = n\} \in \mathcal{F}_n$. □

The first hitting time passes the test, and the anticipating time of Example 4.28 fails it, for the reason already isolated.

Example 4.31 (The first hitting time is a stopping time)**.** For T_j of Example 4.26,

$$\{T_j = n\} = \{X_0 \neq j\} \cap \cdots \cap \{X_{n-1} \neq j\} \cap \{X_n = j\}$$

is built from the coordinates $X_0, \dots, X_n$ only, so it lies in $\mathcal{F}_n$. Hence T_j is a stopping time.

Example 4.32 (The anticipating time is not a stopping time)**.** For $T = U_j - 1$ of Example 4.28, consider $\{T = 1\} = \{U_j = 2\}$, the event that the second visit to j falls at time 2. That requires exactly one visit among times $0, 1$ and a visit at time 2:

$$\{U_j = 2\} = \big(\{X_0 \neq j\} \cap \{X_1 = j\} \cap \{X_2 = j\}\big) \cup \big(\{X_0 = j\} \cap \{X_1 \neq j\} \cap \{X_2 = j\}\big).$$

Both cases name the coordinate X_2, so $\{T = 1\}$ is determined by X_0, X_1, X_2 and lies in $\mathcal{F}_2$, not in $\mathcal{F}_1$. The definition demands $\{T = 1\} \in \mathcal{F}_1$, which fails. This is the earlier anticipation made exact: we cannot declare $T = 1$ until we have seen time 2.

Exercise 4.33 (Shifting a hitting time forward stays a stopping time)**.** With U_j the second-visit time, show that $U_j + 1$ *is* a stopping time, by checking $\{U_j + 1 = n\} = \{U_j = n-1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$. Contrast this with Example 4.32: delaying a time never anticipates, but advancing one may.

Exercise 4.34 (Last visit time)**.** Let L_j be the *last* time the chain visits state j, that is the largest n with $X_n = j$ (on outcomes for which j is visited finitely often). Decide whether L_j is a stopping time, and justify your answer from Definition 4.29. (Consider what you must know about the future to declare that a visit is the last one.)

10. The Strong Markov Property

Stopping times were isolated to rule out the anticipation that broke Example 4.28. The payoff is that for them the Markov property does survive a random present: a chain observed up to a stopping time, and found at state i there, moves on afterwards as if freshly started from i. This is the strong Markov property, and it is the tool the study of first passage rests on.

To state it we need to say what "the history up to a stopping time" is.

DEFINITION 4.35 (History up to a stopping time). For a stopping time T, the **history up to** T is the collection

$$\mathcal{F}_T := \big\{A \in \mathcal{F} : A \cap \{T = t\} \in \mathcal{F}_t \text{ for every } t \geq 0\big\}.$$

These are the events whose occurrence can be decided from the chain's values up to the random time T.

A prescribed path up to T, the event that reads $\{X_0 = i_0, \dots, X_t = i\}$ on each $\{T = t\}$, is one such event; so is $\{X_T = i\}$ itself. With this in hand the theorem is stated for any past event and any starting state.

Theorem 4.36 (Strong Markov property)**.** Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC and let T be a stopping time that is finite almost surely, $\Pr(T < \infty) = 1$. Then for every $s \geq 0$, all states $i, j \in S$, and every event $A \in \mathcal{F}_T$ with $\Pr(A \cap \{X_T = i\}) > 0$,

$$\Pr\big(X_{T+s} = j \mid A,\ X_T = i\big) = p_{ij}^{(s)}.$$

In particular, conditioning on a prescribed history $X_0 = i_0, \dots, X_T = i$ gives $\Pr(X_{T+s} = j \mid X_0 = i_0, \dots, X_T = i) = p_{ij}^{(s)}$.

PROOF. It is enough to prove the joint identity for a prescribed history,

$$\Pr\big(X_0 = i_0,\ \dots,\ X_T = i,\ X_{T+s} = j\big) = p_{ij}^{(s)}\ \Pr\big(X_0 = i_0,\ \dots,\ X_T = i\big);$$

dividing by the positive probability of the conditioning event gives the stated conditional probability, and the general $A \in \mathcal{F}_T$ follows by the same steps with the history replaced by $A \cap \{X_t = i\}$. Abbreviate

$$H_t := \{X_0 = i_0,\ \dots,\ X_t = i\}\ \ (t \geq 0), \qquad H_T := \{X_0 = i_0,\ \dots,\ X_T = i\},$$

so that on the event $\{T = t\}$ one has $H_T = H_t$ and $X_{T+s} = X_{t+s}$. Then

$$\begin{aligned}
\Pr(H_T,\, X_{T+s} = j) &\overset{(1)}{=} \sum_{t\geq 0} \Pr(H_T,\, X_{T+s} = j,\, T = t) \\
&\overset{(2)}{=} \sum_{t\geq 0} \Pr(H_t,\, X_{t+s} = j,\, T = t) \\
&\overset{(3)}{=} \sum_{t\geq 0} \Pr(H_t)\ \Pr(X_{t+s} = j \mid H_t)\ \Pr(T = t \mid H_t,\, X_{t+s} = j) \\
&\overset{(4)}{=} \sum_{t\geq 0} \Pr(H_t)\ \Pr(X_{t+s} = j \mid H_t)\ \Pr(T = t \mid H_t) \\
&\overset{(5)}{=} \sum_{t\geq 0} \Pr(H_t)\, p_{ij}^{(s)}\ \Pr(T = t \mid H_t) \\
&\overset{(6)}{=} p_{ij}^{(s)} \sum_{t\geq 0} \Pr(H_t)\ \Pr(T = t \mid H_t)
\end{aligned}$$

$$\overset{(7)}{=} p_{ij}^{(s)} \sum_{t\geq 0} \Pr(H_t,\, T=t)$$

$$\overset{(8)}{=} p_{ij}^{(s)} \sum_{t\geq 0} \Pr(H_T,\, T=t)$$

$$\overset{(9)}{=} p_{ij}^{(s)}\ \Pr(H_T).$$

The justifications, step by step:

(1): T is finite almost surely, so the events $\{T=t\}$, $t\geq 0$, partition a set of probability one; marginalise over them.

(2): On $\{T=t\}$ the random index equals t, so $H_T = H_t$ and $X_{T+s} = X_{t+s}$.

(3): Definition of conditional probability, factoring the event $H_t \cap \{X_{t+s}=j\} \cap \{T=t\}$ in the order history, then future, then $\{T=t\}$.

(4): T is a stopping time, so $\{T=t\} \in \mathcal{F}_t = \sigma(X_0,\dots,X_t)$; being determined by the coordinates up to time t, its probability does not need the future value X_{t+s}, which therefore drops from the conditioning of the last factor.

(5): Markov property (Exercise 4.10): given $X_t = i$ the future forgets the earlier states, so $\Pr(X_{t+s}=j \mid H_t) = \Pr(X_{t+s}=j \mid X_t=i) = p_{ij}^{(s)}$.

(6): $p_{ij}^{(s)}$ does not depend on t; take it outside the sum.

(7): Definition of conditional probability again, $\Pr(H_t)\ \Pr(T=t \mid H_t) = \Pr(H_t \cap \{T=t\})$.

(8): On $\{T=t\}$, $H_t = H_T$; restore the random index.

(9): Marginalise back over the partition $\{T=t\}$, once more using $\Pr(T<\infty)=1$.

Dividing both sides by $\Pr(H_T) > 0$ gives the claim. □

Almost-sure finiteness entered only at the two ends, steps (1) and (9), where the partition $\{T=t\}$ was allowed to carry full probability: if T could be infinite with positive probability, that mass would be lost from the sum and the identity would fail. The stopping property entered once, at step (4), to make $\{T=t\}$ an event of the past so that the future could be dropped from its conditioning; the Markov property entered once, at step (5).

Remark 4.37 (Why the anticipating time was excluded)**.** The chain of equalities breaks for $T = U_j - 1$ at step (4). There the event $\{T=t\}$, that is $\{U_j = t+1\}$, involves the coordinate X_{t+1}, so it is not $\mathcal{F}_t$-measurable, and the future value X_{t+s} cannot be dropped from the conditioning of $\Pr(T=t \mid H_t,\, X_{t+s}=j)$.

11. First Passage Probabilities

A recurring reason to model a system as a chain is to ask when it first reaches a state: when a queue first empties, when a gambler is first ruined, when a walker first comes home. The first hitting time T_j of Example 4.26 is the random variable that records this, and its distribution is what we now name and compute.

Definition 4.38 (First passage probabilities). Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC, and let $i, j \in S$. For $n \geq 1$ the **first passage probability** from i to j in n steps is

$$f_{ij}^{(n)} := \Pr\big(X_1 \neq j,\ X_2 \neq j,\ \dots,\ X_{n-1} \neq j,\ X_n = j \mid X_0 = i\big).$$

The **eventual passage probability** from i to j is

$$f_{ij} := \Pr\big(\exists\, n \geq 1 : X_n = j \mid X_0 = i\big).$$

Read in words, $f_{ij}^{(n)}$ is the chance that, started at i, the chain lands on j for the first time at step n, having avoided j at every step before. When $i \neq j$ this is the distribution of the first hitting time T_j; when $i = j$ the count starts at time 1, so $f_{jj}^{(n)}$ is the chance of the first *return* to j at step n. The single formula covers both readings. The eventual probability f_{ij} answers a coarser question: started at i, does the chain ever reach j at all? These two are linked by one identity.

Proposition 4.39 (Eventual passage decomposes into first passages)**.** For all states $i, j \in S$,

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}.$$

Proof. Ever reaching j means reaching it for the first time at some step. Decompose the eventual event by the step of first arrival,

$$\{\exists\, n \geq 1 : X_n = j\} = \bigcup_{n=1}^{\infty} \{X_1 \neq j, \ldots, X_{n-1} \neq j,\ X_n = j\}.$$

The events on the right are disjoint: if the first arrival is at step n then $X_n = j$ while $X_m \neq j$ for $m < n$, so no outcome can have its first arrival at two different steps $n \neq n'$. Conditioning throughout on $\{X_0 = i\}$ and using countable additivity,

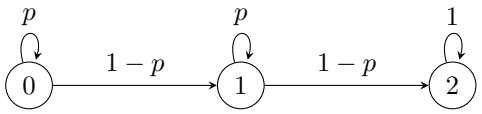
$$f_{ij} = \sum_{n=1}^{\infty} \Pr\big(X_1 \neq j, \ldots, X_{n-1} \neq j,\ X_n = j \mid X_0 = i\big) = \sum_{n=1}^{\infty} f_{ij}^{(n)}. \qquad \square$$

The identity turns a question about ever reaching a state into a sum of one-step distributions. It is easiest to feel on a small chain, where the sum can be carried out by hand. We use the diagram representation of a chain, which draws the states as nodes and the positive transition probabilities as labelled arrows out of them.

Example 4.40 (First passage on a three-state absorbing chain)**.** Take $S = \{0, 1, 2\}$ with, for a fixed $0 \leq p < 1$,

$$P = \begin{pmatrix} p & 1-p & 0 \\ 0 & p & 1-p \\ 0 & 0 & 1 \end{pmatrix},$$

a chain that at each step either stays put with probability p or advances one state with probability $1 - p$, until it reaches the absorbing state 2.



From state 0, the only way to avoid state 1 is to stay at 0, so reaching 1 for the first time at step n means $n - 1$ self-loops followed by the advance:

$$f_{01}^{(1)} = 1 - p, \qquad f_{01}^{(5)} = p^4(1-p), \qquad f_{01}^{(n)} = p^{\,n-1}(1-p).$$

Summing the geometric series, the chain started at 0 reaches 1 with certainty:

$$f_{01} = \sum_{n=1}^{\infty} p^{\,n-1}(1-p) = (1-p)\,\frac{1}{1-p} = 1.$$

The return to state 1 tells a different story. A first return at step 1 is the self-loop, $f_{11}^{(1)} = p$. A first return at step 2 is impossible, and the impossibility must be shown rather than asserted:

$$f_{11}^{(2)} = \Pr(X_1 \neq 1,\ X_2 = 1 \mid X_0 = 1).$$

From state 1 the event $\{X_1 \neq 1\}$ forces $X_1 = 2$, since $p_{10} = 0$ leaves state 2 as the only alternative. Then, conditioning and using the Markov property to drop the irrelevant past $X_0 = 1$,

$$f_{11}^{(2)} = \Pr(X_2 = 1 \mid X_1 = 2,\ X_0 = 1)\ \Pr(X_1 = 2 \mid X_0 = 1) = p_{21}\, p_{12} = 0 \cdot (1 - p) = 0.$$

The naive reading, "once the chain leaves 1 it is at 2, and 2 never returns," is correct, but the words "2 never returns" are the transition fact $p_{21} = 0$ read through the Markov property; without it the step from time 1 to time 2 could not ignore the history. The same argument kills every later return, $f_{11}^{(n)} = 0$ for all $n \geq 2$, so

$$f_{11} = \sum_{n=1}^{\infty} f_{11}^{(n)} = p.$$

Started at 1, the chain returns to 1 with probability p and with probability $1-p$ escapes forever into state 2.

From state 0 the chain is sure to reach 1, so $f_{01} = 1$; from state 1 the chain returns to itself only with probability $p < 1$, so $f_{11} < 1$. Whether $f_{jj} = 1$ or $f_{jj} < 1$, whether a state is sure to be revisited or may be left for good, is the distinction between a *recurrent* and a *transient* state.

Exercise 4.41 (First passages on the absorbing chain)**.** For the chain of Example 4.40, compute $f_{02}^{(n)}$ for $n \geq 2$ and the eventual probability f_{02}, and show $f_{22}^{(n)} = 0$ for all $n \geq 1$ except $n = 1$. Interpret $f_{22} = 1$ in the language of the preceding paragraph.

12. Recurrence and Transience

The absorbing chain of Example 4.40 already sorts its states by a single number. From state 2 the chain returns with certainty, $f_{22} = 1$; from states 0 and 1 it can leave and never come back, $f_{00}, f_{11} < 1$. It is tempting to stop here and treat $f_{jj} = 1$ as the whole story. But $f_{jj} = 1$ certifies only that a return happens, not how long it takes. Two states can both be sure to be revisited, and yet one is revisited quickly while the other keeps you waiting, on average, forever. To separate them we have to measure the return, not merely certify it.

The quantity to measure is the time of the first return. It is the first-hitting construction of Example 4.26 read from state j onward, with the clock starting at time 1 so that sitting at j initially does not count as a return.

Definition 4.42 (First return time, mean return time). Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC and $j \in S$. The **first return time** to j is the random time

$$R_j := \min\{\, n \geq 1 : X_n = j \,\}, \qquad R_j := \infty \text{ if } X_n \neq j \text{ for all } n \geq 1.$$

Started at j, its law is the first-passage sequence: for $n \geq 1$,

$$\Pr(R_j = n \mid X_0 = j) = f_{jj}^{(n)}, \qquad \Pr(R_j = \infty \mid X_0 = j) = 1 - f_{jj}.$$

When $f_{jj} = 1$, the **mean return time** to j is

$$\nu_j := \mathbb{E}[R_j \mid X_0 = j] = \sum_{n=1}^{\infty} n\, f_{jj}^{(n)} \ \in (0, \infty].$$

Each level set $\{R_j = n\} = \{X_1 \neq j, \dots, X_{n-1} \neq j, X_n = j\}$ is a finite intersection of measurable events, so R_j is a random time by the argument of Example 4.26, and $\{R_j \leq n\} \in \mathcal{F}_n$ makes it a stopping time. When $f_{jj} < 1$ the sequence $(f_{jj}^{(n)})_n$ sums to $f_{jj} < 1$: it is a *defective* distribution, with the missing mass $1 - f_{jj}$ sitting on $R_j = \infty$. In that case $\mathbb{E}[R_j \mid X_0 = j] = \infty$ for the trivial

reason that R_j is infinite with positive probability, and the mean carries no information about the size of the finite returns. This is why ν_j is defined only once $f_{jj} = 1$, where R_j is finite almost surely and its mean measures a genuine waiting time.

DEFINITION 4.43 (Classification of states). Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC on a countable state space S, and let $j \in S$.

- j is **transient** if $f_{jj} < 1$.
- j is **recurrent** if $f_{jj} = 1$. Among recurrent states,
  - j is **null recurrent** if $\nu_j = \infty$;
  - j is **positive recurrent** if $\nu_j < \infty$.

Every state is exactly one of transient or recurrent, the split turning on whether f_{jj} falls short of 1 or reaches it. A recurrent state is exactly one of null or positive recurrent, the second split turning on whether the certain return has finite or infinite expected length. A transient state is left for good with positive probability; a null recurrent state is revisited for certain but only after an unbounded average wait; a positive recurrent state is revisited for certain and, on average, soon.

Example 4.44 (Classifying the absorbing chain)**.** Return to Example 4.40. There $f_{00} = f_{11} = p < 1$, so states 0 and 1 are **transient**: once the chain advances it never falls back. State 2 is absorbing, $f_{22}^{(1)} = 1$, so $f_{22} = 1$ and 2 is recurrent. Its return is immediate, $R_2 = 1$ almost surely from state 2, hence

$$\nu_2 = \sum_{n=1}^{\infty} n\, f_{22}^{(n)} = 1 \cdot 1 = 1 < \infty,$$

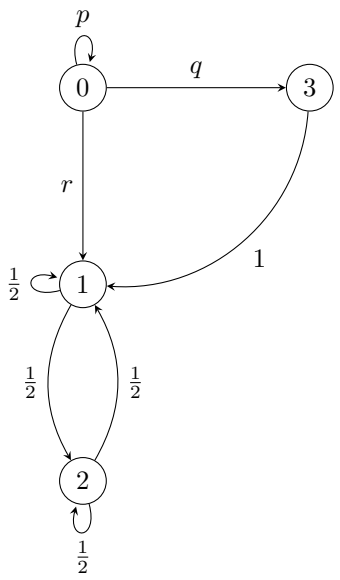
and state 2 is **positive recurrent**.

The absorbing state returns trivially. A recurrent state whose return is genuinely random, and whose mean must be summed, is more instructive.

Example 4.45 (Classifying a four-state chain)**.** Take $S = \{0, 1, 2, 3\}$ with fixed $p + q + r = 1$, $p < 1$, $q > 0$, $r > 0$, and

$$P = \begin{pmatrix} p & r & 0 & q \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

rows and columns indexed $0, 1, 2, 3$.



The four states fall into three kinds, each read off by the same first-return count.

State 0. A first return at step 1 is the self-loop, $f_{00}^{(1)} = p$. Any other departure lands in 1 or 3, from which no path returns to 0: state 1 moves only within $\{1,2\}$, and state 3 moves to 1. So a first return after step 1 is impossible, $f_{00}^{(n)} = 0$ for $n \geq 2$, and $f_{00} = p < 1$, so 0 is **transient**.

State 3. From 3 the chain moves to 1 with certainty, and 3 is reachable only from 0, which is never revisited once left. Started at 3 the chain never comes back, $f_{33} = 0 < 1$, so 3 is **transient**.

States 1 *and* 2. The pair $\{1,2\}$ is closed: from either state the chain stays inside it. A first return to 1 at step 1 is the self-loop, $f_{11}^{(1)} = \frac{1}{2}$; a first return at step $n \geq 2$ leaves to 2, holds there for the intervening $n-2$ steps, and switches back:

$$f_{11}^{(n)} = \underbrace{\tfrac{1}{2}}_{1\to 2} \cdot \underbrace{\left(\tfrac{1}{2}\right)^{n-2}}_{\text{hold at } 2} \cdot \underbrace{\tfrac{1}{2}}_{2\to 1} = \left(\tfrac{1}{2}\right)^{n}, \qquad n \geq 2,$$

which also matches $n = 1$. The chain returns with certainty, $f_{11} = \sum_{n\geq 1}(\frac{1}{2})^n = 1$, so 1 is recurrent. Its mean return time is the arithmetic–geometric sum $\sum_{n\geq 1} nx^n = x/(1-x)^2$ at $x = \frac{1}{2}$,

$$\nu_1 = \sum_{n=1}^{\infty} n\left(\tfrac{1}{2}\right)^n = \frac{\frac{1}{2}}{\left(1-\frac{1}{2}\right)^2} = \frac{\frac{1}{2}}{\frac{1}{4}} = 2 < \infty,$$

so 1 is **positive recurrent**, and by the symmetry of $\{1,2\}$ so is 2. The chain drifts through the transient states 0 and 3 and is drawn into the recurrent pair $\{1,2\}$, where it returns on average every $\nu_1 = 2$ steps, the reciprocal of the invariant weight $\pi_1 = \frac{1}{2}$ it carries there.

Remark 4.46 (Null recurrence needs infinitely many states)**.** Both examples produced positive recurrence, and this is no accident of the arithmetic. On a finite state space no state is null recurrent: a recurrent state there is always positive recurrent. Null recurrence, a certain return with infinite expected length, requires infinitely many states for the return time to spread its mass out far enough. The standard specimen is the simple symmetric random walk on $\mathbb{Z}$, which we take up once the recurrence test of the next section makes the computation tractable.

13. Counting Returns

Classifying a state through f_{jj} means computing first-passage probabilities, and the definition of $f_{jj}^{(n)}$ carries an awkward constraint: the chain must avoid j at *every* earlier step. On the two- and three-state chains above that constraint was easy to enforce by hand, but on a large or infinite chain the bookkeeping of all j-avoiding paths becomes forbidding. Meanwhile the raw n-step probability $p_{jj}^{(n)}$, the (j,j) entry of P^n, carries no such constraint and is exactly what matrix powers hand us for free. The question is whether recurrence can be decided from the unconstrained numbers $p_{jj}^{(n)}$ alone. It can, and the bridge is a single random variable that counts how often the chain comes home.

DEFINITION 4.47 (Number of returns). For $j \in S$, the **number of returns** to j is

$$M_j := \sum_{n=1}^{\infty} \mathbf{1}\{X_n = j\} \ \in \{0, 1, 2, \dots\} \cup \{\infty\},$$

the number of positive times at which the chain occupies j.

The number of returns has two faces. Averaged directly, it sums the transition probabilities; decomposed by successive returns, it is geometric. Setting the two faces equal is the whole point.

Theorem 4.48 (Expected returns sum the diagonal probabilities)**.** For every $j \in S$,

$$\mathbb{E}[M_j \mid X_0 = j] = \sum_{n=1}^{\infty} p_{jj}^{(n)}.$$

PROOF. Compute

$$\begin{aligned}
\mathbb{E}[M_j \mid X_0 = j] &\overset{(1)}{=} \mathbb{E}\Big[\sum_{n=1}^{\infty} \mathbf{1}\{X_n = j\} \,\Big|\, X_0 = j\Big] \\
&\overset{(2)}{=} \sum_{n=1}^{\infty} \mathbb{E}\big[\mathbf{1}\{X_n = j\} \mid X_0 = j\big] \\
&\overset{(3)}{=} \sum_{n=1}^{\infty} \Pr(X_n = j \mid X_0 = j) \\
&\overset{(4)}{=} \sum_{n=1}^{\infty} p_{jj}^{(n)}.
\end{aligned}$$

The justifications:

(1): Definition of M_j.

(2): The partial sums $\sum_{n=1}^{N} \mathbf{1}\{X_n = j\}$ are non-negative and non-decreasing in N, so the monotone convergence theorem permits interchanging the expectation with the infinite sum. This is not the finite linearity of expectation; for an infinite sum the interchange needs a hypothesis, and monotone non-negative summands supply one.

(3): The expectation of an indicator is the probability of its event.

(4): Definition of the n-step transition probability.

□

Remark 4.49 (The interchange is not automatic)**.** Step (2) is where a careless "by linearity of expectation" hides a real theorem. Linearity is proved for finite sums; for infinite sums the expectation and the summation cannot in general be swapped, and measure theory supplies

examples where the swap fails. Three standard hypotheses license it: monotone non-negative terms (monotone convergence), a uniform bound (bounded convergence), or domination by an integrable envelope (dominated convergence). Here the summands are indicators, so monotone convergence applies without strain.

Proposition 4.50 (The number of returns is geometric)**.** For every $j \in S$ and $k = 0, 1, 2, \ldots$,

$$\Pr(M_j = k \mid X_0 = j) = f_{jj}^{k}\,(1 - f_{jj}),$$

a geometric law. Consequently

$$\mathbb{E}[M_j \mid X_0 = j] = \begin{cases} \dfrac{f_{jj}}{1 - f_{jj}}, & f_{jj} < 1, \\ \infty, & f_{jj} = 1. \end{cases}$$

Proof. Work with the tail probabilities $\Pr(M_j \geq k \mid X_0 = j)$, decomposing each by the time of the first return R_j. For $k \geq 1$,

$$\begin{aligned} \Pr(M_j \geq k \mid X_0 = j) &\overset{(1)}{=} \sum_{t=1}^{\infty} \Pr(R_j = t,\ M_j \geq k \mid X_0 = j) \\ &\overset{(2)}{=} \sum_{t=1}^{\infty} \Pr(R_j = t \mid X_0 = j)\ \Pr(M_j \geq k - 1 \mid X_0 = j) \\ &\overset{(3)}{=} \Pr(M_j \geq k - 1 \mid X_0 = j) \sum_{t=1}^{\infty} f_{jj}^{(t)} \\ &\overset{(4)}{=} f_{jj}\ \Pr(M_j \geq k - 1 \mid X_0 = j). \end{aligned}$$

The justifications:

(1): On $\{M_j \geq 1\}$ the first return time R_j is finite, and the events $\{R_j = t\}$, $t \geq 1$, partition it; on $\{R_j = t\}$ one has $M_j = 1 +$ (returns strictly after t), so $\{M_j \geq k\}$ requires at least $k - 1$ further returns.

(2): The event $\{R_j = t\}$ is $\mathcal{F}_t$-measurable and fixes $X_t = j$; by the Markov property the chain restarts at j from time t, so the number of returns after t has the same law as M_j started at j and is independent of the past given $X_t = j$. This is the strong Markov property (Theorem 4.36) applied at R_j, carried out here by conditioning on its value.

(3): $\Pr(R_j = t \mid X_0 = j) = f_{jj}^{(t)}$ by Definition 4.42, and the return-count factor does not depend on t.

(4): $\sum_{t \geq 1} f_{jj}^{(t)} = f_{jj}$ by Proposition 4.39.

With $\Pr(M_j \geq 0 \mid X_0 = j) = 1$, induction gives $\Pr(M_j \geq k \mid X_0 = j) = f_{jj}^{k}$ for all $k \geq 0$. Differencing the tails,

$$\Pr(M_j = k \mid X_0 = j) = f_{jj}^{k} - f_{jj}^{k+1} = f_{jj}^{k}(1 - f_{jj}).$$

Summing the tails gives the mean,

$$\mathbb{E}[M_j \mid X_0 = j] = \sum_{k=1}^{\infty} \Pr(M_j \geq k \mid X_0 = j) = \sum_{k=1}^{\infty} f_{jj}^{k} = \frac{f_{jj}}{1 - f_{jj}} \quad (f_{jj} < 1),$$

which diverges to ∞ when $f_{jj} = 1$. □

Exercise 4.51 (A recurrent state is visited infinitely often)**.** Let j be recurrent, $f_{jj} = 1$. Using Proposition 4.50, show that $\Pr(M_j = \infty \mid X_0 = j) = 1$. Conclude that, started from j, the chain returns to j at an infinite increasing sequence of times, every one of them finite almost surely.

Two expressions for one expectation now sit side by side. Whenever a single quantity admits two representations, equating them yields something neither reveals alone; here it converts a hard question about first passages into an easy one about transition probabilities.

Corollary 4.52 (Recurrence test in transition probabilities)**.** For every $j \in S$,

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} = \frac{f_{jj}}{1 - f_{jj}},$$

with both sides infinite when $f_{jj} = 1$. Consequently:

(1) j is recurrent if and only if $\displaystyle\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty$;

(2) if j is transient then $\displaystyle\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$, and hence $\lim_{n\to\infty} p_{jj}^{(n)} = 0$.

Proof. The identity equates Theorem 4.48 with Proposition 4.50. For (1), $f_{jj} = 1$ makes the right side infinite, so the diagonal series diverges; conversely divergence forces $f_{jj} = 1$, since $f_{jj} < 1$ gives the finite value $f_{jj}/(1 - f_{jj})$. For (2), transience means $f_{jj} < 1$, so the series converges, and a convergent series has vanishing terms. □

The test is one-directional in its useful form. If $p_{jj}^{(n)}$ can be computed and does *not* tend to 0, then by the contrapositive of part (2) the state is recurrent at once, no first-passage sum required. If $p_{jj}^{(n)} \to 0$ the test is silent: the series may still converge or diverge, and only the full sum decides. This is the tool that will settle recurrence for the random walks where f_{jj} is out of reach but $p_{jj}^{(n)}$ is not.

Exercise 4.53 (Finitely many visits to a transient state)**.** Suppose $\sum_{n\geq 1} p_{jj}^{(n)} < \infty$. Using the first Borel–Cantelli lemma with the events $A_n = \{X_n = j\}$ under $\Pr(\,\cdot \mid X_0 = j)$, show that the chain visits j only finitely often with probability one. Reconcile this with $\mathbb{E}[M_j \mid X_0 = j] < \infty$.

Exercise 4.54 (Interarrival times are i.i.d.)**.** Let $R_j^{(1)} < R_j^{(2)} < \cdots$ be the successive return times to a recurrent state j, and let $\tau_k = R_j^{(k)} - R_j^{(k-1)}$ (with $R_j^{(0)} = 0$) be the gaps between them. Show that, started at j, the gaps $\tau_1, \tau_2, \ldots$ are independent and identically distributed, each with law $(f_{jj}^{(n)})_{n\geq 1}$. State precisely where the strong Markov property enters.

Exercise 4.55 (Restarting after a return)**.** For a recurrent state j and any $k \in S$, $s \geq 0$, prove

$$\Pr(X_{R_j+s} = k \mid R_j < \infty,\ X_0 = j) = p_{jk}^{(s)}.$$

Identify which hypothesis of Theorem 4.36 you are using and why R_j satisfies it.

14. Communicating Classes

The recurrence test (Corollary 4.52) settles one state at a time, and on a large chain even that is labour. But look again at the four-state chain of Example 4.45. The two transient states 0 and

3 were exactly the states the chain could leave for good; the two positive recurrent states 1 and 2 were exactly the closed pair it could never escape. Transience and recurrence did not fall at random across the diagram: they respected which states could reach which. This raises a question the state-by-state test cannot answer. Is recurrence a property of a state, or of a whole block of mutually reachable states? The answer, and the labour it saves, is that recurrence is a property of a class.

DEFINITION 4.56 (Reachability and communication). Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous DTMC on S, and $i, j \in S$. We say j is **reachable** from i, written $i \to j$, if $p_{ij}^{(n)} > 0$ for some $n \geq 0$. We say i and j **communicate**, written $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

The index $n = 0$ is allowed, and $p_{ii}^{(0)} = 1 > 0$ gives $i \to i$ for every state. In the transition graph, the nodes S with an arrow $i \to j$ drawn whenever $p_{ij} > 0$, reachability $i \to j$ is the existence of a directed path from i to j, which the eye reads off at a glance. Reachability keeps only that *some* $p_{ij}^{(n)}$ is positive; the number of steps and the size of the probability are discarded.

Proposition 4.57 (Communication is an equivalence relation)**.** The relation $\leftrightarrow$ is reflexive, symmetric, and transitive on S.

PROOF. Reflexivity is $p_{ii}^{(0)} = 1 > 0$, so $i \leftrightarrow i$. Symmetry is immediate, since the defining condition "$i \to j$ and $j \to i$" is unchanged under swapping i and j. For transitivity, suppose $i \leftrightarrow j$ and $j \leftrightarrow k$. Then $i \to j$ and $j \to k$ give $r, s \geq 0$ with $p_{ij}^{(r)} > 0$ and $p_{jk}^{(s)} > 0$, and by Chapman–Kolmogorov (Theorem 4.6), keeping only the intermediate state j in the sum,

$$p_{ik}^{(r+s)} = \sum_{l \in S} p_{il}^{(r)} p_{lk}^{(s)} \;\geq\; p_{ij}^{(r)} p_{jk}^{(s)} > 0,$$

so $i \to k$. The same computation with the roles of i and k reversed gives $k \to i$, whence $i \leftrightarrow k$. □

An equivalence relation partitions the set it lives on into disjoint classes, each the set of all elements related to a given one; this is the standard fact that a relation which is reflexive, symmetric, and transitive carves S into blocks $C_i = \{\, j \in S : i \leftrightarrow j \,\}$ that are either identical or disjoint, and whose union is S.

DEFINITION 4.58 (Communicating classes, closed classes, irreducibility). The equivalence classes of $\leftrightarrow$ are the **communicating classes** of the chain. A communicating class C is **closed** if $p_{ij} = 0$ for every $i \in C$ and every $j \notin C$; a class that is not closed is **open**. The chain (equivalently its tpm P) is **irreducible** if S is a single communicating class.

From a closed class the chain cannot escape: if C is closed, $i \in C$, and $j \notin C$, then $p_{ij}^{(n)} = 0$ for all n, by induction on n using $p_{ij}^{(n+1)} = \sum_l p_{il}\, p_{lj}^{(n)}$, where every term with $p_{il} > 0$ has $l \in C$ and so, inductively, $p_{lj}^{(n)} = 0$. An open class has at least one positive-probability step leading out.

Remark 4.59 (An irreducible chain is closed, vacuously)**.** When S is a single class $C = S$, there is no state $j \notin C$, so the closedness condition "$p_{ij} = 0$ for all $i \in C$, $j \notin C$" holds with no cases to check: the hypothesis $j \notin C$ is never met, and an implication with an unsatisfiable hypothesis is true. Hence every irreducible chain is closed.

Example 4.60 (Communicating classes of the four-state chain)**.** For the chain of Example 4.45, state 0 reaches 1 and 3 but is reached from neither (1 and 2 stay within $\{1,2\}$, and $3 \to 1$ only), so $C_0 = \{0\}$; likewise $C_3 = \{3\}$. The pair communicates, $1 \leftrightarrow 2$, and reaches nothing outside itself, so $C_1 = \{1,2\}$. The three communicating classes are

$$\{0\}, \qquad \{3\}, \qquad \{1,2\}.$$

The classes $\{0\}$ and $\{3\}$ are open (from 0 a step leaves to 1; from 3 a step leaves to 1), while $\{1,2\}$ is closed. The open classes are exactly the transient states found earlier, and the closed class is exactly the positive recurrent pair, the pattern the next results explain.

Exercise 4.61 (Reading classes off the four-state chain)**.** Confirm the three communicating classes of Example 4.60 directly from the tpm of Example 4.45, mark each as open or closed, and state which of Corollary 4.52's two verdicts (transient or recurrent) each class carries.

15. The Long-Run Rate of Visits

The recurrence test reads recurrence off the diagonal series $\sum_n p_{jj}^{(n)}$: finite for a transient state, infinite for a recurrent one. But a recurrent state can be null or positive recurrent, and the test cannot tell them apart, since for *both* the series diverges. The distinction lives one level finer, in the *rate* at which the series diverges. Rescaling the partial sums by n turns that rate into a finite limit, and the limit turns out to be $1/\nu_j$: it sees the mean return time, and so separates null recurrence ($\nu_j = \infty$) from positive recurrence ($\nu_j < \infty$). The same limit has a direct reading as the long-run fraction of time the chain spends at j.

Proving this needs one classical input from probability, the strong law of large numbers (Theorem 1.45), applied to the times between successive returns to a state. Those return times are non-negative with mean $\nu_j \in (0,\infty]$, so the strong law sends their running average to ν_j, read as $+\infty$ in the null recurrent case.

Theorem 4.62 (Long-run visit rate)**.** Let j be a recurrent state ($f_{jj} = 1$) and let i be any state with $f_{ij} = 1$, that is, from which j is reached almost surely. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \frac{1}{\nu_j} =: \gamma_j, \qquad \text{with the convention } \tfrac{1}{\infty} = 0.$$

The limit does not depend on i; it is 0 when j is null recurrent and $1/\nu_j > 0$ when j is positive recurrent. In particular, taking $i = j$, $\frac{1}{n}\sum_{k=1}^{n} p_{jj}^{(k)} \to \gamma_j$.

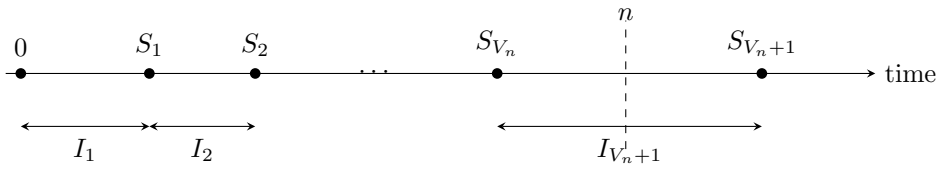
Proof. Work under $\Pr(\cdot \mid X_0 = i)$. Let $V_n := \sum_{k=1}^{n} \mathbf{1}\{X_k = j\}$ be the number of visits to j up to time n. Because $p_{ij}^{(k)} = \mathbb{E}[\mathbf{1}\{X_k = j\} \mid X_0 = i]$, summing over k turns the Cesàro average into a single expectation,

$$\frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \mathbb{E}\left[\frac{V_n}{n} \,\middle|\, X_0 = i \right],$$

and since $0 \le V_n/n \le 1$, bounded convergence lets us prove the theorem through the single almost-sure statement $V_n/n \to \gamma_j$.

The visits to j occur at an increasing sequence of times $S_1 < S_2 < \cdots$. The first, $S_1 = T_j$, is finite almost surely because $f_{ij} = 1$, and it is a stopping time (Example 4.31), so the strong Markov property (Theorem 4.36) restarts the chain afresh at j at time T_j. Writing $I_1 := S_1$ and $I_m := S_m - S_{m-1}$ for the gap before the m-th visit, the gaps $I_2, I_3, \ldots$ after this restart are the successive return times of a chain started at j, hence i.i.d. with mean $\mathbb{E}[I_2] = \nu_j \in (0,\infty]$

(Exercise 4.54, Definition 4.42). Recurrence makes j recur infinitely often (Exercise 4.51), so every S_m is finite and $V_n \to \infty$ almost surely.



The chain stands at j at the marked times, and the horizon n falls between the last visit up to time n and the next, $S_{V_n} \le n < S_{V_n+1}$. Writing each visit time as a sum of gaps, $S_m = I_1 + \cdots + I_m$, this is

$$\sum_{l=1}^{V_n} I_l \;\le\; n \;<\; \sum_{l=1}^{V_n+1} I_l,$$

and dividing by V_n, which is at least 1 once j has been reached, and taking reciprocals, sandwiches the quantity of interest,

$$\frac{V_n}{\sum_{l=1}^{V_n+1} I_l} \;<\; \frac{V_n}{n} \;\le\; \frac{V_n}{\sum_{l=1}^{V_n} I_l}.$$

Both ends converge to $1/\nu_j$. Along the deterministic index L, the strong law of large numbers (Theorem 1.45) applied to the i.i.d. gaps $I_2, I_3, \ldots$ gives $\frac{1}{L}\sum_{l=2}^{L} I_l \to \nu_j$, and the single finite term I_1 washes out, $\frac{1}{L} I_1 \to 0$, so

$$\frac{1}{L}\sum_{l=1}^{L} I_l \longrightarrow \nu_j \qquad \text{almost surely.}$$

Since $V_n \to \infty$, and a real sequence with limit ν_j keeps that limit along any integer indices tending to infinity, we may substitute $L = V_n$ and $L = V_n + 1$:

$$\frac{1}{V_n}\sum_{l=1}^{V_n} I_l \longrightarrow \nu_j, \qquad \frac{1}{V_n+1}\sum_{l=1}^{V_n+1} I_l \longrightarrow \nu_j.$$

The upper bound of the sandwich is the reciprocal of the first limit, hence tends to $1/\nu_j$; the lower bound is $\frac{V_n}{V_n+1}$ times the reciprocal of the second, and $\frac{V_n}{V_n+1} \to 1$, so it too tends to $1/\nu_j$. Therefore

$$\frac{V_n}{n} \longrightarrow \frac{1}{\nu_j} = \gamma_j \qquad \text{almost surely,}$$

read as 0 when $\nu_j = \infty$. Bounded convergence returns this to the average, $\frac{1}{n}\sum_{k\le n} p_{ij}^{(k)} \to \gamma_j$, a value fixed by j alone: zero when j is null recurrent and $1/\nu_j > 0$ when j is positive recurrent. □

Remark 4.63 (What the limit means)**.** The quantity V_n/n is the fraction of the first n steps spent at j, so Theorem 4.62 says the chain visits j a long-run fraction $\gamma_j = 1/\nu_j$ of the time: returns that take on average ν_j steps occupy a $1/\nu_j$ share of the timeline. It also sharpens the recurrence test. Where Corollary 4.52 only reported that $\sum_{k\le n} p_{jj}^{(k)}$ diverges for a recurrent state, Theorem 4.62 gives its growth rate: the partial sum behaves like $\gamma_j\, n$, which is $o(n)$ when j is null recurrent ($\gamma_j = 0$) and grows proportionally to n when j is positive recurrent ($\gamma_j > 0$).

Exercise 4.64 (Visit rate on the recurrent pair)**.** For the closed pair $\{1, 2\}$ of Example 4.45, show $p_{11}^{(k)} = \frac{1}{2}$ for every $k \ge 1$, deduce $\frac{1}{n}\sum_{k=1}^{n} p_{11}^{(k)} = \frac{1}{2}$, and check that this matches $\gamma_1 = 1/\nu_1$ with the value $\nu_1 = 2$ found in Example 4.45.

16. The Class Property Theorem

Everything is now in place to prove that recurrence, and its null and positive refinements, are properties of a whole communicating class rather than of its individual states. One state settles the fate of all the states it communicates with.

Theorem 4.65 (Class property theorem)**.** The states of a communicating class are either all transient, or all null recurrent, or all positive recurrent.

Proof. A class with a single state needs no argument, so take two distinct states $j \leftrightarrow k$ of the class. Communication gives $r, s \geq 0$ with $p_{jk}^{(r)} > 0$ and $p_{kj}^{(s)} > 0$. One way to return from j to j in $r+n+s$ steps is to go $j \to k$ in r steps, $k \to k$ in n, and $k \to j$ in s; this is a single term of the Chapman–Kolmogorov sum (Theorem 4.6), and all terms are non-negative, so

$$(\star) \qquad p_{jj}^{(r+n+s)} \;\geq\; p_{jk}^{(r)}\, p_{kk}^{(n)}\, p_{kj}^{(s)} \qquad \text{for every } n \geq 0.$$

Transience is shared. Suppose j is transient, so $\sum_n p_{jj}^{(n)} < \infty$ (Corollary 4.52). Summing $(\star)$ over $n \geq 1$, and dropping from the diagonal series the finitely many terms of index below $r+1+s$,

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} \;\geq\; \sum_{n=1}^{\infty} p_{jj}^{(r+n+s)} \;\geq\; p_{jk}^{(r)}\, p_{kj}^{(s)} \sum_{n=1}^{\infty} p_{kk}^{(n)}.$$

The far left is finite and the constants $p_{jk}^{(r)}, p_{kj}^{(s)}$ are strictly positive, so $\sum_n p_{kk}^{(n)} < \infty$ and k is transient (Corollary 4.52). Thus j transient forces k transient; by the symmetry of the argument in j and k, transience holds for all states of the class or for none, and the same is therefore true of recurrence.

Null recurrence is shared. Suppose the class is recurrent, so every state in it is recurrent and Theorem 4.62 applies on the diagonal, giving $\frac{1}{N}\sum_{n\leq N} p_{jj}^{(n)} \to \gamma_j = 1/\nu_j$ and likewise $\gamma_k = 1/\nu_k$. Averaging $(\star)$ over $n = 1, \dots, N$,

$$\frac{1}{N}\sum_{n=1}^{N} p_{jj}^{(r+n+s)} \;\geq\; p_{jk}^{(r)}\, p_{kj}^{(s)}\, \frac{1}{N}\sum_{n=1}^{N} p_{kk}^{(n)}.$$

Shifting the window on the left by the fixed offset $c = r+s$ leaves the Cesàro limit unchanged,

$$\frac{1}{N}\sum_{n=1}^{N} p_{jj}^{(n+c)} = \frac{N+c}{N}\cdot\frac{1}{N+c}\sum_{m=1}^{N+c} p_{jj}^{(m)} - \frac{1}{N}\sum_{m=1}^{c} p_{jj}^{(m)} \;\longrightarrow\; \gamma_j,$$

so passing to the limit gives $\gamma_j \geq p_{jk}^{(r)} p_{kj}^{(s)}\, \gamma_k$, that is

$$\gamma_k \;\leq\; \frac{\gamma_j}{p_{jk}^{(r)}\, p_{kj}^{(s)}}.$$

If j is null recurrent then $\gamma_j = 0$, forcing $\gamma_k = 0$; being recurrent with $\nu_k = \infty$, the state k is null recurrent too. One null recurrent state thus makes the whole class null recurrent.

Positive recurrence is what is left. Suppose finally the recurrent class contains a positive recurrent state j, so $\gamma_j > 0$. No state k of the class can be null recurrent: were it, the argument just given, run with the roles of j and k exchanged (communication supplies the positive constants either way), would force $\gamma_j = 0$, contradicting $\gamma_j > 0$. Every state of the class is therefore positive recurrent. Combined with the transient case, the class is uniformly transient, uniformly null recurrent, or uniformly positive recurrent. □

Remark 4.66 (Why the tempting limit argument is not enough)**.** It is natural to prove transience is shared by taking limits in $(\star)$ rather than sums. If j is transient then $p_{jj}^{(r+n+s)} \to 0$

(Corollary 4.52), and $(\star)$ forces $p_{jk}^{(r)} p_{kj}^{(s)} \lim_n p_{kk}^{(n)} \le 0$, hence $\lim_n p_{kk}^{(n)} = 0$. But this does *not* make k transient: vanishing terms are consistent with a divergent series. The harmonic series has $1/n \to 0$ yet $\sum 1/n = \infty$, and the simple symmetric random walk on $\mathbb{Z}$ has $p_{kk}^{(2n)} \sim 1/\sqrt{\pi n} \to 0$ while $\sum_n p_{kk}^{(2n)} = \infty$, so that walk is recurrent. Only the summed form of $(\star)$, which controls the whole series and not merely its tail, decides recurrence; this is why the proof runs through $\sum_n p_{kk}^{(n)}$ and not $\lim_n p_{kk}^{(n)}$.

Solutions to Exercises

Solution to Exercise 4.10.

The defining property (Definition 4.3) conditions on a whole history. First extract from it a one-step fact for any past event. Fix $r \ge 0$ and a state i, and let $B \in \mathcal{F}_r = \sigma(X_0, \dots, X_r)$ satisfy $B \subseteq \{X_r = i\}$ and $\Pr(B) > 0$. Such a B is a countable disjoint union of atoms $A_\alpha = \{X_0 = i_0^\alpha, \dots, X_{r-1} = i_{r-1}^\alpha,\ X_r = i\}$ of positive probability, and Definition 4.3 gives $\Pr(X_{r+1} = j \mid A_\alpha) = p_{ij}$ for each. Hence

$$\Pr(X_{r+1} = j,\ B) = \sum_\alpha \Pr(A_\alpha)\ \Pr(X_{r+1} = j \mid A_\alpha) = p_{ij} \sum_\alpha \Pr(A_\alpha) = p_{ij}\ \Pr(B),$$

so

$$(*) \qquad \Pr(X_{r+1} = j \mid B) = p_{ij} \qquad \text{whenever } B \in \mathcal{F}_r,\ B \subseteq \{X_r = i\},\ \Pr(B) > 0.$$

Now prove the claim

$$\Pr(X_{m+n} = j \mid X_m = k,\ X_0 = i) = \Pr(X_{m+n} = j \mid X_m = k)$$

by induction on $n \ge 1$. For $n = 1$ both events $\{X_0 = i,\ X_m = k\}$ and $\{X_m = k\}$ lie in $\mathcal{F}_m$ and sit inside $\{X_m = k\}$, so $(*)$ with $r = m$ evaluates each side to p_{kj}. For the step, partition on the state at time $m + n$ and split each term with the multiplication rule:

$$\Pr(X_{m+n+1} = j \mid X_m = k,\ X_0 = i) = \sum_{l \in S} \Pr(X_{m+n} = l \mid X_m = k,\ X_0 = i)\ \Pr(X_{m+n+1} = j \mid X_{m+n} = l,\ X_m = k,\ X_0 = i$$

The event $\{X_{m+n} = l,\ X_m = k,\ X_0 = i\}$ lies in $\mathcal{F}_{m+n}$ and inside $\{X_{m+n} = l\}$, so $(*)$ with $r = m+n$ makes the last factor p_{lj}, and the inductive hypothesis replaces $\Pr(X_{m+n} = l \mid X_m = k,\ X_0 = i)$ by $\Pr(X_{m+n} = l \mid X_m = k)$:

$$\Pr(X_{m+n+1} = j \mid X_m = k,\ X_0 = i) = \sum_{l \in S} \Pr(X_{m+n} = l \mid X_m = k)\, p_{lj}.$$

The identical partition applied to $\Pr(X_{m+n+1} = j \mid X_m = k)$, using $(*)$ once more, gives the same sum. The two conditional probabilities agree. □

Solution to Exercise 4.11.

With $P = \left(\begin{smallmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{smallmatrix}\right)$ on states (e, w), sum over the two intermediate pads $e \to \cdot \to \cdot \to w$:

$$\begin{aligned} e\,e\,e\,w: &\quad (0.8)(0.8)(0.2) = 0.128, & e\,e\,w\,w: &\quad (0.8)(0.2)(0.4) = 0.064, \\ e\,w\,e\,w: &\quad (0.2)(0.6)(0.2) = 0.024, & e\,w\,w\,w: &\quad (0.2)(0.4)(0.4) = 0.032, \end{aligned}$$

so $p_{ew}^{(3)} = 0.128 + 0.064 + 0.024 + 0.032 = 0.248$. As a matrix entry, from $P^2 = \left(\begin{smallmatrix} 0.76 & 0.24 \\ 0.72 & 0.28 \end{smallmatrix}\right)$ (Corollary 4.7),

$$p_{ew}^{(3)} = \left(P^2 P\right)_{ew} = p_{ee}^{(2)} p_{ew} + p_{ew}^{(2)} p_{ww} = (0.76)(0.2) + (0.24)(0.4) = 0.152 + 0.096 = 0.248,$$

in agreement. □

Solution to Exercise 4.15.

Set $M_n = h(X_n)$ and $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$. Since X_n is $\mathcal{F}_n$-measurable and h is a function on S, M_n is $\mathcal{F}_n$-measurable, so (M_n) is adapted; and h bounded gives $|M_n| \le \sup_i |h(i)| < \infty$, so M_n is integrable. For the martingale identity, expand $h(X_{n+1}) = \sum_{j \in S} h(j)\, \mathbf{1}\{X_{n+1} = j\}$ and use linearity of conditional expectation (Definition 2.11), justified for the bounded sum by conditional dominated convergence:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \sum_{j \in S} h(j)\, \mathbb{E}\big[\mathbf{1}\{X_{n+1} = j\} \mid \mathcal{F}_n\big] = \sum_{j \in S} h(j)\, \Pr(X_{n+1} = j \mid \mathcal{F}_n).$$

The Markov property (Definition 4.3), read on the atoms generating $\mathcal{F}_n$, gives $\Pr(X_{n+1} = j \mid \mathcal{F}_n) = p_{X_n j}$. Hence, by harmonicity (Definition 4.14),

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \sum_{j \in S} p_{X_n j}\, h(j) = h(X_n) = M_n.$$

So (M_n) is a martingale (Definition 3.19). □

Solution to Exercise 4.23.

With $P = \left(\begin{smallmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{smallmatrix}\right)$, the system $\pi = \pi P$ read column by column is

$$\pi_0 = \tfrac{1}{3}\pi_0 + \tfrac{3}{4}\pi_1, \qquad \pi_1 = \tfrac{2}{3}\pi_0 + \tfrac{1}{4}\pi_1.$$

The first gives $\frac{2}{3}\pi_0 = \frac{3}{4}\pi_1$, that is $\pi_0 = \frac{9}{8}\pi_1$. Normalising $\pi_0 + \pi_1 = 1$ yields $\frac{17}{8}\pi_1 = 1$, so

$$\pi = \left(\tfrac{9}{17},\ \tfrac{8}{17}\right),$$

which satisfies the second equation as a check: $\frac{2}{3} \cdot \frac{9}{17} + \frac{1}{4} \cdot \frac{8}{17} = \frac{6}{17} + \frac{2}{17} = \frac{8}{17}$. The process is stationary when started from $\underline{\mu_0} = \pi = (9/17, 8/17)$. □

Solution to Exercise 4.24.

The path $1 - 2 - 3 - 4 - 5$ has edges $\{12, 23, 34, 45\}$, so $|E| = 4$, $2|E| = 8$, and degrees $\deg = (1, 2, 2, 2, 1)$. By Proposition 4.22,

$$\pi(v) = \frac{\deg(v)}{8}, \qquad \pi = \left(\tfrac{1}{8},\ \tfrac{1}{4},\ \tfrac{1}{4},\ \tfrac{1}{4},\ \tfrac{1}{8}\right).$$

Vertex 3 has neighbours 2 and 4, each of degree 2, so

$$(\pi P)(3) = \sum_{v \sim 3} \frac{\pi(v)}{\deg(v)} = \frac{\pi(2)}{2} + \frac{\pi(4)}{2} = \frac{1/4}{2} + \frac{1/4}{2} = \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4} = \pi(3),$$

confirming $\pi = \pi P$ at the interior vertex. □

Solution to Exercise 4.33.

The second-visit time U_j is a random time, and $\{U_j = m\}$ names the two visits among times $0, \dots, m$ with the later at m, so it is built from $X_0, \dots, X_m$ and lies in $\mathcal{F}_m$. The second visit needs at least two epochs, so $U_j \ge 1$ and $U_j + 1 \ge 2$; thus $\{U_j + 1 = n\} = \varnothing \in \mathcal{F}_n$ for $n \in \{0, 1\}$. For $n \ge 2$,

$$\{U_j + 1 = n\} = \{U_j = n - 1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n.$$

In every case $\{U_j + 1 = n\} \in \mathcal{F}_n$, so $U_j + 1$ is a stopping time (Definition 4.29). Delaying refers the event back to an earlier σ-algebra, which the filtration already contains; advancing, as in Example 4.32, pushes it to a later one that $\mathcal{F}_n$ does not. □

Solution to Exercise 4.34.

L_j is *not* a stopping time. To declare $L_j = n$ one must know both that $X_n = j$ and that j is never visited again:

$$\{L_j = n\} = \{X_n = j\} \cap \bigcap_{m > n} \{X_m \neq j\},$$

an event that names coordinates X_m for all $m > n$. To see it fails Definition 4.29, take any chain in which, from j, a later return to j has positive probability, and pick two outcomes ω, ω' that agree on $X_0, \dots, X_n$ with $X_n = j$, where after time n the chain never revisits j along ω but does along ω'. Then $\omega \in \{L_j = n\}$ while $\omega' \notin \{L_j = n\}$, yet $(X_0, \dots, X_n)$ takes the same values on both. Any set in $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ has an indicator that is a function of $(X_0, \dots, X_n)$, so it cannot separate ω from ω'; hence $\{L_j = n\} \notin \mathcal{F}_n$. The obstruction is exactly the anticipation ruled out in Definition 4.29: lastness is a statement about the entire future. □

Solution to Exercise 4.41.

Take the chain of Example 4.40, $P = \left(\begin{smallmatrix} p & 1-p & 0 \\ 0 & p & 1-p \\ 0 & 0 & 1 \end{smallmatrix}\right)$ on states $\{0, 1, 2\}$. Started at 0, reaching 2 for the first time at step n forces the path to sit at 0 for some s steps ($0 \le s \le n-2$), cross to 1, hold at 1 for the remaining $n - s - 2$ steps, then cross to 2 at step n; state 2 is avoided before step n automatically, since 2 is reached only from 1. Each such path has probability

$$\underbrace{p^s}_{\text{hold at } 0} \; \underbrace{(1-p)}_{0\to 1} \; \underbrace{p^{n-s-2}}_{\text{hold at } 1} \; \underbrace{(1-p)}_{1\to 2} = (1-p)^2 p^{n-2},$$

and there are $n - 1$ choices of $s \in \{0, \dots, n-2\}$, so for $n \ge 2$

$$f_{02}^{(n)} = (n-1)(1-p)^2 p^{n-2}.$$

Summing, with $m = n - 2$,

$$f_{02} = \sum_{n=2}^{\infty} (n-1)(1-p)^2 p^{n-2} = (1-p)^2 \sum_{m=0}^{\infty} (m+1) p^m = (1-p)^2 \cdot \frac{1}{(1-p)^2} = 1,$$

so from 0 the chain reaches the absorbing state 2 with certainty. For f_{22}, state 2 is absorbing, $p_{22} = 1$, so $f_{22}^{(1)} = 1$; a first return at $n \ge 2$ would need $X_1 \ne 2$ started from 2, impossible since $X_1 = 2$ almost surely, hence $f_{22}^{(n)} = 0$ for $n \ge 2$. Therefore $f_{22} = 1$: state 2 is **recurrent** (indeed absorbing), the one revisited-for-certain state against which the transient 0 and 1, with $f_{00}, f_{11} < 1$, are contrasted. □

Solution to Exercise 4.51.

When $f_{jj} = 1$, Proposition 4.50 gives $\Pr(M_j = k \mid X_0 = j) = f_{jj}^k (1 - f_{jj}) = 0$ for every finite k, so $\Pr(M_j < \infty \mid X_0 = j) = \sum_{k \ge 0} \Pr(M_j = k \mid X_0 = j) = 0$, and therefore $\Pr(M_j = \infty \mid X_0 = j) = 1$. Almost surely, then, $X_n = j$ for infinitely many n; listing those times in increasing order gives an infinite sequence $S_1 < S_2 < \cdots$ of finite visit times. □

Solution to Exercise 4.53.

Work throughout under $\Pr(\cdot \mid X_0 = j)$. With $A_n = \{X_n = j\}$,

$$\Pr(A_n) = \Pr(X_n = j \mid X_0 = j) = p_{jj}^{(n)}, \qquad \sum_{n=1}^{\infty} \Pr(A_n) = \sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$$

by hypothesis. The First Borel–Cantelli Lemma (Theorem 1.21) gives $\Pr\big(\limsup_n A_n\big) = 0$. Now $\limsup_n A_n = \bigcap_N \bigcup_{n \ge N} A_n$ is exactly the event that $X_n = j$ for infinitely many n, so the chain visits j only finitely often with probability one.

To reconcile with $\mathbb{E}[M_j \mid X_0 = j] < \infty$: the number of returns is $M_j = \sum_{n \ge 1} \mathbf{1}\{X_n = j\}$, so $\{M_j = \infty\} = \limsup_n A_n$, which we have just shown has probability zero. This is consistent with Theorem 4.48, since a non-negative random variable with finite expectation is finite almost surely; both routes conclude that j is visited finitely often. □

Solution to Exercise 4.54.

Write $t_i = n_1 + \cdots + n_i$ for the partial sums, so that the event $\{\tau_1 = n_1, \ldots, \tau_i = n_i\}$ fixes $R_j^{(i)} = t_i$ and $X_{t_i} = j$. We prove, by induction on k, that for all $n_1, \ldots, n_k \geq 1$

$$\Pr\big(\tau_1 = n_1, \ldots, \tau_k = n_k \mid X_0 = j\big) = \prod_{i=1}^{k} f_{jj}^{(n_i)}.$$

For $k = 1$, $\tau_1 = R_j$ and $\Pr(\tau_1 = n_1 \mid X_0 = j) = f_{jj}^{(n_1)}$ by Definition 4.42. For the step, condition on the first $k-1$ gaps, which fix $R_j^{(k-1)} = t_{k-1}$. The next gap is

$$\{\tau_k = n_k\} = \{X_{t_{k-1}+1} \neq j, \ldots, X_{t_{k-1}+n_k-1} \neq j,\ X_{t_{k-1}+n_k} = j\}.$$

Because $R_j^{(k-1)}$ is a stopping time, the conditioning event $\{\tau_1 = n_1, \ldots, \tau_{k-1} = n_{k-1}\}$ is $\mathcal{F}_{t_{k-1}}$-measurable and fixes $X_{t_{k-1}} = j$. The Markov property at the now-determined time t_{k-1} therefore drops the earlier history and restarts the chain at j:

$$\Pr\big(\tau_k = n_k \mid \tau_1 = n_1, \ldots, \tau_{k-1} = n_{k-1},\ X_0 = j\big) = \Pr\big(X_1 \neq j, \ldots, X_{n_k} = j \mid X_0 = j\big) = f_{jj}^{(n_k)}.$$

Multiplying by the inductive hypothesis gives the product form. Since the joint law factors into identical marginals $f_{jj}^{(\cdot)}$, the gaps are independent and identically distributed with law $(f_{jj}^{(n)})_{n\geq 1}$. The strong Markov property enters at exactly the italicised step: it is what lets the chain restart afresh at the random time $R_j^{(k-1)}$, realised here by summing the ordinary Markov property over the value $\{R_j^{(k-1)} = t_{k-1}\}$, precisely the device of Theorem 4.36 and Proposition 4.50. Recurrence ($f_{jj} = 1$) makes each $R_j^{(k)}$ finite almost surely, so the construction never stalls. □

Solution to Exercise 4.55.

Apply the strong Markov property (Theorem 4.36) at $T = R_j$. Its hypothesis is $\Pr(T < \infty) = 1$; here j is recurrent, so $f_{jj} = 1$ and $\Pr(R_j < \infty \mid X_0 = j) = 1$ by Definition 4.42, which is exactly that hypothesis. The time R_j is a stopping time (Definition 4.42), and $\{X_0 = j\} \in \mathcal{F}_{R_j}$, since $\{X_0 = j\} \cap \{R_j = t\} \in \mathcal{F}_t$ for every t. By construction $X_{R_j} = j$ always. Taking the event $A = \{X_0 = j\}$ and state $i = j$ in Theorem 4.36,

$$\Pr\big(X_{R_j+s} = k \mid R_j < \infty,\ X_0 = j\big) = \Pr\big(X_{R_j+s} = k \mid A,\ X_{R_j} = j\big) = p_{jk}^{(s)},$$

the first equality because $X_{R_j} = j$ is automatic and $\Pr(R_j < \infty \mid X_0 = j) = 1$. □

Solution to Exercise 4.61.

From the tpm of Example 4.45, the positive entries are p_{00}, p_{01}, p_{03}; p_{11}, p_{12}; p_{21}, p_{22}; and p_{31}. Reading arrows, $0 \to 1, 3$ but nothing outside $\{1, 2\}$ returns to 0, so $C_0 = \{0\}$; similarly $3 \to 1$ only and nothing returns to 3, so $C_3 = \{3\}$; and $1 \leftrightarrow 2$ with no exit, so $C_1 = \{1, 2\}$. The classes are $\{0\}$, $\{3\}$, $\{1, 2\}$. Since $p_{01} > 0$ leaves $\{0\}$ and $p_{31} > 0$ leaves $\{3\}$, both singletons are open; $\{1, 2\}$ is closed, as its rows put all mass inside $\{1, 2\}$. By Corollary 4.52 the open classes carry the transient verdict ($f_{00}, f_{33} < 1$) and the closed class the recurrent one ($f_{11} = f_{22} = 1$). □

Solution to Exercise 4.64.

Restricted to the closed pair $\{1, 2\}$ the tpm is $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, which is idempotent: $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, so P^k restricted to $\{1, 2\}$ equals P for every $k \geq 1$ and $p_{11}^{(k)} = \frac{1}{2}$. Hence

$$\frac{1}{n}\sum_{k=1}^{n} p_{11}^{(k)} = \frac{1}{n}\sum_{k=1}^{n} \tfrac{1}{2} = \tfrac{1}{2} \qquad \text{for every } n,$$

so the limit is $\frac{1}{2}$. This matches Theorem 4.62: $\gamma_1 = 1/\nu_1 = 1/2$ with $\nu_1 = 2$ from Example 4.45, and $\frac{1}{2}$ is indeed the long-run fraction of time the chain spends at state 1. □