Introduction to Stochastic Processes

Version July 26, 2020

Timo Seppäläinen

Benedek Valkó

NOT FOR DISTRIBUTION!

Contents

Chapter 1

Random variables and stochastic processes

1.1. Probability spaces, random variables, and stochastic processes

Kolmogorov's axioms. Random outcomes are modeled mathematically in terms of a probability space $(\Omega, \mathcal{F}, P)$. These are Kolmogorov's axioms for probability theory.

Definition 1.1. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ with the following three components.

(a) Ω is a set, called the **sample space**. It is the set of all the possible outcomes. Elements of Ω are called **sample points** and typically denoted by ω.

(b) Subsets of Ω are called **events**. The collection of events in Ω is denoted by $\mathcal{F}$. Technically speaking $\mathcal{F}$ is something called a σ**-algebra**, but this point is not relevant for our course.

(c) P is a function from $\mathcal{F}$ into real numbers, called the **probability measure**. Each event A has a probability $P(A)$, and P satisfies the following axioms.

(c.1) $0 \le P(A) \le 1$ for each event $A \in \mathcal{F}$.

(c.2) $P(\varnothing) = 0$ and $P(\Omega) = 1$.

(c.3) If $\{A_k\}_{1 \le k < \infty}$ is a sequence of pairwise disjoint events then

$$P\biggl(\bigcup_{k=1}^{\infty} A_k\biggr) = \sum_{k=1}^{\infty} P(A_k). \tag{1.1}$$

△

Often used consequences of the axioms include the equation for complementary probability

$$P(A^c) = 1 - P(A), \tag{1.2}$$

monotonicity

$$P(A) \le P(B) \quad \text{if} \quad A \subset B, \tag{1.3}$$

and countable subadditivity: for any sequence of events $\{A_k\}_{1 \le k < \infty}$,

$$P\biggl(\bigcup_{k=1}^{\infty} A_k\biggr) \le \sum_{k=1}^{\infty} P(A_k). \tag{1.4}$$

This last statement is proved as Lemma B.1 in Appendix B.

We illustrate the concepts with this finite probability space.

Example 1.2 (Fair coin flips)**.** We model three fair coin flips. Encode the outcomes as 0 for heads and 1 for tails. Then the sample space of all outcomes is the Cartesian product space of triples of 0s and 1s:

$$\begin{aligned}\Omega = \{0,1\}^3 &= \{\omega = (\omega_1, \omega_2, \omega_3) : \text{ each } \omega_i = 0 \text{ or } 1\} \\ &= \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}.\end{aligned}$$

The individual outcomes $\omega = (\omega_1, \omega_2, \omega_3)$ are the sample points. The assumption of fair coin tosses dictates equal probability for each outcome: $P\{\omega\} = 1/8$ for each $\omega \in \Omega$. Events can be expressed in English or in set notation, for example,

$$B = \{\text{exactly two tails}\} = \{(0,1,1), (1,0,1), (1,1,0)\}.$$

The probability of an event comes by additivity:

$$P(B) = P(\text{exactly two tails}) = \sum_{\omega :\, \omega \in B} P\{\omega\} = \tfrac{3}{8}.$$

△

Random variables and their distributions.

Definition 1.3. A function X from a sample space Ω into a set $\mathcal{S}$ is an $\mathcal{S}$-valued **random variable**. The **probability distribution** of the random variable X is the function μ defined for subsets B of $\mathcal{S}$ by $\mu(B) = P(X \in B)$, the probability that the value of X lies in the set B. △

The probability distribution μ of a random variable satisfies properties (c.1)–(c.3) of Definition 1.1. So μ itself is a probability measure on the space $\mathcal{S}$.

The random variables most commonly discussed in introductory probability texts are real-valued, that is, where $\mathcal{S}$ is the real line $\mathbb{R}$ or some subset of it. If $\mathcal{S}$ is a subset of n-dimensional Euclidean space $\mathbb{R}^n$, then X can be called a *random vector* and sometimes the notation is altered to $\mathbf{X}$. An $\mathbb{R}^n$-valued random vector is of the form $\mathbf{X} = (X_1, X_2, \dots, X_n)$ where the coordinates $X_1, X_2, \dots, X_n$ are real-valued random variables. The probability distribution of an n-dimensional random vector $\mathbf{X}$ is the function μ on subsets of $\mathbb{R}^n$ defined by $\mu(B) = P(\mathbf{X} \in B)$ for $B \subset \mathbb{R}^n$. The probability distribution of the random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is also called the *joint distribution* of the random variables $X_1, X_2, \dots, X_n$.

Remark 1.4 (Technical remark about Borel sets)**.** When $\mathcal{S}$ is a finite or countably infinite set, $\mu(B) = P(X \in B)$ can usually be defined for all subsets $B \subset \mathcal{S}$. But if $\mathcal{S}$ is the real line or some other uncountable space, $\mu(B)$ typically cannot be defined for *all* subsets B of $\mathcal{S}$. In such cases there is a class of sets B for which the values $\mu(B)$ are defined, for *any* probability distribution μ. These are called *Borel sets* and their collection is another σ-algebra, the Borel σ-algebra of $\mathbb{R}$. This technical point is valid for all our subsequent statements about probabilities on uncountable spaces. But this issue can be safely ignored for the treatment in this book because *all reasonable sets that arise in practice are Borel sets.* △

The natural descriptions of the distribution μ of a random variable X depend on the type of the range space $\mathcal{S}$. We go over the most important cases.

Every real-valued random variable X has a *cumulative distribution function* defined by

$$F(t) = P(X \le t) \qquad \text{for real } t. \tag{1.5}$$

The probabilities of left-open right-closed intervals $B = (a, b]$ can be read directly from the cumulative distribution function through the identity

$$P(X \in B) = P(X \in (a, b]) = F(b) - F(a). \tag{1.6}$$

It is a deeper measure-theoretic fact that the function F determines $P(X \in B)$ for *all* (Borel) subsets B of $\mathbb{R}$. Precisely speaking, this means that if the cumulative distribution functions F_X and F_Y of two random variables X and Y satisfy $F_X(t) = F_Y(t)$ for all $t \in \mathbb{R}$, then all probabilities of these two random variables agree: $P(X \in B) = P(Y \in B)$ for *all* (Borel) subsets B of $\mathbb{R}$.

The (*absolutely*) *continuous* random variables form a special class of real-valued random variables. The defining property of a continuous random variable X is that its cumulative distribution function F satisfies

$$F(t) = \int_{-\infty}^{t} f(y)\,dy \qquad \text{for all } t \in \mathbb{R}$$

for a function f called the *probability density function* of X. It follows that every probability $P(X \in B)$ for a (Borel) subset B of $\mathbb{R}$ can be computed by integrating the density function:

$$P(X \in B) = \int_B f(y)\,dy.$$

The cumulative distribution function and the probability density function have natural extensions to $\mathbb{R}^n$-valued random vectors.

If the range space $\mathcal{S}$ is a finite or countably infinite set, then any $\mathcal{S}$-valued random variable X is called a *discrete random variable*. Such a set $\mathcal{S}$ is also called a *discrete set*. The *probability mass function* p of an $\mathcal{S}$-valued discrete random variable X is defined by

$$p(x) = P(X = x) \quad \text{for points } x \in \mathcal{S}. \tag{1.7}$$

The distribution μ of X is determined by the identity

$$\mu(B) = P(X \in B) = \sum_{x \in B} p(x)$$

for subsets B of $\mathcal{S}$. The set $\mathcal{S}$ can be a subset of real numbers but it does not have to be.

Remark 1.5 (Notation)**.** As already indicated above, these objects can be decorated with a subscript to indicate the random variable which they describe. For example, μ_X is the probability distribution of X, f_Y is the probability density function of Y, and p_W is the probability mass function of W. This becomes necessary if the discussion concerns many different random variables. △

Example 1.6 (Fair coin flips, continuing Example 1.2)**.** Random variables that can be defined on the sample space of Example 1.2 include the following.

For $i \in \{1, 2, 3\}$, we let X_i denote the outcome of the ith flip. As a function on Ω, X_i is defined by

$$X_i(\omega) = \omega_i \quad \text{for sample point } \omega = (\omega_1, \omega_2, \omega_3).$$

X_i is a function from Ω into the set $\{0, 1\}$, abbreviated by $X_i : \Omega \to \{0, 1\}$. The probability distribution of X_i can be represented as follows:

$$\mu_{X_i}(k) = P(X_i = k) = \tfrac{1}{2} \quad \text{for } k \in \{0, 1\}.$$

X_i is an example of a Bernoulli random variable with parameter $\frac{1}{2}$, abbreviated $X_i \sim \text{Ber}(\frac{1}{2})$.

Let S denote the total number of tails in the three flips. As a function on Ω,

$$S(\omega) = \omega_1 + \omega_2 + \omega_3 \quad \text{for sample point } \omega = (\omega_1, \omega_2, \omega_3)$$

and hence $S : \Omega \to \{0, 1, 2, 3\}$. S can also be defined in terms of the X_i random variables by $S = X_1 + X_2 + X_3$. S has the binomial distribution with parameters 3 and $\frac{1}{2}$, given by

$$\mu_S(k) = P(S = k) = \binom{3}{k} \left(\tfrac{1}{2}\right)^3 \quad \text{for } k \in \{0, 1, 2, 3\}.$$

This is abbreviated by $S \sim \text{Bin}(3, \frac{1}{2})$. △

Expectation. The *expectation* or *mean* EX of a real-valued random variable X has two representations, depending on whether X is discrete or continuous: in the discrete case

$$E[X] = \sum_k k\, p(k)$$

where the sum runs over those values k for which $P(X = k) > 0$, and in the continuous case

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

where f is the density function of X. Important extensions of these formulas work for functions of random variables. Let g be a function defined on the range of X. Then if X is discrete,

$$E[g(X)] = \sum_x g(x) P(X = x)$$

where the sum ranges over the possible values of X which can now be points in an arbitrary space $\mathcal{S}$. If X is real-valued and has density function f, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

There is a completely general definition of EX that does not rely on having either a probability mass function or a probability density function, but this definition belongs in the realm of measure-theoretic probability. The formulas above are sufficient for our purposes.

Example 1.7 (Fair coin flips, continuing Example 1.6)**.** In the three fair coin flips of Example 1.6, the expected number of tails equals

$$E[S] = E[X_1 + X_2 + X_3] = E[X_1] + E[X_2] + E[X_3] = \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} = \tfrac{3}{2}.$$

Above we used the additivity of the expectation and then the calculation

$$E[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = \tfrac{1}{2}.$$

Suppose we win a prize of $g(k)$ dollars when we get k tails altogether. Then the random prize amount equals $g(S)$. The expected prize amount is

$$E[g(S)] = \sum_{k=0}^{3} g(k) P(S = k) = \sum_{k=0}^{3} g(k) \binom{3}{k} \tfrac{1}{8}.$$

△

Example 1.8 (Random hotel)**.** To give an example of a random variable whose values are not numbers, suppose $\mathcal{S} = \{a, b, c\}$ is a set of three distinct hotels and X is the hotel that the traveler chooses. A possible probability distribution of X is $P(X = a) = \frac{1}{2}$, $P(X = b) = \frac{1}{3}$, and $P(X = c) = \frac{1}{6}$, reflecting the probability that a particular hotel is chosen. The expectation EX does not make sense now since the values of X are not numbers. But if $g : \mathcal{S} \to \mathbb{R}$ is a real-valued function on $\mathcal{S}$, then the expectation $E[g(X)]$ does make sense. For example, if g gives the nightly rate of each hotel as $g(a) = 80$, $g(b) = 60$ and $g(c) = 120$, then the expected cost of the hotel night is

$$E[g(X)] = \sum_{x \in \mathcal{S}} g(x) P(X = x) = 80 \cdot \tfrac{1}{2} + 60 \cdot \tfrac{1}{3} + 120 \cdot \tfrac{1}{6} = 80.$$

A random vector can also have other than real-valued coordinates. In the hotel example above, the random vector $X = (X_1, X_2, X_3)$ could represent the hotel choices of three successive trips. This random vector X is $\mathcal{S}^3$-valued. △

Stochastic processes.

Definition 1.9. A **stochastic process** is a collection of random variables $\{X_i : i \in \mathcal{I}\}$ all defined on the same probability space $(\Omega, \mathcal{F}, P)$.

In the general definition above the *index set* $\mathcal{I}$ is allowed to be completely arbitrary. When the index set is *finite*, for example $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$, we usually do not call the collection $\{X_1, X_2, \ldots, X_n\}$ a “stochastic process” but instead use the term random vector $\mathbf{X} = (X_{i_1}, X_{i_2}, \ldots, X_{i_n})$.

In our examples the index often represents time. There are two different types of time-indexing, discrete time and continuous time.

Discrete time is typically represented by the index set $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$ of nonnegative integers or some subset thereof. The stochastic process $\{X_k\}_{k \in \mathbb{Z}_{\geq 0}} = \{X_0, X_1, X_2, \ldots\}$ is then a *sequence* of random variables. In applications the random variables $X_0, X_1, X_2, \dots$ can represent for example outcomes of successive trials such as flips of a coin or rolls of a die, or some measurements such as the closing daily values of the S&P 500 stock index.

Discrete-time stochastic processes studied in this and later chapters of this book include *i.i.d. processes*, *random walks*, *discrete-time Markov chains*, and *martingales*.

Continuous time is represented by the index set $\mathbb{R}_{\geq 0} = [0, \infty)$ of nonnegative reals. The process is then denoted by $\{X_t\}_{0 \leq t < \infty}$. In an application the random variable X_t can represent for example the number of customers that have arrived at a service station during time interval $[0, t]$.

Continuous-time stochastic processes studied in this and later chapters of this book include *renewal processes*, *Poisson processes*, and *continuous-time Markov chains*.

Often in a stochastic process all the random variables have the same range space $\mathcal{S}$. Then $\mathcal{S}$ is called the *state space* of the stochastic process. For example, if $X_0, X_1, X_2, \dots$ represent the outcomes of successive rolls of a die, then the value of each X_k is an integer in the range $\{1, 2, 3, 4, 5, 6\}$, and we say that the state space of the process $\{X_k\}_{k \geq 0}$ is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. Elements of $\mathcal{S}$ can be called *states* or *points*.

Exactly as for random variables, our ultimate interest lies in calculating probabilities of various events and expectations of various functions determined by a stochastic process.

When we describe the probabilities $P(X \in B)$ associated to a single random variable X, we often focus on intervals $B = (a, b]$ and $B = (-\infty, b]$ whose probabilities are determined by the cumulative distribution function through (1.5) and (1.6). This is justified by the crucial fact that these simple cases determine the probabilities $P(X \in B)$ of *all* (Borel) subsets B of $\mathbb{R}$.

There is no counterpart of the cumulative distribution function in the mathematical theory of stochastic processes. But there is again a theoretical fact that allows us to focus on probabilities of relatively simple, tractable events. To illustrate, consider the case where the index set is $\mathbb{Z}_{\geq 0}$ and the state space $\mathcal{S} = \mathbb{R}$, the real line. Then realizations of the stochastic process $\{X_k : k \in \mathbb{Z}_{\geq 0}\}$ are sequences $(x_k)_{k \in \mathbb{Z}_{\geq 0}} = (x_0, x_1, x_2, \dots)$ where each entry x_k is a real number. The notion of the probability distribution extends naturally from random variables and random vectors to stochastic processes: knowing the probability distribution of the process $\{X_k : k \in \mathbb{Z}_{\geq 0}\}$ means knowing all the probabilities of the form

$$P\big\{(X_0, X_1, X_2, \dots) \in B\big\}$$

where B is a set of sequences of real numbers.

Such sets B are harder to imagine than subsets of $\mathbb{R}$ or $\mathbb{R}^n$. Fortunately it turns out that *all* the probabilities of a stochastic process are entirely determined by the finite-dimensional probabilities of the form

$$P\big\{(X_0, X_1, \dots, X_n) \in B\big\}$$

where n varies across positive integers and B varies across subsets of $\mathbb{R}^{n+1}$. In fact, even more is true: all the probabilities are uniquely determined by knowing probabilities of the form

$$P\big(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n\big) \tag{1.8}$$

where n varies across positive integers and $A_0, A_1, \dots, A_n$ vary across subintervals of $\mathbb{R}$. These are deep facts from measure theory. Their practical significance is enormous. To repeat: the probabilities of all events of a stochastic process are completely determined if we know all probabilities of the form (1.8). Thus, even though we have an infinite index set, we can work mostly with probabilities that involve only finitely many random variables.

1.2. Independent identically distributed random variables

Structurally, the simplest type of process is the one where the random variables $X_0, X_1, X_2, \dots$ are independent and they all have the same probability distribution. The acronym *i.i.d.* is short for *independent and identically distributed*.

Definition 1.10. A stochastic process $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ is an **i.i.d. process** if the random variables $X_0, X_1, X_2, \dots$ are independent and identically distributed.

The property that characterizes an i.i.d. process is that

$$P\big(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n\big) = \prod_{i=0}^{n} P(X_0 \in A_i) \tag{1.9}$$

for all positive integers n and all subsets A_i of the state space $\mathcal{S}$. Identity (1.9) actually combines two steps: independence gives the product property

$$P\big(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n\big) = \prod_{i=0}^{n} P(X_i \in A_i) \tag{1.10}$$

and then identical distribution implies that $P(X_i \in A_i) = P(X_0 \in A_i)$ for each i.

In particular, identical distribution of the random variables $X_0, X_1, X_2, \dots$ implies that

- for a given subset B of $\mathcal{S}$, the probability $P(X_k \in B)$ has the same value for all indices k;
- for a given function $g : \mathcal{S} \to \mathbb{R}$, the expectation $E[g(X_k)]$ has the same value for all indices k.

Remark 1.11 (Notation and terminology)**.** The notation $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ for a stochastic process indexed by $\mathbb{Z}_{\geq 0}$ is quickly simplified to $\{X_k\}$ when the index set is understood or not important for the discussion. We shall reduce the notation even further and write phrases such as "the stochastic process X_n" even though strictly speaking X_n is just a single random variable. But the sentence indicates that an entire collection $\{X_n\}$ is meant. This mild misuse of language is analogous to saying "the function $f(x)$" even though $f(x)$ is just one particular value of the function. The more accurate expression would be "the function f". △

When the random variables of an i.i.d. process are discrete with state space $\mathcal{S}$, identical distribution is equivalent to saying that they all have a common probability mass function $P(X_k = x)$ for all indices k. Then we have a very convenient formula for any *finite-dimensional joint distribution* of the process: for any $n \in \mathbb{Z}_{\geq 0}$ and any states $x_0, x_1, \ldots, x_n \in \mathcal{S}$,

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=0}^{n} P(X_0 = x_k). \tag{1.11}$$

Example 1.12 (Independent trials)**.** Suppose we repeat independent trials with success probability p. Each trial has two outcomes, success and failure. Particular examples could be flips of a fair coin where success means the outcome heads ($p = \frac{1}{2}$) or rolls of a fair die where success means rolling a six ($p = \frac{1}{6}$). For $k = 1, 2, 3, \ldots$, let

$$X_k = \begin{cases} 1, & \text{if trial } k \text{ is a success} \\ 0, & \text{if trial } k \text{ is a failure.} \end{cases}$$

Then $\{X_k\}_{k \in \mathbb{Z}_{>0}}$ is a discrete-time, discrete-space i.i.d. stochastic process with state space $\mathcal{S} = \{0, 1\}$. Finite-dimensional joint distributions obey the equation

$$P(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n) = p^{\sum_{i=1}^n a_i}(1-p)^{n - \sum_{i=1}^n a_i} \tag{1.12}$$

for any n-tuple $(a_1, a_2, \ldots, a_n)$ of zeros and ones. To illustrate, here is a particular case of the equation (1.12) above:

$$\begin{aligned} &P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1, X_5 = 0) \\ &\qquad = p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3(1-p)^2. \end{aligned}$$

This process can be fully characterized by saying that *the random variables* $\{X_k\}$ *are i.i.d. Ber*(p) *random variables.*

Out of these same ingredients we can define another stochastic process $\{S_n : n \geq 0\}$ that keeps track of the cumulative number of successes. Put $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$ for $n \geq 1$. This process S_n has neither the independence nor the identical distribution property. △

Strong law of large numbers. A classical problem in probability is the study of the asymptotic behavior of the partial sum $X_1 + \cdots + X_n$ of i.i.d. real-valued random variables. The *strong law of large numbers* (SLLN) is one of the most fundamental limit theorem of probability theory. It says that the average of independent, identically distributed observations converges to the mean as the number of observations tends to infinity. We state the basic version below and then an extension useful for our purposes. For the statements employ the very standard notation

$$S_n = X_1 + \cdots + X_n$$

for the sum of n random variables. The assumption is that the common mean of the i.i.d. random variables is finite.

Theorem 1.13. (Strong law of large numbers) *Let* $\{X_k\}$ *be i.i.d. random variables and assume that the absolute mean* $E|X_k|$ *is finite. Set* $\mu = E(X_1)$*. Then the limit*

$$\lim_{n \to \infty} \frac{S_n}{n} = \mu \tag{1.13}$$

holds with probability one.

The linearity of expectation implies that if the random variables $\{X_n\}$ are identically distributed with a common expected value $\mu = E[X_1]$, then the mean of S_n/n is the same:

$$E[n^{-1}S_n] = n^{-1}\big(E[X_1] + \cdots + E[X_n]\big) = n^{-1} \cdot n\mu = \mu.$$

Thus the message of the SLLN is that the average $\frac{S_n}{n}$ of a large number of i.i.d. variables is itself close to its mean.

Example 1.14 (Independent trials, continuing Example 1.12)**.** For independent trials, the SLLN gives the limit $n^{-1}S_n \to p$ as $n \to \infty$, with probability one. △

The assumption of a finite mean can be dropped if the random variables are nonnegative. This is useful for us in the sequel.

Theorem 1.15. *Let $\{X_k\}$ be nonnegative i.i.d. random variables. Then with probability one,*

$$\lim_{n\to\infty} \frac{S_n}{n} = E(X_1) \tag{1.14}$$

regardless of whether the expectation $E(X_1)$ is finite or infinite.

Example 1.16. Let $\{X_k\}_{k\in\mathbb{Z}_{\geq 1}}$ be i.i.d. random variables with probability mass function

$$P(X_k = m) = \frac{1}{m(m+1)} \quad \text{for integers } m \geq 1.$$

First note that the values sum to one so this is a legitimate probability mass function:

$$\begin{aligned}\sum_{m=1}^{\infty} \frac{1}{m(m+1)} &= \sum_{m=1}^{\infty}\left(\frac{1}{m} - \frac{1}{m+1}\right) = \lim_{n\to\infty}\sum_{m=1}^{n}\left(\frac{1}{m} - \frac{1}{m+1}\right)\\ &= \lim_{n\to\infty}\left(1 - \frac{1}{n+1}\right) = 1.\end{aligned}$$

Next calculate the expectation:

$$E[X_k] = \sum_{m=1}^{\infty} m\,P(X_k = m) = \sum_{m=1}^{\infty}\frac{m}{m(m+1)} = \sum_{m=1}^{\infty}\frac{1}{m+1} = \infty.$$

Theorem 1.16 now gives the limit

$$\lim_{n\to\infty}\frac{S_n}{n} = \infty \quad \text{with probability one.}$$

△

1.3. Renewal process limit theorem

The remainder of this chapter discusses our first class of stochastic processes, the renewal processes. These are constructed from i.i.d. random variables. From the SLLN we can already derive interesting and nontrivial asymptotics for this class of processes.

Imagine that we track the times of occurrences of some event that repeats randomly but always in the same circumstances. For concrete examples, think about replacements of a burned-out light bulb, arrivals of a bus, or phone calls coming to a help desk. For $n \geq 1$ let S_n denote the time of the nth occurrence. For convenience put also $S_0 = 0$ to mark the beginning of the process, but S_0 is *not* counted as an occurrence. The times between successive occurrences are $X_n = S_n - S_{n-1}$ for $n \geq 1$. These can be called *inter-arrival times*, *waiting times*, or *cycle times*. The fundamental assumption is that they are i.i.d. random variables. For real $t \geq 0$, let N_t denote the number of occurrences that have happened by time t. $\{N_t : t \in \mathbb{R}_{\geq 0}\}$ is called a renewal process. Here is a precise definition.

Definition 1.17. Let $X_1, X_2, X_3, \ldots$ be strictly positive i.i.d. random variables. Set $S_0 = 0$ and $S_n = \sum_{k=1}^{n} X_k$ for $n \geq 1$. Then the **renewal process** with waiting times $\{X_k : k \geq 1\}$ is the process N_t defined by

$$N_t = \sup\{n \geq 0 : S_n \leq t\} \quad \text{for real } t \geq 0. \tag{1.15}$$

Above, sup means "supremum" and in this case it is the same as a maximum. That X_k is strictly positive means that $P(X_k > 0) = 1$. This assumption is convenient because then two arrivals never come at exactly the same time. N_t is an example of a continuous-time process. We summarize some of its basic properties below. In the next lemma the abbreviation *a.s.* is for "almost surely", which is an equivalent way of saying "with probability one".

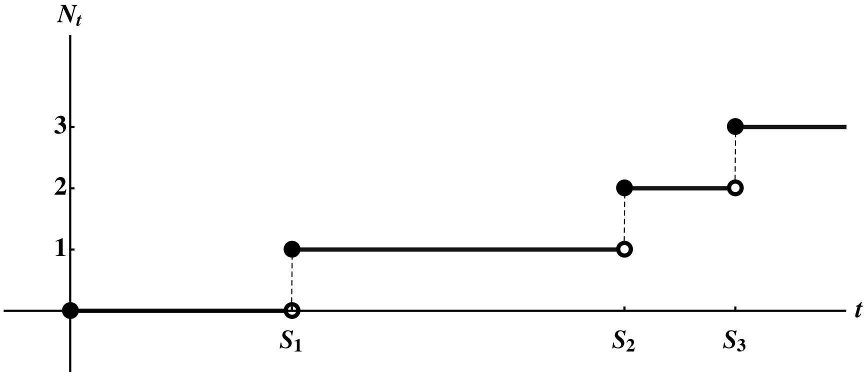

Figure 1. Graph of a renewal process N_t. The figure shows the first three occurrences that correspond to jumps of N_t.

Lemma 1.18. *Suppose that $\{N_t : t \geq 0\}$ is a renewal process with strictly positive i.i.d. waiting times $\{X_k : k \geq 1\}$. Then $N_0 = 0$, and with probability one, N_t is a finite non-negative integer for each $t \geq 0$. The random function $t \mapsto N_t$ is a.s. piecewise constant with jumps of size one at the points S_n for $n = 1, 2, 3, \ldots$. With probability one we have $\lim_{t\to\infty} N_t = \infty$.*

Proof. Since $S_0 = 0$ and $S_1 > 0$, the definition shows that $N_0 = 0$. The only way that N_t can fail to be an integer is that the maximum in (1.15) blows up to infinity. Thus it is enough to show that, with probability one, for any given t there are only finitely many n with $S_n \leq t$. This is true because $\lim_{n\to\infty} S_n = \infty$ almost surely, as follows from the version of the SLLN in Theorem 1.15. Note that we do not need to assume the finiteness of $E[X_1]$, for since $X_1 > 0$ we either have $E[X_1] = \infty$ or $0 < E[X_1] < \infty$, and hence (1.14) implies $\lim_{n\to\infty} S_n = \infty$.

Definition (1.15) implies that $N_t = 0$ for $t \in [0, S_1)$, $N_t = 1$ for $t \in [S_1, S_2)$ and in general $N_t = n$ for $t \in [S_n, S_{n+1})$ for $n \geq 0$. The assumption $P(X_k > 0) = 1$ implies that each interval $[S_n, S_{n+1})$ has strictly positive length. Consequently N_t is piecewise constant, with a jump of size one at each time point S_n for $n \geq 1$. This also shows that $N_t \to \infty$ as $t \to \infty$, since for any $n \geq 1$, $N_t \geq n$ for all $t \geq S_n$. □

In this section we address the long-term behavior of N_t by appeal to the SLLN. Further questions about renewal processes will be taken up in a later chapter.

Theorem 1.19 (Strong law of large numbers for renewal processes)**.** *Let $\{N_t : t \geq 0\}$ be a renewal process with strictly positive i.i.d. waiting times $\{X_k : k \geq 1\}$. Let $\mu = E[X_1]$.* (*We allow the possibility $\mu = \infty$.*) *Then with probability one,*

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu}. \tag{1.16}$$

If $\mu = \infty$ the limit is zero.

Before a rigorous proof, let us give a heuristic justification. If the cycle times were not random at all, but deterministically equal to their mean μ, then the number N_t of cycles completed by time t would be approximately t/μ. For large t, N_t/t would be very close to $1/\mu$. Then, as in the SLLN, we can imagine that random fluctuations away from the mean tend to average out over the long run. So it is sensible to expect that even with random cycle times, $N_t/t \approx 1/\mu$, with improving accuracy as t becomes large.

Proof of Theorem 1.19. The definition of N_t implies $S_{N_t} \leq t < S_{N_t+1}$. Take t large enough so that $N_t > 0$ and divide by it to get

$$\frac{S_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{S_{N_t+1}}{N_t} = \frac{N_t+1}{N_t} \cdot \frac{S_{N_t+1}}{N_t+1}. \tag{1.17}$$

By Lemma 1.18, $N_t \to \infty$ almost surely as $t \to \infty$. This implies that $\frac{N_t+1}{N_t} \to 1$ as $t \to \infty$. By Theorem 1.15, $S_n/n \to \mu$ as $n \to \infty$. Since $N_t \to \infty$, the same limit holds if we replace n with N_t or $N_t + 1$ and let $t \to \infty$. Thus with probability one

$$\frac{S_{N_t}}{N_t} \to \mu \quad \text{and} \quad \frac{S_{N_t+1}}{N_t} = \frac{N_t+1}{N_t} \cdot \frac{S_{N_t+1}}{N_t+1} \to 1 \cdot \mu = \mu \qquad \text{as } t \to \infty.$$

Thus both extremes of the inequality (1.17) converge almost surely to μ as $t \to \infty$. This forces $t/N_t \to \mu$. Taking reciprocals of this limit gives the conclusion $N_t/t \to 1/\mu$ in (1.16). □

Example 1.20. Suppose the lifetimes of light bulbs used in a storage room have i.i.d. exponential distribution with mean 100 hours.

(a) Suppose it takes on average one hour for a custodian to notice a burned-out bulb, at which point the bulb is immediately replaced with a new one. What is the long term rate at which bulbs are consumed?

(b) Suppose a custodian checks the bulb exactly on the hour, every hour, around the clock. A burned-out bulb is replaced immediately upon discovery. What is now the long term rate at which bulbs are consumed?

In both cases we take the cycle length as the time between the replacements of bulbs. Let $\{Z_k : k \geq 1\}$ be the lifetimes of the bulbs. The assumption is that $Z_k \sim \text{Exp}(\frac{1}{100})$ and $EZ_k = 100$.

In part (a) the length of the kth cycle is

$$X_k = Z_k + \text{ the time until the burned-out bulb is replaced.}$$

By the assumption, $\mu = EX_k = EZ_k + 1 = 101$. N_t is the number of bulbs replaced by time t. By Theorem 1.19, the long term rate is

$$\lim_{t\to\infty} \frac{N}{t} = \frac{1}{\mu} = \frac{1}{101}.$$

In part (b) the assumption is that a burned-out bulb is replaced at the next integer time. For example, if the bulb burns out at time 56.71, it is replaced at time 57. We can capture this with the *ceiling function*: the length of the kth cycle is $X_k = \lceil Z_k \rceil$, where the ceiling function is defined for real x by

$$\lceil x \rceil = \min\{n \in \mathbb{Z} : n \geq x\} = \text{ the smallest integer } \geq x.$$

We have to find the distribution of $X_k = \lceil Z_k \rceil$. For integers $m \geq 1$,

$$\begin{aligned} P(X_k = m) &= P(m-1 < Z_k \leq m) = P(Z_k > m-1) - P(Z_k > m) \\ &= e^{-\frac{m-1}{100}} - e^{-\frac{m}{100}} = \big(e^{-\frac{1}{100}}\big)^{m-1}\big(1 - e^{-\frac{1}{100}}\big). \end{aligned}$$

This indicates that X_k is a geometric random variable with success probability $p = 1 - e^{-\frac{1}{100}}$. The mean of a Geom($p$) random variable is $\mu = \frac{1}{p}$. Hence again by Theorem 1.19, the long term rate is

$$\lim_{t\to\infty} \frac{N}{t} = \frac{1}{\mu} = p = 1 - e^{-\frac{1}{100}}.$$

The answers to (a) and (b) are close to each other. Recall that $e^x \approx 1 + x$ for small $|x|$. Thus the answer to (b) is approximately $1 - (1 - \frac{1}{100}) = \frac{1}{100}$. △

Remark 1.21. The reader may wonder why we take a limit of infinite time to capture an average rate. Why not simply calculate the average number of events in some time interval? There are two reasons.

(i) Calculating $E[N_t]$ can be surprisingly tricky.

(ii) A more fundamental reason is that the mean rate of arrivals may vary depending on which particular time interval is observed.

To illustrate this, suppose a bus comes by at exactly every 40 minutes. Since 40 min = 2/3 hr, the average number of arrivals per hour is 3/2. But this is not necessarily what we see from a limited portion of the process. If we start observing immediately after the previous bus, then in the next hour we see one bus (suggesting a rate of 1/hr), in the next 1.5 hrs we see 2 buses (rate 1.33/hr) and in the next 2 hours we see 3 buses (rate 1.5/hr). If we come to observe 30 minutes after the previous bus, the next hour sees 2 buses (rate 2/hr).

Let N_t denote the number of buses up to time t where t is in units of hours. Then, regardless of when we started to observe the process,

$$\frac{60t}{40} - 1 \le N_t \le \frac{60t}{40} + 1.$$

In the $t \to \infty$ limit we get $t^{-1}N_t \to 3/2$, the correct average rate.

This elementary example contains an important idea: the behavior of a process in the beginning might not be representative of its long term behavior. For this reason *long term limits* are fundamental to any study of temporal evolutions, both stochastic and deterministic. △

In the next example we include an additional random outcome to each cycle. This model is a special case of a renewal-reward process. After the example we state the general definition and theorem.

Example 1.22. Buses arrive at the terminal with an average inter-arrival time of 10 minutes. Each bus carries on average 15 passengers. Let R_t denote the number of passengers that have arrived in t hours. What can we say about the long-term average $\lim_{t\to\infty} t^{-1}R_t$ number of passengers arriving per hour?

The answer should be intuitively clear: on average 6 buses arrive per hour, so on average there should be $6 \cdot 15 = 90$ passengers arriving per hour. To derive the answer rigorously we model this problem with a version of a renewal process.

Let N_t be the renewal process that represents the number of buses that have arrived in t hours. Here we are assuming that the inter-arrival times X_k between buses are i.i.d. and strictly positive. By assumption $E[X_1] = 1/6$ hours. Let Y_k denote the number of passengers on the kth bus. We have $E[Y_1] = 15$. In order to use the SLLN, we assume that $\{Y_k : k \ge 1\}$ are i.i.d.

The first task is to find an expression for R_t. The total number of passengers on the first n buses is $\sum_{k=1}^{n} Y_k$. N_t buses have arrived by time t, so the total number of passenger arrivals up to time t is given by $R_t = \sum_{k=1}^{N_t} Y_k$.

To find the limit of $t^{-1}R_t$ as $t \to \infty$, first rewrite this quantity as

$$\frac{1}{t}R_t = \frac{1}{t}\sum_{k=1}^{N_t} Y_k = \frac{N_t}{t} \cdot \frac{1}{N_t}\sum_{k=1}^{N_t} Y_k.$$

Above we consider large enough t so that $N_t \ge 1$ and we can divide by it.

Next record the almost sure limits we know. Theorem 1.19 gives $\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{E[X_1]} = 6/\text{hr}$. Since $\{Y_k : k \ge 1\}$ are i.i.d., the strong law of large numbers gives

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_k = E[Y_1] = 15.$$

Then the limit $N_t \to \infty$ as $t \to \infty$ implies that

$$\lim_{t\to\infty} \frac{1}{N_t}\sum_{k=1}^{N_t} Y_k = E[Y_1] = 15.$$

Putting all these together gives this limit:

$$\frac{1}{t}\sum_{k=1}^{N_t} Y_k = \frac{N_t}{t}\cdot\frac{1}{N_t}\sum_{k=1}^{N_t} Y_k \to \frac{1}{E[X_1]}\cdot E[Y_1] = \frac{6}{\text{hour}}\cdot 15 = \frac{90}{\text{hour}}$$

with probability one. △

Definition 1.23. Let $\{N_t : t \ge 0\}$ be a renewal process with strictly positive i.i.d. waiting times $\{X_k : k \ge 1\}$. Let $\{Y_k : k \ge 1\}$ be another sequence of i.i.d. random variables on the same probability space. Define

$$R_t = \sum_{k=1}^{N_t} Y_k \qquad \text{for real } t \ge 0.$$

$\{R_t : t \ge 0\}$ is the **renewal-reward process** corresponding to $\{(X_k, Y_k) : k \ge 1\}$. △

Example 1.22 derived a law of large numbers result for a specific example of a renewal-reward process. The theorem below contains the general case. The proof is essentially the same as the derivation in Example 1.22.

Theorem 1.24. *Suppose that R_t is a renewal-reward process corresponding to strictly positive i.i.d. waiting times $\{X_k : k \ge 1\}$ and i.i.d. rewards $\{Y_k : k \ge 1\}$. Assume that $E[X_1] = \mu \in (0, \infty]$ and $E[Y_1] = m \in \mathbb{R}$. Then with probability one*

$$\frac{R_t}{t} \to \frac{m}{\mu} \qquad \textit{as} \quad t \to \infty. \tag{1.18}$$

The next two examples illustrate how the notion of "reward" can be applied flexibly, and also how long term limits can ignore the contribution of the cycle that is currently under way at time t.

Example 1.25 (On-off process)**.** Suppose that a machine alternates between two states: ON and OFF. It spends a random time in the ON state, then a random time in the OFF state, then again in the ON state, and so on. Suppose that the average duration in the ON state is 5 hours, and in the OFF state 3 hours. What is the long term fraction of time that the machine is on?

Intuitively the answer must be $\frac{5}{5+3} = \frac{5}{8}$. We can use renewal processes to provide a rigorous justification of this answer under an i.i.d. assumption.

Denote by A_n the length of the nth ON period, and by B_n the length of the nth OFF period. By the setup of the problem $E[A_n] = 5$ and $E[B_n]$=3. To use our results, we assume that the pairs $\{(A_n, B_n) : n \ge 1\}$ form an i.i.d. sequence. We construct a renewal process N_t by using the events when the machine switches from OFF to ON. The cycle lengths are the random variables $\{A_n + B_n : n \ge 1\}$. These are now i.i.d. as a consequence of the i.i.d. assumption on $\{(A_n, B_n)\}$. Let M_t denote the time spent in the ON phase up to time t. Our goal is to prove the existence of the limit of $t^{-1}M_t$ as $t \to \infty$ and to find its value.

By considering the number of completed ON-OFF cycles up to time t we get

$$\sum_{k=1}^{N_t} A_k \le M_t \le \sum_{k=1}^{N_t+1} A_k. \tag{1.19}$$

We can consider A_n as the "reward" of the nth ON-OFF cycle. Theorem 1.24 gives the limit

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{N_t} A_k = \frac{E[A_1]}{E[A_1+B_1]} \quad \text{a.s.}$$

Similarly we derive an almost sure limit on the right-hand side of (1.19):

$$\begin{aligned}\frac{1}{t}\sum_{k=1}^{N_t+1} A_k &= \frac{N_t}{t}\cdot\frac{N_t+1}{N_t}\cdot\frac{1}{N_t+1}\sum_{k=1}^{N_t+1} A_k \\ &\underset{t\to\infty}{\longrightarrow} \frac{1}{E[A_1+B_1]}\cdot 1\cdot E[A_1] = \frac{E[A_1]}{E[A_1+B_1]}.\end{aligned}$$

It now follows from (1.19) that $\frac{M_t}{t}$ must converge to this same limit with probability one. We have shown that the long term fraction of time spent in the ON phase is indeed $\frac{E[A_1]}{E[A_1+B_1]} = \frac{5}{8}$. △

Example 1.26 (Continuation of Example 1.20)**.** In scenario (a) of Example 1.20, calculate the long term fraction of time that there is a functioning light bulb in the storage room.

Take the lifetime Z_k of the kth bulb as the reward of the kth cycle. Then $R_t = \sum_{k=1}^{N_t} Z_k$ is the total amount of time that already replaced bulbs functioned in the storage room. By Theorem 1.24,

$$\lim_{t\to\infty} t^{-1}R_t = \frac{EZ_1}{EX_1} = \frac{100}{101}.$$

The calculation above ignores the contribution of the current light bulb. This contribution is at most the elapsed length of the current cycle (the cycle that is under way at time t) which is $t-\sum_{k=1}^{N_t} X_k$. This contribution vanishes in the limit when we divide by t:

$$\lim_{t\to\infty}\frac{1}{t}\Bigl(t-\sum_{k=1}^{N_t} X_k\Bigr) = 1-\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{N_t} X_k = 1-\frac{EX_1}{EX_1} = 0.$$

To summarize, if we let U_t denote the amount of time up to time t that the storage room has a functioning light bulb, then we have the bounds

$$\frac{1}{t}R_t \;\le\; \frac{1}{t}U_t \;\le\; \frac{1}{t}R_t + \frac{1}{t}\Bigl(t-\sum_{k=1}^{N_t} X_k\Bigr).$$

From the limits above, $t^{-1}U_t \to \frac{100}{101}$ as $t\to\infty$, with probability one. △

In Example 1.22 the reward came in discrete amounts exactly at the renewal times. By contrast, in Examples 1.25 and 1.26 the reward accrued partly continuously in time. Still, the limit is the same m/μ given in Theorem 1.24.

1.4. Technical appendix

As stated in Theorem 1.13, the SLLN is true under the assumption of a finite mean. If we strengthen the assumption to a finite fourth moment $E[X_1^4]$, we can give a fairly quick proof. It is a fact that if $0<a<b$, then $E[\,|X|^b\,]<\infty$ implies

$E[\,|X|^a\,] < \infty$. Thus a finite fourth moment implies a finite expectation, and our assumption of a finite fourth moment is stronger than the assumption of a finite mean. But we do not lose much by assuming this stronger condition because most of the random variables that arise in practice satisfy this assumption. Proofs under the finite expectation assumption can be found in graduate probability textbooks, for example in Durrett [Dur19].

Proof of Theorem 1.13 under a finite fourth moment. Suppose first that $\mu = 0$. The general case of $\mu \neq 0$ follows easily from the $\mu = 0$ case at the end of the proof.

We show that the random variable $\sum_{n=1}^{\infty}(S_n/n)^4$ has a finite expectation. This implies the conclusion due to the following:

(i) If $E\big[\sum_{n=1}^{\infty}(S_n/n)^4\big]$ is finite, then the nonnegative random variable $\sum_{n=1}^{\infty}(S_n/n)^4$ must be finite with probability one. For otherwise the expectation would be infinite, as proved in Lemma B.3 in Appendix B.2.

(ii) In general, if a series $\sum_{n=1}^{\infty} a_n$ converges, then $a_n \to 0$. Thus, by point (i) we have $(S_n/n)^4 \to 0$ with probability one, which implies that $S_n/n \to 0$ with probability one. This is our desired result.

We turn to bounding $E[S_n^4]$. Expanding $S_n^4 = (X_1 + \cdots + X_n)^4$ gives terms of the form

$$X_i^4,\ \ X_i^2X_j^2,\ \ X_i^3X_j,\ \ X_i^2X_jX_k,\ \text{ and }\ X_iX_kX_jX_\ell,\ \text{ for distinct } i,j,k,\ell \in \{1,\dots,n\}.$$

Because the random variables $X_1, \dots, X_n$ are i.i.d. and have mean zero, only the terms X_i^4 and $X_i^2X_j^2$ have nonzero expectation. Denote these by $E[X_i^4] = E[X_1^4] = c_1$ and $E[X_i^2X_j^2] = (E[X_1^2])^2 = c_2$. Then after collecting terms

$$\begin{aligned} E[S_n^4] &= E[(X_1 + \cdots + X_n)^4] = nE[X_1^4] + 6\binom{n}{2}(E[X_1^2])^2 \\ &= nc_1 + 3n(n-1)c_2 \le nc_1 + 3n^2c_2. \end{aligned}$$

From the estimate above we get the desired result:

$$E\left[\sum_{n=1}^{\infty}\frac{S_n^4}{n^4}\right] = \sum_{i=1}^{\infty} E\left[\frac{S_n^4}{n^4}\right] \le \sum_{n=1}^{\infty}\frac{c_1}{n^3} + \sum_{n=1}^{\infty}\frac{3c_2}{n^2} < \infty.$$

Switching around the expectation and the infinite series above is justified by the monotone convergence theorem (Theorem B.5 in Appendix B.2).

We have now proved the SLLN for the case $\mu = 0$. If the mean is nonzero, we apply the result already proved to the mean zero random variables $X_i - \mu$. This yields the almost sure limit

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{(X_i-\mu)}{n} = 0 \quad \text{or equivalently} \quad \lim_{n\to\infty}\sum_{i=1}^{n}\frac{X_i}{n} = \mu.$$

□

The SLLN gives convergence $n^{-1}S_n(\omega) \to \mu$ for sample points ω on an event of probability one, but not necessarily for all sample points ω. We illustrate the necessity of this restriction through the example of fair coin flips.

Example 1.27. Let $\{X_k : k \geq 1\}$ be i.i.d. random variables with probability mass function $P(X_k = 0) = P(X_k = 1) = \frac{1}{2}$ and $S_n = X_1 + \cdots + X_n$. The SLLN gives the limit

$$n^{-1} S_n(\omega) \to \tfrac{1}{2} \quad \text{with probability one.} \tag{1.20}$$

The natural sample space for the coin flips is the space

$$\Omega = \{\omega = (\omega_k)_{k \geq 1} : \omega_k \in \{0,1\} \text{ for all } k \geq 1\}$$

of $\{0,1\}$-valued sequences. There are many sequences ω that do not satisfy (1.20). For example, if we substitute the constant zero sequence $\omega = (0, 0, 0, \ldots)$ into (1.20), we get $n^{-1} S_n(\omega) = n^{-1} n \cdot 0 = 0$ which certainly does not converge to $\frac{1}{2}$. △

Thus another way to express the message of the SLLN is that under the probability measure that corresponds to i.i.d. random variables, those sample points that satisfy the limit make up an event of probability one.

Proof of Theorem 1.15, assuming the basic SLLN Theorem 1.13. If the expectation $E(X_1)$ is finite, then the claim is already contained in Theorem 1.13. So suppose $E(X_1) = \infty$. We use a *truncation*, a common technique in probability. For a real number $c > 0$, let $X_k^{(c)} = X_k \wedge c$. The notation here is that $a \wedge b = \min\{a, b\}$ is the smaller one of the two numbers a and b.

The truncated random variables are still i.i.d. (because they are produced by applying a function to the original i.i.d. variables) and their expectation satisfies $0 \leq E[X_k^{(c)}] \leq c$. Thus the original SLLN applies to the random variables $\{X_k^{(c)}\}_{k \geq 1}$. Given any real $b > 0$, since $E(X_1) = \infty$, it is possible to choose a large enough c such that $E[X_1^{(c)}] > b$. (This is again justified by the monotone convergence theorem.) Now we can reason as follows: since $X_k \geq X_k^{(c)}$ is true for all k, we have

$$\frac{1}{n} \sum_{k=1}^{n} X_k \geq \frac{1}{n} \sum_{k=1}^{n} X_k^{(c)}.$$

By Theorem 1.13, the last expression above converges to a number larger than b. Since we can take b as large as we please, the first average must converge to infinity. □

Chapter 2

Discrete-time Markov chains

2.1. Markov property and transition probabilities

The most obvious generalization away from a sequence of independent random variables is to allow dependence one step into the past. To lead us to this Markovian dependence we first recall the general multiplication formula.

Finite-dimensional distributions and the general multiplication formula. To see how to naturally generalize from a process of independent random variables, let us recall the general *multiplication formula* for probabilities: for any events $A_0, A_1, \dots, A_n$ on $(\Omega, \mathcal{F}, P)$, provided that $P(A_0 \cap \cdots \cap A_{n-1}) > 0$,

$$\begin{aligned} &P(A_0 \cap A_1 \cap \cdots \cap A_n) \\ &\qquad = P(A_0)\, P(A_1 \,|\, A_0)\, P(A_2 \,|\, A_1 \cap A_0) \cdots P(A_n \,|\, A_{n-1} \cap \cdots \cap A_0). \end{aligned} \tag{2.1}$$

This is proved by first multiplying and dividing by the probabilities of all the conditioning events and by then applying the definition of conditional probability:

$$\begin{aligned} &P(A_0 \cap A_1 \cap \cdots \cap A_n) \\ &\quad = P(A_0) \cdot \frac{P(A_0 \cap A_1)}{P(A_0)} \cdot \frac{P(A_0 \cap A_1 \cap A_2)}{P(A_0 \cap A_1)} \cdots \frac{P(A_0 \cap \cdots \cap A_{n-1} \cap A_n)}{P(A_0 \cap \cdots \cap A_{n-1})} \\ &\quad = P(A_0)\, P(A_1 \,|\, A_0)\, P(A_2 \,|\, A_1 \cap A_0) \cdots P(A_n \,|\, A_{n-1} \cap \cdots \cap A_0). \end{aligned} \tag{2.2}$$

Apply (2.1) to the joint distribution of a completely arbitrary (that is, without any further assumptions made) discrete-time stochastic process $\{X_k\}_{k \ge 0}$ with discrete state space $\mathcal{S}$: for any $n \in \mathbb{Z}_{\ge 0}$ and any points $x_0, x_1, \dots, x_n \in \mathcal{S}$ such that $P(X_0 = x_0, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) > 0$,

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &\qquad = P(X_0 = x_0) \cdot \prod_{k=0}^{n-1} P(X_{k+1} = x_{k+1} \,|\, X_k = x_k, \dots, X_0 = x_0). \end{aligned} \tag{2.3}$$

Independence of the random variables $\{X_k\}$ means exactly that conditioning on the values of $X_0, \dots, X_k$ does not change the probabilities of X_{k+1}. Thus in the independent case equation (2.3) simplifies to

$$P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \prod_{k=0}^{n} P(X_k = x_k). \tag{2.4}$$

Markovian dependence. As we begin to generalize from independent random variables to more complicated joint distributions, a natural next step of complexity is to assume that X_{k+1} depends on X_k but not on the earlier past. The mathematical statement of this assumption is that

$$\begin{aligned} &P(X_{k+1} = x_{k+1} \,|\, X_k = x_k, X_{k-1} = x_{k-1}, \dots, X_0 = x_0) \\ &\qquad = P(X_{k+1} = x_{k+1} \,|\, X_k = x_k). \end{aligned} \tag{2.5}$$

Note that if $\{X_n : n \geq 0\}$ is an sequence of independent random variables then (2.5) is satisfied, as both sides are equal to $P(X_{k+1} = x_{k+1})$. Hence this is indeed a generalization of independent sequences.

If we assume (2.5) then equation (2.3) simplifies to

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &\qquad = P(X_0 = x_0) \cdot \prod_{k=0}^{n-1} P(X_{k+1} = x_{k+1} \,|\, X_k = x_k). \end{aligned} \tag{2.6}$$

This new assumption turns out to be an incredibly rich source of deep and practically useful mathematics. We have in fact just stated a version of the definition of a Markov chain. We repeat the definition precisely again below, after first showing through two examples that this type of process arises very naturally.

Example 2.1 (Simple random walk on the integers)**.** Fix $0 < p < 1$. Let $\{Y_k\}_{k \in \mathbb{Z}_{>0}}$ be an i.i.d. process with common probability mass function

$$p_Y(1) = P(Y_k = 1) = p \quad \text{and} \quad p_Y(-1) = P(Y_k = -1) = 1 - p.$$

Define the *simple random walk* as the following process:

$$S_0 = 0, \text{ and for } n \geq 1, \ S_n = Y_1 + \cdots + Y_n. \tag{2.7}$$

We check that $\{S_n\}_{n \geq 0}$ has the following property: for any integers $x_0, x_1, \dots, x_{n+1}$ such that $x_0 = 0$ and $P(S_0 = x_0, S_1 = x_1, \dots, S_n = x_n) > 0$, we have the equality

$$\begin{aligned} &P(S_{n+1} = x_{n+1} \,|\, S_0 = x_0, \, S_1 = x_1, \dots, S_n = x_n) \\ &\qquad\qquad = P(S_{n+1} = x_{n+1} \,|\, S_n = x_n). \end{aligned} \tag{2.8}$$

In other words, if the state $S_n = x_n$ at time n is given, conditioning on the entire earlier past does not influence the next state at time $n+1$. The verification of (2.8) uses the independence of the Y_ks.

First evaluate the left-hand side of (2.8):

$$
\begin{aligned}
&P(S_{n+1} = x_{n+1} \,|\, S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n)\\
&\qquad = \frac{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n, S_{n+1} = x_{n+1})}{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n)}\\
&\qquad = \frac{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n, S_{n+1} - S_n = x_{n+1} - x_n)}{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n)}
\end{aligned}
$$

Note that $S_0, \dots S_n$ are all computable from $Y_1, \dots, Y_n$, and $S_{n+1} - S_n = Y_{n+1}$ is independent of $Y_1, \dots, Y_n$. Hence

$$
\begin{aligned}
&P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n, S_{n+1} - S_n = x_{n+1} - x_n)\\
&\qquad = P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n, Y_{n+1} = x_{n+1} - x_n)\\
&\qquad = P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n) P(Y_{n+1} = x_{n+1} - x_n)
\end{aligned}
$$

and cancellation gives

$$
\begin{aligned}
&\frac{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n, S_{n+1} - S_n = x_{n+1} - x_n)}{P(S_0 = x_0,\, S_1 = x_1, \dots, S_n = x_n)}\\
&\qquad = P(Y_{n+1} = x_{n+1} - x_n).
\end{aligned}
$$

For the right side of (2.8) we have, similarly,

$$
\begin{aligned}
&P(S_{n+1} = x_{n+1} \,|\, S_n = x_n) = \frac{P(S_n = x_n,\, S_{n+1} = x_{n+1})}{P(S_n = x_n)}\\
&\quad = \frac{P(S_n = x_n, Y_{n+1} = x_{n+1} - x_n)}{P(S_n = x_n)} = \frac{P(S_n = x_n)\, P(Y_{n+1} = x_{n+1} - x_n)}{P(S_n = x_n)}\\
&\quad = P(Y_{n+1} = x_{n+1} - x_n).
\end{aligned}
$$

Above we used the independence of S_n and Y_{n+1}. Thus we have verified (2.8).

Additionally, we found a formula for the *transition probability* of the random walk:

$$
\begin{aligned}
&P(S_{n+1} = x + 1 \,|\, S_n = x) = p\,,\\
&P(S_{n+1} = x - 1 \,|\, S_n = x) = 1 - p\,, \qquad (2.9)\\
\text{and} \qquad &P(S_{n+1} = y \,|\, S_n = x) = 0 \quad \text{if } y \notin \{x \pm 1\}.
\end{aligned}
$$

Simple random walk is the most fundamental discrete-time stochastic process. When $p = \frac{1}{2}$, S_n is a *symmetric* simple random walk (abbreviated SSRW), and if $p \neq \frac{1}{2}$, S_n is an *asymmetric* simple random walk.

SSRW can be described in words as follows. Start at the origin. At each integer time, flip a fair coin to determine whether you take a step right or a step left. For obvious reasons, SSRW is facetiously also called the *drunkard's walk.* △

Example 2.2 (The last two coin flips)**.** Flip a fair coin repeatedly. Record the outcome of the nth flip as $Y_n = 0$ for heads and $Y_n = 1$ for tails, for $n = 1, 2, 3, \dots$. For $n \geq 2$ let $X_n = (Y_{n-1}, Y_n)$ be the ordered pair of the $(n-1)$st and nth outcomes. For example, if the first four coin flips are (heads, tails, tails, heads) then $X_2 = (0, 1)$, $X_3 = (1, 1)$, $X_4 = (1, 0)$. $\{X_n\}_{n \geq 2}$ is a discrete-time stochastic process with finite state space $\mathcal{S} = \{0, 1\} \times \{0, 1\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

Although $\{Y_k\}_{k\geq 1}$ are i.i.d., the random variables $X_2, X_3, \dots$ are not independent. This intuitively obvious fact can be seen rigorously for example as follows: for any $a, b \in \{0,1\}$ and $n \geq 2$,

$$P(X_n = (a,b)) = P(Y_{n-1} = a, Y_n = b) = P(Y_{n-1} = a)P(Y_n = b) = \tfrac{1}{4},$$

but $P(X_n = (a,b), X_{n+1} = (c,d)) = 0$ unless $b = c$. Thus independence is lost, but the process $\{X_n\}_{n\geq 2}$ satisfies the Markov property (2.6).

To verify the Markov property we need to check that for any $a_k, b_k \in \{0,1\}$, $k = 2, \dots, n+1$, for which

$$P\big(X_2 = (a_2,b_2), X_3 = (a_3,b_3), \dots, X_n = (a_n,b_n)\big) > 0, \tag{2.10}$$

we have the identity

$$\begin{aligned} &P\big(X_{n+1} = (a_{n+1},b_{n+1}) \,\big|\, X_2 = (a_2,b_2), \dots, X_n = (a_n,b_n)\big) \\ &\qquad = P\big(X_{n+1} = (a_{n+1},b_{n+1}) \,\big|\, X_n = (a_n,b_n)\big). \end{aligned} \tag{2.11}$$

Note that condition (2.10) holds exactly when $b_2 = a_3, b_3 = a_4, \dots, b_{n-1} = a_n$, and then

$$\begin{aligned} &P\big(X_2 = (a_2,b_2), X_3 = (a_3,b_3), \dots, X_n = (a_n,b_n)\big) \\ &\qquad = P\big(Y_1 = a_1, Y_2 = a_2, \dots, Y_{n-1} = a_n, Y_n = b_n\big) = 2^{-n}. \end{aligned} \tag{2.12}$$

Checking (2.11) involves two cases.

Case 1. $a_{n+1} = b_n$. In this case, on the left-hand side of (2.11), (2.12) applies to both numerator (with n replaced by $n+1$) and denominator to give

$$\begin{aligned} &P(X_{n+1} = (a_{n+1},b_{n+1}) \,|\, X_1 = (a_1,b_1), X_2 = (a_2,b_2), \dots, X_n = (a_n,b_n)) \\ &\qquad = \frac{2^{-(n+1)}}{2^{-n}} = \tfrac{1}{2}. \end{aligned} \tag{2.13}$$

On the right-hand side of (2.11) we have

$$\begin{aligned} P\big(X_{n+1} = (a_{n+1},b_{n+1}) \,\big|\, X_n = (a_n,b_n)\big) &= \frac{P\big(X_n = (a_n,b_n), X_{n+1} = (a_{n+1},b_{n+1})\big)}{P(X_n = (a_n,b_n))} \\ &= \frac{P(Y_{n-1} = a_n, Y_n = b_n, Y_n = a_{n+1}, Y_{n+1} = b_{n+1})}{P(Y_{n-1} = a_n, Y_n = b_n)} = \frac{2^{-3}}{2^{-2}} = \tfrac{1}{2}. \end{aligned}$$

We have verified (2.11) in the case $a_{n+1} = b_n$.

Case 2. $a_{n+1} \neq b_n$. In this case both sides of (2.11) equal zero because the event $X_n = (a_n,b_n), X_{n+1} = (a_{n+1},b_{n+1})$ cannot happen.

To summarize, the Markov property (2.11) has been verified for this process $\{X_n\}_{n\geq 2}$. △

Remark 2.3 (Irrelevance of start time)**.** In Example 2.2 the process X_n was indexed from time $n = 2$ onwards. The reader should understand that the choice of the initial time index value is usually *not* an essential feature of any stochastic process. We use time zero as a generic initial time simply for convenience. In any case, the time index can be shifted. In Example 2.2 we can define a new process $\widetilde{X}_n = X_{n+2}$ if we insist on starting at $n = 0$. △

It is time for precise definitions. A *countably infinite* set is one whose elements can be arranged in a sequence. Finite and countably infinite sets together are called *at most countable*, *countable*, and *discrete*.

Definition 2.4. A stochastic process $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ with a countable state space $\mathcal{S}$ is a **Markov chain** if for any time index $n \in \mathbb{Z}_{\geq 0}$ and any states $x_0, x_1, \dots, x_{n+1} \in \mathcal{S}$ such that $P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) > 0$, we have the equality

$$\begin{aligned} P(X_{n+1} = x_{n+1} \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ = P(X_{n+1} = x_{n+1} \,|\, X_n = x_n). \end{aligned} \tag{2.14}$$

Equation (2.14) is called the **Markov property**.

If the conditional probability $P(X_{n+1} = b \,|\, X_n = a)$ is the *same* for all n, then we call the Markov chain **time-homogeneous**. In that case the conditional probability $P(X_{n+1} = b \,|\, X_n = a)$ is called the **transition probability (function)** of the Markov chain and regarded as a function of the pair (a, b). △

The transition probability summarizes all the information about the evolution of a Markov chain. It turns out convenient to describe all examples in terms of their transition probability. For this purpose we formalize the concept of a transition probability as a function p on the product space $\mathcal{S} \times \mathcal{S}$ in the next definition.

Definition 2.5. Let $\mathcal{S}$ be a countable set. A **transition probability** p on $\mathcal{S}$ is a function $p : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ such that

$$0 \leq p(x,y) \leq 1 \quad \forall x, y \in \mathcal{S}, \tag{2.15}$$

$$\text{and} \quad \sum_{y\in\mathcal{S}} p(x,y) = 1 \quad \forall x \in \mathcal{S}. \tag{2.16}$$

△

The transition probability function of a time-homogeneous Markov chain satisfies $p(x,y) = P(X_{n+1} = y \,|\, X_n = x)$ when $P(X_n = x) > 0$. In other words, $p(x,y)$ is the conditional probability that the Markov chain jumps next to state y, given that it is currently in state x. Properties (2.15)–(2.16) are then natural: conditional probabilities are numbers in $[0,1]$ and they satisfy

$$1 = \sum_{y\in\mathcal{S}} P(X_{n+1} = y \,|\, X_n = x) = \sum_{y\in\mathcal{S}} p(x,y)$$

as long as $P(X_n = x) > 0$.

It is natural to represent the transition probability as a matrix when the state space $\mathcal{S}$ is finite. For example, if $\mathcal{S} = \{0,1,2\}$, the general form of the *transition probability matrix* looks like this:

$$\mathbf{P} = \begin{bmatrix} p(0,0) & p(0,1) & p(0,2) \\ p(1,0) & p(1,1) & p(1,2) \\ p(2,0) & p(2,1) & p(2,2) \end{bmatrix}. \tag{2.17}$$

The rows and the columns of the matrix $\mathbf{P}$ are indexed by the states of the Markov chain. The matrix language and notation are so convenient that we adopt them even when the state space is infinite. Then the possibly infinite transition probability matrix is denoted by $\mathbf{P} = \{p(x,y)\}_{x,y\in\mathcal{S}}$. This convention will acquire further significance when we see that parts of matrix algebra work for infinite transition probability matrices and these operations have important meaning for the Markov chain.

Let us look at the transition probabilities of the two examples that we discussed.

Example 2.6. As stated in (2.9) in Example 2.1, simple random walk $\{S_n\}$ is a Markov chain with state space $\mathbb{Z}$ and transition probability

$$p(x,y) = \begin{cases} p, & y = x+1 \\ 1-p, & y = x-1 \\ 0, & y \notin \{x \pm 1\}. \end{cases}$$

In the particular case in Example 2.1 the initial state was $S_0 = 0$. We could have chosen to start the random walk from any integer. △

Example 2.7. In Example 2.2 of the last two coin flips the state space is $\mathcal{S} = \{(0,0),(0,1),(1,0),(1,1)\}$. We derived the transition probabilities for $(a,b),(c,d) \in \mathcal{S}$ as

$$p((a,b),(c,d)) = P(X_{n+1} = (c,d) \,|\, X_n = (a,b)) = \begin{cases} \frac{1}{2}, & \text{if } b = c \\ 0, & \text{otherwise.} \end{cases}$$

The matrix form is

$$\mathbf{P} = \begin{array}{c} \\ \scriptstyle(0,0) \\ \scriptstyle(0,1) \\ \scriptstyle(1,0) \\ \scriptstyle(1,1) \end{array} \begin{array}{c} \begin{array}{cccc} \scriptstyle(0,0) & \scriptstyle(0,1) & \scriptstyle(1,0) & \scriptstyle(1,1) \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array} \tag{2.18}$$

△

Next we present the final definition of a time-homogeneous Markov chain that combines the Markov property, the transition probability, and a specification of how the process starts at time zero. To allow a random initial state, we use a probability distribution on $\mathcal{S}$.

A *probability measure* or *probability distribution* μ on a countable space $\mathcal{S}$ can be thought of as a function $\mu : \mathcal{S} \to \mathbb{R}$ with the properties $0 \le \mu(x) \le 1$ for all $x \in \mathcal{S}$ and $\sum_{x\in\mathcal{S}} \mu(x) = 1$. This terminology is somewhat inaccurate because properly speaking a probability measure is a function on *subsets* of the state space. In the case of a countable space $\mathcal{S}$, the probability of a subset $B \subset \mathcal{S}$ is then $\mu(B) = \sum_{x\in B} \mu(x)$.

Definition 2.8. Let $\mathcal{S}$ be a countable state space, p a transition probability on $\mathcal{S}$, and μ a probability measure on $\mathcal{S}$. Let $\{X_k\}_{k\in\mathbb{Z}_{\ge 0}}$ be a stochastic process on $(\Omega, \mathcal{F}, P)$ with state space $\mathcal{S}$. Then $\{X_k\}$ is a **Markov chain with transition probability** p **and initial distribution** μ if

$$P(X_0 = x) = \mu(x) \quad \forall\, x \in \mathcal{S} \tag{2.19}$$

and

$$P(X_{n+1} = y \,|\, X_n = x, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = p(x,y) \tag{2.20}$$

for all states $x_0, \dots, x_{n-1}, x, y$ whenever

$$P(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = x) > 0.$$

If μ is degenerate, that is, supported on a single state z: $\mu(z) = 1$ and $\mu(x) = 0$ for $x \neq z$, then $\{X_k\}$ is a **Markov chain with transition probability** p **and initial state** z. △

Definition 2.8 was written so that it is easy to check. But as it stands it does not imply the Markov property (2.14) unless we show that (2.20) implies also that $P(X_{n+1} = y \,|\, X_n = x) = p(x,y)$ whenever $P(X_n = x) > 0$. This is the special case $m = n$ of the next lemma.

Lemma 2.9. *Suppose $\{X_k\}$ is a Markov chain with transition probability p on state space $\mathcal{S}$. Then for all integers $n \ge m \ge 0$ and states $x_m, \dots, x_n, y$ such that*

$$P(X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n) > 0$$

we have

$$P(X_{n+1} = y \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_m = x_m) = p(x_n, y). \tag{2.21}$$

Proof. For ease of notation denote y by x_{n+1}. The assumption made above ensures that the denominators below are nonzero.

$$\begin{aligned}
&P(X_{n+1} = x_{n+1} \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_m = x_m) \\
&= \frac{P(X_m = x_m, \dots, X_{n+1} = x_{n+1})}{P(X_m = x_m, \dots, X_n = x_n)} \\
&= \frac{\sum_{z_0,\dots,z_{m-1}} P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_{n+1} = x_{n+1})}{\sum_{z_0,\dots,z_{m-1}} P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_n = x_n)} \\
&\overset{(2.20)}{=} \frac{\sum_{z_0,\dots,z_{m-1}} P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_n = x_n)\, p(x_n, x_{n+1})}{\sum_{z_0,\dots,z_{m-1}} P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_n = x_n)} \\
&= p(x_n, x_{n+1}).
\end{aligned}$$

Assumption (2.20) was applied to the numerator in the form

$$p(x_n, x_{n+1}) = \frac{P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_{n+1} = x_{n+1})}{P(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, X_m = x_m, \dots, X_n = x_n)}.$$

The last step is cancellation. □

In Examples 2.1 and 2.2 we basically checked property (2.20) for all states $x_0, \dots, x_n, y \in \mathcal{S}$. But this can be cumbersome. Here is a short-cut that works for some models.

Lemma 2.10. *Suppose that* $\{X_n : n \geq 0\}$ *is a stochastic process with state space* $\mathcal{S}$*. Let* $\{Y_n : n \geq 1\}$ *be a sequence of i.i.d. random variables with values from a set* $\mathcal{A}$*. Assume that* $\{Y_n : n \geq 1\}$ *are independent of* X_0*, and that there is a function* $H : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ *such that*

$$X_{n+1} = H(X_n, Y_{n+1}) \qquad \textit{for all } n \geq 0. \tag{2.22}$$

Then $\{X_n : n \geq 0\}$ *is a Markov chain with transition probability*

$$p(x, y) = P(H(x, Y_1) = y).$$

This lemma says that if X_{n+1} can be computed from X_n and a new independent random input Y_{n+1} (with a fixed distribution) through a fixed function H, then $\{X_n : n \geq 0\}$ is a Markov chain. Exercise 2.14 shows that in a certain sense any Markov chain on a countable state space $\mathcal{S}$ can be represented this way.

Proof. We check (2.20). Suppose that $P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0$, then

$$\begin{aligned}
&P(X_{n+1} = y \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
&\quad = \frac{P(X_{n+1} = y, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)}{P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)} \\
&\quad = \frac{P(H(x_n, Y_{n+1}) = y, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)}{P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)}.
\end{aligned}$$

By (2.22) the random variables $X_0, X_1, \ldots, X_n$ can be computed from $X_0, Y_1, \ldots, Y_n$. Since Y_{n+1} is independent of these random variables, we have that $H(x_n, Y_{n+1})$ is independent of $X_0, \ldots, X_n$. This means that

$$\begin{aligned}
&P(H(x_n, Y_n) = y, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
&\quad = P(H(x_n, Y_{n+1}) = y) P(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0).
\end{aligned}$$

But this implies

$$\begin{aligned}
&P(X_{n+1} = y \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
&\quad = P(H(x_n, Y_{n+1}) = y) = P(H(x_n, Y_1) = y) = p(x_n, y),
\end{aligned}$$

which is exactly what we needed. □

Example 2.11. Our two first examples both fall into this pattern.

The random walk S_n in Example 2.1 is constructed from an i.i.d. sequence $\{Y_k : k \geq 1\}$ with $S_{n+1} = S_n + Y_{n+1}$ and $S_0 = 0$. Lemma 2.10 implies that S_n is a Markov chain.

The process in Example 2.2 satisfies $X_n = (Y_{n-1}, Y_n)$ and $X_{n+1} = (Y_n, Y_{n+1})$ where $\{Y_n : n \geq 1\}$ is an i.i.d. sequence. Thus $X_{n+1} = H(X_n, Y_{n+1})$ for the function $H((a, b), c) = (b, c)$. Lemma 2.10 implies that X_n is a Markov chain. △

The next theorem observes that the properties in Definition 2.8 are enough to completely determine all the finite-dimensional distributions of the stochastic process $\{X_n\}$, and thereby the probabilities of all events concerning this process. There are no denominators in equation (2.23) below, and hence there is no need to assume a priori that any probability is nonzero.

Theorem 2.12. *Suppose $\{X_k\}$ is a Markov chain with transition probability p on state space $\mathcal{S}$. Then for all integer time points $n > m \geq 0$ and all states $x_m, \dots, x_n \in \mathcal{S}$,*

$$\begin{aligned} &P(X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n) \\ &\qquad = P(X_m = x_m) \prod_{k=m}^{n-1} p(x_k, x_{k+1}). \end{aligned} \tag{2.23}$$

Proof. *Case 1.* First we take care of the case where

$$P(X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_{n-1} = x_{n-1}) = 0. \tag{2.24}$$

This makes the left-hand side of (2.23) zero. We show that the right-hand side of (2.23) is also zero. Let $\ell \in \{m, m+1, \dots, n\}$ be the first integer such that $P(X_m = x_m, \dots, X_\ell = x_\ell) = 0$. The argument splits into two cases.

(i) If $\ell = m$ then $P(X_m = x_m) = 0$ makes the right-hand side of (2.23) vanish.

(ii) If $\ell > m$ then $\{X_m = x_m, \dots, X_{\ell-1} = x_{\ell-1}\}$ is a positive probability event. Equation (2.21) gives

$$\begin{aligned} p(x_{\ell-1}, x_\ell) &= P(X_\ell = x_\ell \,|\, X_m = x_m, \dots, X_{\ell-1} = x_{\ell-1}) \\ &= \frac{P(X_m = x_m, \dots, X_{\ell-1} = x_{\ell-1}, X_\ell = x_\ell)}{P(X_m = x_m, \dots, X_{\ell-1} = x_{\ell-1})} \\ &= 0. \end{aligned}$$

In the last quotient the denominator is strictly positive while the numerator vanishes. Since $p(x_{\ell-1}, x_\ell)$ is one of the factors on the right-hand side of (2.23), this side is also zero. Thus identity (2.23) holds under assumption (2.24) in the sense that both sides vanish.

Case 2. Now assume

$$P(X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_{n-1} = x_{n-1}) > 0.$$

This makes the conditionings below legitimate. Apply the general multiplication rule (2.3) and then (2.21).

$$\begin{aligned} &P(X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n) \\ &= P(X_m = x_m) \cdot \prod_{k=m}^{n-1} P(X_{k+1} = x_{k+1} \,|\, X_m = x_m, \dots, X_k = x_k) \\ &\overset{(2.21)}{=} P(X_m = x_m) \prod_{k=m}^{n-1} p(x_k, x_{k+1}). \end{aligned}$$

Identity (2.23) has now been proved in all cases. □

Note that if (2.23) holds then (2.20) follows immediately: assuming that the denominator does not vanish,

$$
\begin{aligned}
&P(X_{n+1}=y \,|\, X_n=x_n, X_{n-1}=x_{n-1},\dots,X_0=x_0)\\
&\quad=\frac{P(X_0=x_0,X_1=x_1,\dots,X_n=x_n,X_{n+1}=y)}{P(X_0=x_0,X_1=x_1,\dots,X_n=x_n)}\\
&\quad=\frac{P(X_0=x_0)\cdot\bigl(\prod_{k=0}^{n-1}p(x_k,x_{k+1})\bigr)\cdot p(x_n,y)}{P(X_0=x_0)\prod_{k=0}^{n-1}p(x_k,x_{k+1})}\\
&\quad=p(x_n,y).
\end{aligned}
\tag{2.25}
$$

Thereby the process is a Markov chain with transition probability p. Thus *validity of equation* (2.23) *for finite-dimensional distributions is an alternative, equivalent characterization of being a Markov chain with transition probability* p.

Definition 2.4 tells us how to decide if a *given* stochastic process is a Markov chain. We used this definition in Examples 2.1 and 2.2 to determine that the introduced stochastic processes are Markov chains.

However, most Markov chains do not arise as a given stochastic process. Instead, we want a Markov chain to model something, and the application determines what the transition probabilities should be. Thus we need to be able to *construct* Markov chains from given initial distributions and transition probabilities.

What does it mean to construct a Markov chain from a given transition probability and initial distribution? To have a Markov chain means to have a probability space $(\Omega,\mathcal{F},P)$ and on that probability space a stochastic process $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ that satisfies Definition 2.8. With the mathematical tools at our disposal we cannot give a completely rigorous construction, and this is not needed for our purposes, but we describe the main ideas in the technical appendix of this chapter (Section 2.8). We summarize the main point in the following theorem. It gives all we need for the theoretical developments and calculations in the sequel.

Theorem 2.13. *Suppose a transition probability p on a countable state space $\mathcal{S}$ is given. Then there exists a sample space Ω with $\mathcal{S}$-valued random variables $\{X_k\}_{k\geq 0}$ defined on Ω such that, for every initial distribution μ, there exists a probability measure P_μ on Ω that satisfies the following equation for all states $x_0,\dots,x_n\in\mathcal{S}$:*

$$
P_\mu(X_0=x_0,X_1=x_1,\dots,X_n=x_n)=\mu(x_0)\prod_{k=0}^{n-1}p(x_k,x_{k+1}).
\tag{2.26}
$$

The calculation in (2.25) *then shows that the stochastic process $\{X_k\}_{k\geq 0}$ defined on $(\Omega,\mathcal{F},P_\mu)$ is a Markov chain with initial distribution μ and transition probability p.*

We address some further notational and foundational issues.

The degenerate probability measure δ_z that puts all its probability on a single point z satisfies

$$
\delta_z(x)=\begin{cases}1, & x=z\\ 0, & x\neq z.\end{cases}
$$

It is called the *point mass* at z or the *Dirac delta* at z. When the initial distribution $\mu = \delta_z$, the initial state is not really random because $P_\mu(X_0 = z) = 1$. In this case P_μ is denoted by P_z. The equation corresponding to (2.26) now reads

$$P_z(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \delta_z(x_0) \prod_{k=0}^{n-1} p(x_k, x_{k+1}). \tag{2.27}$$

The following is a useful formula for calculations: for any event A concerning the Markov chain and any initial distribution μ,

$$P_\mu(A) = \sum_x \mu(x) P_x(A). \tag{2.28}$$

If $A = \{X_0 = x_0, \ldots, X_n = x_n\}$ then this identity follows from (2.26) and (2.27). It can be verified for all events, but the tools for doing so require measure theory. (2.28) can also be understood as a consequence of conditioning on the first state:

$$P_\mu(A) = \sum_x P_\mu(X_0 = x) P_\mu(A \,|\, X_0 = x) = \sum_x \mu(x) P_x(A).$$

If Y is a real-valued random variable on the probability space $(\Omega, \mathcal{F}, P_\mu)$, then the expectation of Y is denoted by $E_\mu[Y]$. If $\mu = \delta_x$, then the expectation is denoted by $E_x[Y]$.

We illustrate these formulas with examples. By appeal to Theorem 2.13, we can produce examples of Markov chains simply by writing down a transition probability on a finite or countably infinite state space.

Example 2.14. Let the state space be $\mathcal{S} = \{0, 1\}$ and the transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{bmatrix} \frac14 & \frac34 \\ \frac12 & \frac12 \end{bmatrix} \end{array}.$$

(a) Suppose the chain starts in state 0. What is the probability that $X_1 = 1$?

$$P_0(X_1 = 1) = p(0,1) = \tfrac34.$$

(b) Suppose the chain starts in state 0. What is the probability that $X_2 = 0$?

$$\begin{aligned} P_0(X_2 = 0) &= P_0(X_1 = 0, X_2 = 0) + P_0(X_1 = 1, X_2 = 0) \\ &= p(0,0)p(0,0) + p(0,1)p(1,0) = \tfrac14 \cdot \tfrac14 + \tfrac34 \cdot \tfrac12 = \tfrac{7}{16}. \end{aligned}$$

(c) Suppose the chain is equally likely to start in either state 0 or 1. What is the probability that it remains in its starting state for at least two further time steps?

The initial distribution is $\mu(0) = \mu(1) = \frac12$.

$$\begin{aligned} &P_\mu(X_2 = X_1 = X_0) = P_\mu(X_2 = X_1 = X_0 = 0) + P_\mu(X_2 = X_1 = X_0 = 1) \\ &\quad = \mu(0)p(0,0)p(0,0) + \mu(1)p(1,1)p(1,1) = \tfrac12 \cdot \tfrac14 \cdot \tfrac14 + \tfrac12 \cdot \tfrac12 \cdot \tfrac12 = \tfrac{1}{32} + \tfrac18 = \tfrac{5}{32}. \end{aligned}$$

(d) Suppose the chain is equally likely to start in either state 0 or 1. Suppose I get a reward of 2 dollars if $X_1 = 0$ and a reward of 3 dollars if $X_1 = 1$. Find the expected reward.

Let Y denote the amount of the reward.

$$\begin{aligned} E_\mu[Y] &= \sum_k kP_\mu(Y=k) = 2\cdot P_\mu(X_1=0) + 3\cdot P_\mu(X_1=1) \\ &= 2\big(\mu(0)p(0,0)+\mu(1)p(1,0)\big) + 3\big(\mu(0)p(0,1)+\mu(1)p(1,1)\big) \\ &= 2\big(\tfrac12\cdot\tfrac14+\tfrac12\cdot\tfrac12\big) + 3\big(\tfrac12\cdot\tfrac34+\tfrac12\cdot\tfrac12\big) = \tfrac68+\tfrac{15}{8} = \tfrac{21}{8}. \end{aligned}$$

△

Example 2.15. Let the state space be $\mathcal{S}=\{1,2,3\}$ and the transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac13 & \frac13 & \frac13 \\ 0 & \frac12 & \frac12 \\ 0 & \frac14 & \frac34 \end{bmatrix} \end{array}.$$

Examples of probabilities:

$$\begin{aligned} P_1(X_1=1,X_2=2,X_3=2,X_4=3) &= p(1,1)p(1,2)p(2,2)p(2,3) \\ &= \tfrac13\cdot\tfrac13\cdot\tfrac12\cdot\tfrac12 = \tfrac{1}{36}. \end{aligned}$$

$$P_3(X_1=2,X_2=1) = p(3,2)p(2,1) = \tfrac14\cdot 0 = 0.$$

△

The next two examples are well-known ones with names.

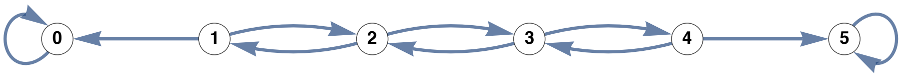

Figure 1. Transition diagram for gambler's ruin in Example 2.16.

Example 2.16 (Gambler's ruin)**.** You play repeatedly a game of chance where you win a dollar with probability p and lose a dollar with probability $1-p$, where $0<p<1$ is a fixed parameter of the model. Successive games are independent. You begin with x dollars in your pocket with $0\le x\le 5$. If you lose all your money you stop playing. Similarly, if you reach 5 dollars, you stop playing and keep your winnings. Construct a Markov chain that keeps track of the amount of money in your pocket.

The state space is the set $\mathcal{S}=\{0,1,2,3,4,5\}$. If you are in state 0 or 5 then everything stops: these are *absorbing states* and we define the transition probability for these states by $p(0,0)=p(5,5)=1$.

From the other states $x\in\{1,2,3,4\}$ you go up or down with probability p and $1-p$, so $p(x,x+1)=p$ and $p(x,x-1)=1-p$. The other values $p(x,y)$ not specified are zero.

The possible transitions are illustrated by the arrows in Figure 1. Here is the transition probability matrix.

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 & 0 \\ 0 & 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \tag{2.29}$$

Away from the boundary points this Markov chain evolves like a simple random walk. Hence this example can also be called *simple random walk with absorbing boundaries.*

Does the player eventually get absorbed either in state 0 or state 5, or can the Markov chain remain in $\{1, 2, 3, 4\}$ forever? We prove that absorption is certain. For this purpose it is convenient to introduce again the i.i.d. random walk steps $\{Y_k\}$ used in Example 2.1 that satisfy $P(Y_k = 1) = p = 1 - P(Y_k = -1)$. Now $Y_k = -1$ means that game k was a loss of a dollar and $Y_k = 1$ that game k was a win of a dollar. The entire process can be defined in terms of these step variables by

$$X_n = \begin{cases} X_{n-1} + Y_n, & \text{if } X_{n-1} \in \{1, 2, 3, 4\} \\ X_{n-1}, & \text{if } X_{n-1} \in \{0, 5\}. \end{cases}$$

Make the simple observation that five losses in a row end the game, no matter what the current state is. Thus we can reason as follows.

$$\begin{aligned} &P_x(\text{absorption never happens}) \\ &\le P_x(\text{absorption has not happened by time } 5n) \\ &\le P_x\big\{(Y_{5k+1}, \dots, Y_{5(k+1)}) \ne (-1, \dots, -1) \ \text{ for } k = 0, 1, \dots, n-1\big\} \\ &= \prod_{k=0}^{n-1} P_x\big\{(Y_{5k+1}, \dots, Y_{5(k+1)}) \ne (-1, \dots, -1)\big\} \\ &= \big(1 - (1-p)^5\big)^n. \end{aligned}$$

The crucial step above was the independence of the Y_ks. This upper bound is true for all n. We can let $n \to \infty$ and conclude that

$$P_x(\text{absorption never happens}) \le \lim_{n\to\infty} \big(1 - (1-p)^5\big)^n = 0.$$

In conclusion, for all initial states $x \in \{0, \dots, 5\}$,

$$P_x(\text{absorption happens eventually}) = 1.$$

△

Example 2.17 (Success run chain)**.** Independent trials are repeated. Each trial is a success with probability α and a failure with probability $1 - \alpha$. The Markov chain X_n keeps track of the length of the current run of successes. Each success increases X_n by one, and each failure sends X_n back to zero.

To define the model, take the state space $\mathcal{S} = \mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \dots\}$, the set of nonnegative integers. The transition probability is given for all $k \in \mathcal{S}$ by

$$p(k, k+1) = \alpha, \quad p(k, 0) = 1 - \alpha, \quad \text{and} \quad p(k, \ell) = 0 \ \text{ for } \ell \notin \{0, k+1\}.$$

We can wrap this example in a story of shooting free throws with a basketball. My shots succeed with probability p independently of every other shot. Suppose I have succeeded three times in a row since my last miss. What is the expected maximal success run I reach before my next miss?

Let Z be the random variable of interest, that is, the maximal success run I reach before my next miss. The question concerns the expectation $E_3[Z]$ of Z when the process starts in state 3. Under P_3 the possible values of Z are $n \geq 3$, and the probability mass function is

$$P_3(Z = n) = p^{n-3}(1 - p).$$

Thus

$$\begin{aligned} E_3[Z] &= \sum_{n=3}^{\infty} n\, P_3(Z = n) = \sum_{n=3}^{\infty} np^{n-3}(1-p) = \sum_{m=0}^{\infty} (3+m)p^m(1-p) \\ &= 3\sum_{m=0}^{\infty} p^m(1-p) + \sum_{m=1}^{\infty} mp^m(1-p) = 3 + \sum_{m=1}^{\infty}\sum_{n=1}^{m} p^m(1-p) \\ &= 3 + \sum_{n=1}^{\infty}\sum_{m=n}^{\infty} p^m(1-p) = 3 + \sum_{n=1}^{\infty} p^n = 3 + \frac{p}{1-p}. \end{aligned}$$

The switch in the order of summation done above is one of the several ways of evaluating a series of the type $\sum np^n$. △

Not every process is a Markov chain. Here is an example.

Example 2.18. Define the process $\{X_k\}_{k \in \mathbb{Z}_{\geq 0}}$ as follows. Let X_0 and X_1 be independent Bernoulli(1/2) random variables, that is, their probability mass function is

$$P(X_i = 0) = P(X_i = 1) = \tfrac{1}{2} \quad \text{for } i = 0, 1.$$

For $n \geq 2$ define $X_n = X_{n-2}$. This means that $X_0, X_2, X_4, \dots$ and $X_1, X_3, X_5, \dots$ are both constant sequences but their values are chosen randomly and independently. We check that this is not a Markov chain by showing that (2.14) fails for some n and $x_0, \dots, x_n$.

From the construction of the process it follows that for any fixed $n \geq 1$ the random variables X_{n-1}, X_n are independent. In particular,

$$P(X_2 = 1 \,|\, X_1 = 1) = P(X_2 = 1) = \tfrac{1}{2}.$$

On the other hand, since $X_2 = X_0$,

$$P(X_2 = 1 \,|\, X_1 = 1, X_0 = 0) = 0.$$

Thus (2.14) fails for $n = 2$ with $x_1 = 1, x_0 = 0$. $\{X_k\}_{k \in \mathbb{Z}_{\geq 0}}$ is not a Markov chain. △

Multistep transition probabilities. Equation (2.20) gives the meaning of the transition probability $p(x, y)$ as the conditional probability of moving to y in the next time step, given that we are presently in state x. We derive a formula for multistep transition probabilities, that is, transition probabilities that look more than one step ahead.

Suppose $\{X_n\}$ is a Markov chain on $(\Omega, \mathcal{F}, P)$ with state space $\mathcal{S}$ and transition probability p and assume that $P(X_n = x) > 0$. Let $k \geq 2$. By (2.23),

$$\begin{aligned}
&P(X_{n+k} = y \,|\, X_n = x) \\
&= \sum_{z_1,\dots,z_{k-1}\in\mathcal{S}} P(X_{n+1} = z_1, \dots, X_{n+k-1} = z_{k-1}, X_{n+k} = y \,|\, X_n = x) \\
&= \sum_{z_1,\dots,z_{k-1}\in\mathcal{S}} \frac{P(X_n = x, X_{n+1} = z_1, \dots, X_{n+k-1} = z_{k-1}, X_{n+k} = y)}{P(X_n = x)} \\
&\overset{(2.23)}{=} \sum_{z_1,\dots,z_{k-1}\in\mathcal{S}} \frac{P(X_n = x)p(x, z_1)p(z_1, z_2)\cdots p(z_{k-1}, y)}{P(X_n = x)} \\
&= \sum_{z_1,\dots,z_{k-1}\in\mathcal{S}} p(x, z_1)p(z_1, z_2)\cdots p(z_{k-1}, y).
\end{aligned}$$

We take the outcome of the calculation above and define the *k-step transition probability* for $x, y \in \mathcal{S}$ and $k \geq 2$ as

$$p^{(k)}(x, y) = \sum_{z_1,\dots,z_{k-1}\in\mathcal{S}} p(x, z_1)\, p(z_1, z_2) \cdots p(z_{k-1}, y). \tag{2.30}$$

To have $p^{(k)}(x, y)$ defined for all $k \geq 0$ we complete this definition by setting

$$p^{(0)}(x, y) = \delta_x(y) \quad \text{and} \quad p^{(1)}(x, y) = p(x, y) \quad \text{for } x, y \in \mathcal{S}. \tag{2.31}$$

Setting $p^{(0)}(x, y) = 1$ for $y = x$ and zero otherwise agrees with common sense since "zero steps ahead" means not looking beyond the present state x. With these definitions the following *Chapman-Kolmogorov equations* are in force: for all integers $m, n \geq 0$ and all states $x, y \in \mathcal{S}$,

$$p^{(m+n)}(x, y) = \sum_{z\in\mathcal{S}} p^{(m)}(x, z)p^{(n)}(z, y). \tag{2.32}$$

Exercise 2.3 asks you to verify (2.32) algebraically from the definition of $p^{(k)}(x, y)$. Probabilistically (2.32) has a straightforward meaning as a decomposition according to the state z visited after m steps on the way from x to y in $m + n$ steps.

We write the multistep transition probabilities as $p^{(k)}(x, y)$ instead of $p^k(x, y)$ to highlight the fact that $p^{(k)}(x, y)$ is *not* the quantity $p(x, y)$ raised to the power k. However, an algebraic power behind this definition becomes visible from the matrix point of view. Consider now the transition probability as the matrix $\mathbf{P} = \{p(x, y)\}_{x,y\in\mathcal{S}}$ indexed by the state space $\mathcal{S}$, and recall the definition of matrix multiplication. Equations (2.30)–(2.31) then imply that $p^{(k)}(x, y)$ is precisely the (x, y)-entry of the kth power $\mathbf{P}^k$ of the matrix $\mathbf{P}$.

Theorem 2.19. *Suppose $\{X_n\}$ is a Markov chain with transition probability matrix $\mathbf{P} = \{p(x, y)\}_{x,y\in\mathcal{S}}$. Then the multistep transition probabilities of $\{X_k\}$ are given* *by the powers of* $\mathbf{P}$*. Namely, for* $n, k \geq 0$ *and states* x, y *such that* $P(X_n = x) > 0$*,*

$$P(X_{n+k} = y \,|\, X_n = x) = p^{(k)}(x, y)$$

where $\mathbf{P}^k = \{p^{(k)}(x, y)\}_{x,y \in \mathcal{S}}$*.*

Example 2.20. Consider the Markov chain with state space $\mathcal{S} = \{0, 1\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}.$$

Start the process in state 0. What is the probability that the process is in state 0 at time n?

The probability in question is $P_0(X_n = 0) = p^{(n)}(0, 0)$. Calculating the powers of the transition matrix is straightforward. In fact, we claim that for all $n \geq 0$,

$$\mathbf{P}^n = \begin{bmatrix} \left(\frac{1}{2}\right)^n & 1 - \left(\frac{1}{2}\right)^n \\ 0 & 1 \end{bmatrix}.$$

The claim is true for $n = 0$ and $n = 1$ just by inspection. We prove the induction step, that is, assuming the claim true for n we check that it holds also for $n + 1$.

$$\begin{aligned} \mathbf{P}^{n+1} = \mathbf{P}^n \cdot \mathbf{P} &= \begin{bmatrix} \left(\frac{1}{2}\right)^n & 1 - \left(\frac{1}{2}\right)^n \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \left(\frac{1}{2}\right)^n \cdot \frac{1}{2} & \left(\frac{1}{2}\right)^n \cdot \frac{1}{2} + 1 - \left(\frac{1}{2}\right)^n \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \left(\frac{1}{2}\right)^{n+1} & 1 - \left(\frac{1}{2}\right)^{n+1} \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

Having verified the formula for all powers $\mathbf{P}^n$, we can read off the answer from the $(0, 0)$-element of the matrix:

$$p^{(n)}(0, 0) = (\mathbf{P}^n)_{0,0} = \left(\tfrac{1}{2}\right)^n.$$

△

As the last item of the present discussion, we derive the probability distribution of the state X_n at time n and a formula for expectations of functions of X_n. Let $\{X_n\}$ be a Markov chain with transition probability matrix $\mathbf{P} = \{p(x, y)\}_{x,y \in \mathcal{S}}$ and initial distribution μ. Then utilizing formula (2.23), for any $n \geq 0$ and state $y \in \mathcal{S}$,

$$\begin{aligned} P_\mu(X_n = y) &= \sum_{x_0, \dots, x_{n-1} \in \mathcal{S}} P_\mu(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = y) \\ &= \sum_{x_0, \dots, x_{n-1} \in \mathcal{S}} \mu(x_0) p(x_0, x_1) \cdots p(x_{n-1}, y) \\ &= \sum_{x_0 \in \mathcal{S}} \mu(x_0) \sum_{x_1, \dots, x_{n-1} \in \mathcal{S}} p(x_0, x_1) \cdots p(x_{n-1}, y) \\ &= \sum_{x_0 \in \mathcal{S}} \mu(x_0) p^{(n)}(x_0, y) \end{aligned}$$

Think of the probability measure μ on $\mathcal{S}$ as a row vector. Then the last expression above is $(\mu\mathbf{P}^n)_y$, namely, the element indexed by y in the row vector $\mu\mathbf{P}^n$, obtained by multiplying the row vector μ by the matrix $\mathbf{P}^n$ from the right. If we define the row vector $\mu^{(n)}(y) = P_\mu(X_n = y)$ as the distribution of X_n, then the equation above can be expressed succinctly as

$$\mu^{(n)} = \mu\mathbf{P}^n. \tag{2.33}$$

Let now $f : \mathcal{S} \to \mathbb{R}$ be a nonnegative or a bounded function. (These assumptions are made to guarantee that expectations are well-defined.) Take two integers $m, n \ge 0$, a state $x \in \mathcal{S}$, and assume that $P(X_m = x) > 0$. We compute first a conditional expectation.

$$\begin{aligned} E[f(X_{m+n}) \,|\, X_m = x] &= \sum_{y\in\mathcal{S}} P(X_{m+n} = y \,|\, X_m = x) f(y) \\ &= \sum_{y\in\mathcal{S}} p^{(n)}(x, y) f(y). \end{aligned}$$

The last formula above is the same as $(\mathbf{P}^n f)_x$, namely, the x-entry of the column vector obtained by multiplying the column vector f by the matrix $\mathbf{P}^n$ from the left. In particular, we get the formula

$$E_x[f(X_n)] = (\mathbf{P}^n f)_x. \tag{2.34}$$

Markov property to the infinite future. In calculations we frequently need to consider complicated future events, well beyond the one-step look ahead $X_{n+1} = y$ of (2.20), and we also need to condition on past events that do not simply specify a particular evolution for $X_0, \dots, X_n$. We state below a version of the Markov property that captures the most general and useful formulation. The key idea is that if we know the present state, then the past does not influence the probabilities of the future, and each time the Markov chain restarts itself anew, using the current state as the new initial state. The theorem below is proved in the technical appendix Section 2.8.

The get an idea of the formula to expect, consider first the case that looks two steps into the future. Let $y_0, \dots, y_n, x_{n+1}, x_{n+2}$ be states such that the conditioning events below have positive probability.

$$\begin{aligned} &P_\mu(X_{n+1} = x_1, X_{n+2} = x_{n+2} \,|\, Y_0 = y_0, \dots, Y_n = y_n) \\ &= \frac{P_\mu(Y_0 = y_0, \dots, Y_n = y_n, X_{n+1} = x_1, X_{n+2} = x_{n+2})}{P_\mu(Y_0 = y_0, \dots, Y_n = y_n)} \\ &= \frac{\mu(y_0)\big(\prod_{i=1}^n p(y_{i-1}, y_i)\big) p(y_n, x_1) p(x_1, x_2)}{\mu(y_0) \prod_{i=1}^n p(y_{i-1}, y_i)} \\ &= p(y_n, x_1) p(x_1, x_2) \\ &= P_{y_n}(X_1 = x_1, X_2 = x_2). \end{aligned}$$

The reader can easily imagine how the formula generalizes to looking m steps into the future:

$$\begin{aligned} &P_\mu(X_{n+1} = x_1, \dots, X_{n+m} = x_m \,|\, Y_0 = y_0, \dots, Y_n = y_n) \\ &= P_{y_n}(X_1 = x_1, \dots, X_m = x_m). \end{aligned} \tag{2.35}$$

In plain English, the formula tells us that when we condition the future of a Markov chain on the past, the conditional probabilities are exactly the same as the probabilities of a new Markov chain started from the present state. Equation (2.35) is the basis for the derivation of the most general formulation presented below in (2.36). The derivation is sketched in Section 2.8.

The formulation of equation (2.36) below is abstract because it is tailored to serve all our needs. For this reason it needs some explanation. Start with the basic point that any event concerning a random variable X can be expressed in the form $\{X \in B\}$ where B is a set of real numbers. For example, $\{a \le X \le b\} = \{X \in [a,b]\}$. Similarly, any event concerning an n-dimensional random vector can be written in the form $\{(X_1, \dots, X_n) \in B\}$ where B is a subset of $\mathbb{R}^n$.

Entirely analogously, any event concerning the process $(X_n, X_{n+1}, X_{n+2}, \dots)$ is of the form $\{(X_n, X_{n+1}, X_{n+2}, \dots) \in U\}$ where U is some particular set of sequences of states. For example, the event that the stochastic process $\{X_k\}_{k\ge 0}$ is eventually absorbed in state 1 can be expressed as $\{(X_0, X_1, X_2, \dots) \in U\}$ where

$$U = \{(x_0, x_1, x_2, \dots) : \text{for some } m,\ x_k = 1 \text{ for all } k \ge m\}$$

is a set-theoretic representation of the set of sequences that eventually become constant 1.

Theorem 2.21. *Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$ and initial distribution μ. Let $x \in \mathcal{S}$, $B \subseteq \mathcal{S}^{n+1}$ and let U be any set of sequences of states. Then for any initial distribution μ,*

$$\begin{aligned} P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U \bigm| X_n = x, (X_0, \dots, X_n) \in B\big] \\ = P_x\big[(X_0, X_1, X_2, \dots) \in U\big] \end{aligned} \tag{2.36}$$

provided the conditioning event has positive probability.

As an application we calculate the probability of winning in gambler's ruin.

Example 2.22. Consider gambler's ruin from Example 2.16. This time we take the state space $\mathcal{S} = \{0, 1, \dots, N\}$ for some positive integer N and symmetric transition probabilities:

$$p(0,0) = p(N,N) = 1, \quad p(x, x-1) = p(x, x+1) = \tfrac{1}{2} \quad \text{for } 1 \le x \le N-1. \tag{2.37}$$

We ask about the probability of winning, in other words, that the Markov chain absorbs in state N rather than in state 0. As a function of the initial state x, denote this probability by

$$h(x) = P_x(\text{state } N \text{ is reached before state } 0).$$

The objective is to solve for $h(x)$. The boundary values satisfy

$$h(0) = 0 \quad \text{and} \quad h(N) = 1. \tag{2.38}$$

If we start at state $x \in \{1, 2, \dots, N-1\}$, then first we take a step left or right with probability $\frac{1}{2}$ each, and then restart the Markov chain from the new state. Hence

$$h(x) = \tfrac{1}{2}h(x-1) + \tfrac{1}{2}h(x+1) \qquad \text{for } 0 < x < N. \tag{2.39}$$

Here is a precise derivation to justify this somewhat heuristic argument. Let U denote the set of sequences of states that visit N before 0. Then $h(x)$ can be expressed as

$$h(x) = P_x\{(X_0, X_1, X_2, \dots) \in U\}.$$

If $x \neq 0, N$ then the first state makes no difference to the condition of visiting 0 or N and it is equivalent to write

$$h(x) = P_x\{(X_1, X_2, X_3, \dots) \in U\}.$$

With these preliminaries, we derive (2.39). Let $0 < x < N$.

$$\begin{aligned} h(x) &= P_x\{(X_1, X_2, X_3, \dots) \in U\} \\ &= P_x\{X_1 = x-1, (X_1, X_2, X_3, \dots) \in U\} \\ &\qquad + P_x\{X_1 = x+1, (X_1, X_2, X_3, \dots) \in U\} \\ &= P_x(X_1 = x-1)\, P_x\{(X_1, X_2, X_3, \dots) \in U \,|\, X_1 = x-1\} \\ &\qquad + P_x(X_1 = x+1)\, P_x\{(X_1, X_2, X_3, \dots) \in U \,|\, X_1 = x+1\} \\ &\overset{(a)}{=} \tfrac{1}{2}\, P_{x-1}\{(X_0, X_1, X_2, \dots) \in U\} + \tfrac{1}{2}\, P_{x+1}\{(X_0, X_1, X_2, \dots) \in U\} \\ &= \tfrac{1}{2} h(x-1) + \tfrac{1}{2} h(x+1). \end{aligned}$$

Step (a) above used the Markov property in the form (2.36).

From (2.38) and (2.39) we can solve h in a couple steps. Rearrange (2.39) into an equation for the increments of h and iterate it:

$$h(x+1) - h(x) = h(x) - h(x-1) = h(x-1) - h(x-2) = \cdots = h(1) - h(0).$$

Since all the increments are equal, we get

$$1 = h(N) - h(0) = \sum_{k=1}^{N} \bigl(h(k) - h(k-1)\bigr) = N\bigl(h(x) - h(x-1)\bigr)$$

for any particular x, from which

$$h(x) - h(x-1) = \frac{1}{N}.$$

Now we can calculate every value:

$$h(x) = h(x) - h(0) = \sum_{k=1}^{x} \bigl(h(k) - h(k-1)\bigr) = \frac{x}{N}.$$

The conclusion is that if the player starts with x dollars, their probability of reaching N dollars before going broke is x/N. △

2.2. Strong Markov property

The next stage is a conceptually nontrivial enhancement of the Markov property. We begin with a motivating question.

Consider a two-state Markov chain with state space $\mathcal{S} = \{0, 1\}$ and transition matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

The transition matrix tells us that $P(X_{n+1} = 1 \,|\, X_n = 0) = \frac{3}{4}$. It is important to realize that n represents here a *fixed deterministic time.* In other words, the previous statement represents statements such as $P(X_9 = 1 \,|\, X_8 = 0) = \frac{3}{4}$ and $P(X_{22} = 1 \,|\, X_{21} = 0) = \frac{3}{4}$, for any integer value of n. The important distinction is that the time index n is *not* random.

An example of the same question at a random time would be the following. What is the probability that the process jumps to state 1 right after the second visit to 0? The time of the second visit to 0 is *random.* The Markov property as developed so far does not cover this case. Yet intuition tells us that the answer must be the same $\frac{3}{4}$ since the past should not matter. In mathematical symbols, if T denotes the random time of the second visit to 0, we would want to conclude that $P(X_{T+1} = 1 \,|\, X_T = 0) = \frac{3}{4}$.

The answer to the question above is deceptively obvious and may lead the reader to think there is no issue here. So consider this question. Let N be the random time of that first visit to state 0 that is immediately followed by a visit to state 1. What is now the probability that at the next time after N we have $X_{N+1} = 1$? The probability must be 1, simply because the definition of N forces $X_{N+1} = 1$.

In the first situation the answer came from the transition matrix, and in the second situation it did not. What is the precise difference between the two that accounts for this? Resolution of this issue involves the concept of a stopping time.

Definition 2.23. Let $\{X_n\}_{n\geq 0}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$. Let T be a random variable on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{Z}_{\geq 0} \cup \{\infty\}$. Then T is a **stopping time** (for the process $\{X_n\}$) if for each nonnegative integer n, there is a subset $C_n \subset \mathcal{S}^{n+1}$ such that this equality of events holds:

$$\{T = n\} = \{(X_0, \dots, X_n) \in C_n\}.$$

△

More informally, the definition can be stated as follows: for each nonnegative integer n, the values $(X_0, \dots, X_n)$ determine whether $T = n$ happens or not, or, that the event $\{T = n\}$ can be expressed in terms of $(X_0, \dots, X_n)$. In colloquial terms, a stopping time is a random time that is not allowed to look into the future.

Example 2.24. Consider from above the time T of the second visit to 0 for the Markov chain with state space $\{0, 1\}$. We can express the event $\{T = n\}$ as

$$\{T = n\} = \{(X_0, \dots, X_{n-1}) \text{ contains exactly one } 0, X_n = 0\}.$$

The key point is that the event $\{T = n\}$ involves the random variables $(X_0, \dots, X_n)$ and nothing beyond time n. Thus T is a stopping time.

A notational comment: expressing events with a mixture of mathematical notation and English sentences works well in probability, as above for $\{T = n\}$. A purely symbolic formula can also be used, such as this version:

$$\begin{gathered}\{T = n\} = \{\exists k \in \{0, \dots, n-1\} : X_0 = \cdots = X_{k-1} = 1, X_k = 0, \\ X_{k+1} = \cdots = X_{n-1} = 1, X_n = 0\}.\end{gathered}$$

The quantifier ∃ is short for "there exists". A more basic but tedious way is to list the alternatives:

$$\begin{aligned}\{T = n\} =&\{(X_0, \dots, X_n) = (0, 1, \dots, 1, 0)\}\\ &\cup \{(X_0, \dots, X_n) = (1, 0, 1, \dots, 1, 0)\}\\ &\cup \{(X_0, \dots, X_n) = (1, 1, 0, 1, \dots, 1, 0)\}\\ &\cup \dots \cup \{(X_0, \dots, X_n) = (1, \dots, 1, 0, 0)\}.\end{aligned}$$

The ellipsis in $1, \dots, 1$ means that the omitted entries are all 1s. △

Example 2.25. A more general version of the example above is a stopping time that records the first time the process enters some state or a subset of the state space. For a subset $A \subset \mathcal{S}$, this stopping time is defined by

$$\tau_A = \inf\{n \geq 0 : X_n \in A\}.$$

The symbol inf is short for *infimum*. In this context it means the same as *minimum*, the smallest element of the set $\{n \geq 0 : X_n \in A\}$. The convention is that if a set is empty, then its infimum is infinite. In other words, the value $\tau_A = \infty$ means that the process never enters the set A.

The proof that τ_A is a stopping time goes simply like this:

$$\{\tau_A = n\} = \{X_0 \in A^c, \dots, X_{n-1} \in A^c, X_n \in A\}.$$

The equality shows that the event $\{\tau_A = n\}$ is equal to an event that depends only on $X_0, \dots, X_n$. △

Example 2.26. The second example from the introductory discussion above was the time N of the first visit to state 0 that is immediately followed by a visit to state 1. In order to know that $N = n$, we must know that $X_n = 0$ and $X_{n+1} = 1$, in addition to past information. Since the event $\{N = n\}$ depends on the future value X_{n+1}, N cannot be a stopping time. A precise expression for the event $\{N = n\}$ is

$$\{N = n\} = \{(X_k, X_{k+1}) \neq (0, 1) \text{ for } k = 0, \dots, n-1,\ (X_n, X_{n+1}) = (0, 1)\}.$$

△

After this discussion, we come to the main theorem about stopping times and Markov chains.

Theorem 2.27 (Strong Markov property)**.** *Let X_n be a Markov chain with arbitrary initial distribution μ and transition probability $p(x, y)$. Let T be a stopping time, and let A be any event determined by $(X_0, \dots, X_T)$, that is, by the process up to the stopping time T. Then, for any state x and states $u_1, \dots, u_m$,*

$$\begin{aligned}P_\mu[\, X_{T+1} = u_1, \dots, X_{T+m} = u_m \,|\, T < \infty, A, X_T = x\,]\\ = P_x[\, X_1 = u_1, \dots, X_m = u_m\,],\end{aligned} \tag{2.40}$$

provided the conditioning event has positive probability.

Here are some special cases of equation (2.40). If it happens that $P_\mu(T < \infty) = 1$, then the condition "$T < \infty$" makes no difference to any calculation and it can be dropped from the conditioning event in (2.40). If we look only one step ahead ($m = 1$) and assume $P_\mu(T < \infty) = 1$, then (2.40) becomes

$$P_\mu[X_{T+1} = y \,|\, A, X_T = x] = p(x, y). \tag{2.41}$$

We can take $A = \Omega$ in (2.40). Since Ω has probability 1 it does not affect any probability calculation and can be dropped. Then (2.40) becomes

$$\begin{aligned} &P_\mu[\, X_{T+1} = u_1, \dots, X_{T+m} = u_m \,|\, T < \infty, X_T = x \,] \\ &\qquad = P_x[\, X_1 = u_1, \dots, X_m = u_m \,]. \end{aligned} \tag{2.42}$$

Equation (2.40) extends to any event concerning the entire future process after time T, including also the value at time T: for any (measurable) subset $U \subset \mathcal{S}^{\mathbb{Z}_{\geq 0}}$,

$$\begin{aligned} &P_\mu[\, (X_T, X_{T+1}, X_{T+2}, \dots) \in U \,|\, T < \infty, A, X_T = x \,] \\ &\qquad = P_x\big[(X_0, X_1, X_2, \dots) \in U\big]. \end{aligned} \tag{2.43}$$

In plain words this says that if T is a finite stopping time then the process $Z_k = X_{T+k}$ is a Markov chain with the same transition probability p as X_n.

Example 2.28. Return to the question posed at the beginning of this section. Let T be the time of the second visit to state 0 in a Markov chain with state space $\{0, 1\}$. Find the probability that $X_{T+1} = 1$. In the next section we learn that $T < \infty$ with probability 1, no matter how the process is started. We take that for granted here. The calculation goes by arranging the probability into a form to which we can apply the strong Markov property.

$$\begin{aligned} P_\mu(X_{T+1} = 1) &= P_\mu(X_T = 0,\ X_{T+1} = 1) \\ &= P_\mu(X_T = 0)\, P_\mu(X_{T+1} = 1 \,|\, X_T = 0) \\ &= 1 \cdot p(0, 1) = \tfrac{3}{4}. \end{aligned}$$

$P_\mu(X_T = 0) = 1$ because the definition of T says that $X_T = 0$. Then we used the strong Markov property in the form (2.41). △

Now for the proof of the strong Markov property. This proof is textbook appropriate because it utilizes a decomposition of an event, followed by the Markov property, followed by undoing the decomposition.

Proof of Theorem 2.27. We prove the equivalent statement

$$\begin{aligned} &P_\mu[\, X_{T+1} = u_1, \dots, X_{T+m} = u_m,\ T < \infty, A, X_T = x \,] \\ &\qquad = P_x[\, X_1 = u_1, \dots, X_m = u_m \,] \cdot P_\mu[\, T < \infty, A, X_T = x \,]. \end{aligned} \tag{2.44}$$

Statement (2.40) follows by dividing by $P_\mu[\, T < \infty, A, X_T = x \,]$.

To prove (2.44), decompose the probability on the left according to the value of T and apply the basic Markov property in the form (2.36).

$$
\begin{aligned}
&P_\mu[\,X_{T+1}=u_1,\dots,X_{T+m}=u_m,\,T<\infty,A,X_T=x\,]\\
&=\sum_{n=0}^{\infty}P_\mu[\,X_{T+1}=u_1,\dots,X_{T+m}=u_m,\,T=n,A,X_T=x\,]\\
&=\sum_{n=0}^{\infty}P_\mu[\,X_{n+1}=u_1,\dots,X_{n+m}=u_m,\,T=n,A,X_n=x\,]\\
&=\sum_{n=0}^{\infty}P_\mu[\,X_{n+1}=u_1,\dots,X_{n+m}=u_m\,|\,T=n,A,X_n=x\,]\\
&\qquad\qquad\qquad\cdot P_\mu[\,T=n,A,X_n=x\,]\\
&\overset{(2.36)}{=}\sum_{n=0}^{\infty}P_x[\,X_1=u_1,\dots,X_m=u_m\,]\cdot P_\mu[\,T=n,A,X_n=x\,]\\
&=P_x[\,X_1=u_1,\dots,X_m=u_m\,]\sum_{n=0}^{\infty}P_\mu[\,T=n,A,X_n=x\,]\\
&=P_x[\,X_1=u_1,\dots,X_m=u_m\,]\cdot P_\mu[\,T<\infty,A,X_T=x\,].
\end{aligned}
$$

The use of the Markov property in the fourth equality is justified by the assumption that the event $\{T=n,A\}$ involves the random variables $(X_0,\dots,X_n)$ and not beyond. □

2.3. Recurrence and transience

Our tools are in place and we can turn to study the behavior of Markov chains. The first issue is the question of recurrence and transience of a particular state. A recurrent state has the property that if the process starts from this state, the process is sure to return to it at some later time. Transience means the opposite, namely that there is some positive probability of never returning. After the introduction of the necessary notation, this notion is made precise in Definition 2.29 below. This section culminates in the canonical decomposition of the state space given in Theorem 2.48.

Let

$$T_x=\inf\{n\ge 1: X_n=x\} \tag{2.45}$$

be the first time after time zero that the Markov chain enters state x. T_x is a stopping time, as is evident from this identity of events:

$$\{T_x=n\}=\{X_1\ne x,\dots,X_{n-1}\ne x, X_n=x\}.$$

Note that the stopping time T_x does not care about whether the process started in state x or not. For states $x,y\in\mathcal{S}$, let

$$\rho_{xy}=P_x(T_y<\infty) \tag{2.46}$$

denote the probability that the process, started from x, at some later time visits state y. In particular, ρ_{xx} is the probability that the process returns to x once it starts at x.

Quantity ρ_{xy} is not exactly a transition probability but it is related. A lower bound comes by dropping events from a union:

$$\begin{aligned}\rho_{xy} &= P_x(X_n = y \text{ for some } n \ge 1) = P_x\Bigl(\bigcup_{n\ge1}\{X_n = y\}\Bigr) \\ &\ge P_x(X_m = y) = p^{(m)}(x,y), \qquad \text{for every } m \ge 1.\end{aligned} \tag{2.47}$$

An upper bound comes by subadditivity (1.4):

$$\rho_{xy} = P_x\Bigl(\bigcup_{n\ge1}\{X_n = y\}\Bigr) \le \sum_{n\ge1} P_x(X_n = y) = \sum_{n\ge1} p^{(n)}(x,y). \tag{2.48}$$

A logical equivalence follows from these inequalities:

$$\rho_{xy} > 0 \text{ if and only if } p^{(n)}(x,y) > 0 \text{ for some } n \ge 1. \tag{2.49}$$

Definition 2.29. A state x is **recurrent** if $\rho_{xx} = 1$ and **transient** if $\rho_{xx} < 1$. △

Example 2.30. Start with some examples without randomness.

An absorbing state x satisfies by definition $p(x,x) = 1$. It is recurrent.

The *2-periodic deterministic chain* has state space $\mathcal{S} = \{0,1\}$ and transition probabilities $p(0,1) = p(1,0) = 1$. This Markov chain simply flips forever between states 0 and 1 without any randomness. Both states are recurrent since the process is sure to return after two time units.

The *deterministically monotone Markov chain* has state space equal to the set $\mathbb{Z}$ of all integers or the set $\mathbb{Z}_{\ge0}$ of nonnegative integers. The transition probability is $p(x,x+1) = 1$ for all states x. Every state is transient because each state is left behind and never visited again. △

Example 2.31. Suppose a state x satisfies $p(x,x) < 1$ and $p(y,x) = 0$ for all states $y \ne x$. Then the process started at x will eventually leave x and never return.

To prove this rigorously, we calculate the probability $P_x(T_x < \infty)$ by summing over all the mutually exclusive possibilities $P_x(T_x = n)$ as n ranges from 1 through all the positive integers. The situation is simple: since the only way to return from x to x is by immediately jumping back to x, $P_x(T_x = 1) = p(x,x)$ and $P_x(T_x = n) = 0$ for $n \ge 2$.

$$\rho_{xx} = P_x(T_x < \infty) = \sum_{n=1}^{\infty} P_x(T_x = n) = p(x,x) < 1.$$

Thus x is transient. △

Example 2.32. This example illustrates the use of the structure of the Markov chain to calculate ρ_{xy} quantities. Let the state space be $\mathcal{S} = \{1,2,3\}$ and the transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} a & 1-a & 0 \\ c & d & 1-c-d \\ 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Assume that $0 < a < 1$, $c > 0$, $d > 0$ and $c + d < 1$ so that only the transition probabilities marked as zero vanish.

To calculate ρ_{11}, observe that immediate return to 1 happens with probability a, while return in $k \geq 2$ steps means a jump to 2, followed by $k-2$ jumps from 2 to 2, and finally a jump to 1.

$$\rho_{11} = \sum_{k=1}^{\infty} P_1(T_1 = k) = a + \sum_{k=2}^{\infty} (1-a)d^{k-2}c = a + \frac{(1-a)c}{1-d}.$$

Using similar reasoning we find the following probabilities.

$$\rho_{12} = \sum_{k=1}^{\infty} P_1(T_2 = k) = \sum_{k=1}^{\infty} a^{k-1}(1-a) = 1.$$

To go from 2 to 1, the process must remain at 2 for some number of steps and then jump to 1. If a jump to 3 happens, the process is absorbed at 3.

$$\rho_{21} = \sum_{k=1}^{\infty} P_2(T_1 = k) = \sum_{k=1}^{\infty} d^{k-1}c = \frac{c}{1-d}.$$

Similarly, a return from 2 back to 2 is possible only by staying at 2 or by spending some time at 1 and then returning:

$$\rho_{22} = \sum_{k=1}^{\infty} P_2(T_2 = k) = d + \sum_{k=2}^{\infty} ca^{k-2}(1-a) = d + c.$$

The answer above makes sense because it is the probability of not moving from 2 to 3.

Since 3 is absorbing, $\rho_{33} = 1$ and $\rho_{31} = \rho_{32} = 0$. Calculating ρ_{13} and ρ_{23} by going through all the possible strategies can be done with a little more work. But we can derive easily $\rho_{13} = \rho_{23} = 1$ in Example 2.37 below after some more theory.

We conclude from above that 1 and 2 are transient states and 3 is recurrent. △

Example 2.33 (Success run chain)**.** Recall from Example 2.17 that the success run chain has state space $\mathcal{S} = \mathbb{Z}_{\geq 0}$ of nonnegative integers and transition probability

$$p(k, k+1) = \alpha,\ \ p(k,0) = 1-\alpha \text{ for all states } k \geq 0 \tag{2.50}$$

where $0 < \alpha < 1$ is the fixed parameter of the model.

To decide whether state 0 is transient or not, we calculate $P_0(T_0 < \infty)$. $P_0(T_0 = n) = \alpha^{n-1}(1-\alpha)$ because the only way for the first return to 0 to happen in exactly n steps is that there are $n-1$ consecutive successes followed by a failure. Thus, when the process starts at 0, T_0 is a geometric random variable with parameter $1-\alpha$. Such a random variable is finite with probability one. Verification amounts to checking that the probability mass function of a geometric random variable sums up to one:

$$\rho_{00} = P_0(T_0 < \infty) = \sum_{n=1}^{\infty} P_0(T_0 = n) = \sum_{n=1}^{\infty} \alpha^{n-1}(1-\alpha) = 1.$$

The last equality above is the evaluation of the geometric series. The conclusion is that 0 is a recurrent state. △

Example 2.34 (Success run chain with varying success probability)**.** To look for more variety of behavior, we let the success probability change with the length of the success run. For some sequence of constants $\{\alpha_k\}_{k\ge 0}$ with values in $[0,1]$ define the transition probability as

$$p(k,k+1)=\alpha_k,\ \ p(k,0)=1-\alpha_k \text{ for all states } k\ge 0. \tag{2.51}$$

Now α_k is the probability of a success when the length of the current success run is k. For a possible story, imagine free throws executed by a device utilizing machine learning whose accuracy improves as successes mount.

Start the process again at 0 and ask whether it is sure to return. The answer depends on the choice of the constants α_k. In one extreme case each $\alpha_k=1$ so every throw succeeds. Obviously the process never returns to 0. On the other hand, if some $\alpha_k=0$, then the process is sure to return to 0 because it cannot go beyond state $k-1$.

To get a generally valid answer, we calculate again the probability $\rho_{00}=P_0(T_0<\infty)$ by adding up the probabilities

$$P_0(T_0=n)=\alpha_0\cdot\alpha_1\cdots\alpha_{n-2}\cdot(1-\alpha_{n-1})=\prod_{k=0}^{n-2}\alpha_k-\prod_{k=0}^{n-1}\alpha_k$$

as n ranges over positive integers. A telescoping series develops:

$$\begin{aligned}P_0(T_0<\infty)=\sum_{1\le n<\infty}P_0(T_0=n)&=\lim_{m\to\infty}\sum_{n=1}^{m}\Bigl(\prod_{k=0}^{n-2}\alpha_k-\prod_{k=0}^{n-1}\alpha_k\Bigr)\\&=\lim_{m\to\infty}\Bigl(1-\prod_{k=0}^{m-1}\alpha_k\Bigr)=1-\lim_{m\to\infty}\prod_{k=0}^{m}\alpha_k.\end{aligned}$$

The last limit exists because the products form a monotone nonincreasing sequence of nonnegative numbers, due to $0\le\alpha_k\le 1$. We can write the answer in terms of an infinite product as

$$P_0(T_0<\infty)=1-\prod_{k=0}^{\infty}\alpha_k.$$

Thus we get a precise criterion of recurrence: state 0 is recurrent if $\prod_{k=0}^{\infty}\alpha_k=0$, and transient if $\prod_{k=0}^{\infty}\alpha_k>0$.

Both cases can happen even if we avoid α_k values zero and one. The constant success probability case is covered here: if $\alpha_k=\alpha\in(0,1)$ then $\prod_{k=0}^{\infty}\alpha_k=\lim_{m\to\infty}\alpha^m=0$. In contrast, if we let α_k converge to 1 fast enough, we can get the opposite conclusion. Take $\alpha_k=1-2^{-k-1}$. Then

$$\begin{aligned}\lim_{m\to\infty}\prod_{k=0}^{m}\alpha_k=\lim_{m\to\infty}\prod_{k=0}^{m}(1-2^{-k-1})&=\lim_{m\to\infty}e^{\sum_{k=0}^{m}\ln(1-2^{-k-1})}\\&\ge\lim_{m\to\infty}e^{-2\sum_{k=0}^{m}2^{-k-1}}=e^{-2}.\end{aligned}$$

(For the inequality above, use calculus to check that $\ln(1-x)\ge -2x$ for $x\in(0,\frac12]$.) The lower bound above is not the exact value of the infinite product but it is good enough to tell us that $\prod_{k=0}^{\infty}\alpha_k>0$. Hence in this case 0 is transient. △

To appreciate the full significance of recurrence and transience, we need further theoretical development. To record times of successive visits to the state x set $T_x^1 = T_x$ for the first visit after time zero, and then for integers $k \geq 2$,

$$T_x^k = \begin{cases} \inf\{n > T_x^{k-1} : X_n = x\}, & \text{if } T_x^{k-1} < \infty \\ \infty, & \text{if } T_x^{k-1} = \infty. \end{cases} \tag{2.52}$$

In plain English, if there was a $(k-1)$st visit to state x, then T_x^k is the time of the next visit to x after that. If the kth visit to x is the last one, then $T_x^j = \infty$ for all $j > k$.

We calculate the probability that the kth visit takes place. The proof is another application of the strong Markov property.

Theorem 2.35. *For any states $x, y \in \mathcal{S}$ and integers $k \geq 1$,*

$$P_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}. \tag{2.53}$$

Proof. The proof goes by induction on k. The case $k = 1$ is the definition (2.46). Now let $k \geq 1$ and assume that $P_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}$.

$$\begin{aligned} P_x(T_y^{k+1} < \infty) &\overset{(a)}{=} P_x(X_{T_y^k} = y, T_y^k < \infty, T_y^{k+1} < \infty) \\ &\overset{(b)}{=} P_x(X_{T_y^k} = y, T_y^k < \infty)\, P_x\big(T_y^{k+1} < \infty \,\big|\, X_{T_y^k} = y, T_y^k < \infty\big) \\ &\overset{(c)}{=} P_x(T_y^k < \infty)\, P_y(T_y^1 < \infty) \overset{(d)}{=} \rho_{xy}\rho_{yy}^{k-1} \cdot \rho_{yy} = \rho_{xy}\rho_{yy}^k. \end{aligned}$$

Step (a) added superfluous conditions into the event: if T_y^{k+1} is finite, then so must be T_y^k, and then by definition $X_{T_y^k} = y$. Step (b) used the product rule. Step (c) dropped the superfluous condition $X_{T_y^k} = y$ from the first probability and applied the strong Markov property to the second probability. The strong Markov step can be clarified by switching to process variables:

$$\begin{aligned} &P_x\big(T_y^{k+1} < \infty \,\big|\, X_{T_y^k} = y, T_y^k < \infty\big) \\ &= P_x\big\{\text{process } (X_{T_y^k+1}, X_{T_y^k+2}, X_{T_y^k+3}, \dots) \text{ visits } y \,\big|\, X_{T_y^k} = y, T_y^k < \infty\big\} \\ &= P_y\big\{\text{process } (X_1, X_2, X_3, \dots) \text{ visits } y\big\} = P_y(T_y^1 < \infty). \end{aligned}$$

Step (d) applied the induction assumption and the definition of ρ_{yy}.

This completes the induction step. The claim (2.53) now holds for all $k \geq 1$. □

With the aid of the theorem we can prove a striking dichotomy: the process returns to a recurrent state infinitely often, while for a transient state there is always a point in time after which it is never seen again. "Infinitely often" means that the visits never end: there is always a next one. Note though that everything here is random. In general, we cannot know ahead of time when the next visit to a recurrent state occurs, and we cannot know which visit to a transient state is the last one.

Theorem 2.36. *If x is recurrent then*

$$P_x(T_x^k < \infty \textit{ for all } k \geq 1) = 1 \tag{2.54}$$

while if x is transient then

$$P_x(T_x^k < \infty \textit{ for all } k \ge 1) = 0. \tag{2.55}$$

Proof. By de Morgan's law, the complement of the event $\{T_x^k < \infty \text{ for all } k \ge 1\}$ is the union $\bigcup_{k=1}^\infty \{T_x^k = \infty\}$. If x is recurrent, then from $\rho_{xx} = 1$ and by (2.53),

$$P_x(T_x^k = \infty) = 1 - P_x(T_x^k < \infty) = 1 - \rho_{xx}^k = 1 - 1 = 0.$$

By subadditivity (1.4),

$$P\Big(\bigcup_{k=1}^\infty \{T_x^k = \infty\}\Big) \le \sum_{k=1}^\infty P_x(T_x^k = \infty) = 0.$$

This establishes (2.54).

By monotonicity, for any $n \ge 1$,

$$P_x(T_x^k < \infty \text{ for all } k \ge 1) \le P_x(T_x^n < \infty) = \rho_{xx}^n.$$

If x is transient, then $\rho_{xx}^n \to 0$ as $n \to \infty$. The inequality then forces the first probability above to be zero. This establishes (2.55). □

We apply the theorem above to derive further facts of the earlier Example 2.32.

Example 2.37 (Continuation of Example 2.32)**.** We argue indirectly that $\rho_{23} = 1$. Start the process at state 2. We already concluded that 2 is a transient state. Then by (2.55), there is a last visit to 2. That is, with P_2-probability one, there is a finite random time $0 \le R_2 < \infty$ such that $X_{R_2} = 2$ and $X_{R_2+n} \ne 2$ for all $n \ge 1$. X_{R_2+1} cannot equal 1 because the process cannot remain in 1 forever and the only way to leave 1 is via 2. Here is a rigorous derivation. (Needless to say, R_2 is not a stopping time because it looks into the future. So we cannot use the strong Markov property. Hence we decompose according to the value of R_2 and then use the basic Markov property.)

$$\begin{aligned} P_2(X_{R_2+1} = 1) &= \sum_{k=0}^\infty P_2(R_2 = k, X_{R_2+1} = 1) \\ &= \sum_{k=0}^\infty P_2(X_k = 2, X_{k+1} = 1, X_{k+n} \ne 2 \,\forall n \ge 2) \\ &= \sum_{k=0}^\infty p^{(k)}(2,2)p(2,1)(1 - \rho_{12}) = 0. \end{aligned}$$

The last equality comes from $\rho_{12} = 1$ derived in Example 2.32. We conclude that at time $R_2 + 1$ the process has been absorbed in state 3 and consequently $\rho_{23} = 1$.

Exercise 2.7 asks you to prove $\rho_{xz} \ge \rho_{xy}\rho_{yz}$. Thus without further calculation we have $\rho_{13} \ge \rho_{12}\rho_{23} = 1$. △

Introduce the $\mathbb{Z}_{\ge 0} \cup \{\infty\}$-valued random variable N_y that counts the number of visits to state y after time zero:

$$N_y = \sum_{n=1}^\infty I_{\{X_n = y\}}. \tag{2.56}$$

The formula above follows the process X_n through times $n = 1, 2, 3\dots$ and adds a 1 to the sum whenever state y is hit. Alternatively, we can count visits to y by checking how many stopping times T_y^k are finite. This gives the equivalent formula

$$N_y = \sum_{k=1}^{\infty} I_{\{T_y^k < \infty\}}. \tag{2.57}$$

The second version (2.57) leads to tidy formula for the expected number of visits.

$$\begin{aligned} E_x[N_y] &= E_x\Big[\sum_{k=1}^{\infty} I_{\{T_y^k<\infty\}}\Big] = \sum_{k=1}^{\infty} E_x[I_{\{T_y^k<\infty\}}] \\ &= \sum_{k=1}^{\infty} P_x(T_y^k < \infty) \overset{(2.53)}{=} \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1}. \end{aligned}$$

Complete the calculation by observing that (i) if $\rho_{xy} = 0$ then the value is zero and (ii) the geometric series converges if and only if $\rho_{yy} < 1$. We summarize the outcome in this theorem.

Theorem 2.38. *For any states $x, y \in \mathcal{S}$,*

$$E_x[N_y] = \begin{cases} 0, & \text{if } \rho_{xy} = 0 \\ \infty, & \text{if } \rho_{xy} > 0 \text{ and } y \text{ is recurrent} \\ \frac{\rho_{xy}}{1-\rho_{yy}} \in (0,\infty), & \text{if } \rho_{xy} > 0 \text{ and } y \text{ is transient.} \end{cases} \tag{2.58}$$

By using both formulas (2.56) and (2.57) for N_y we establish a criterion for recurrence and transience.

Theorem 2.39. *The state x is recurrent if and only if $\sum_{n=1}^{\infty} p^{(n)}(x,x) = \infty$.*

Proof. From (2.56) we get another representation for the expected number of returns to x:

$$\begin{aligned} E_x[N_x] &= E_x\Big[\sum_{n=1}^{\infty} I_{\{X_n=x\}}\Big] = \sum_{n=1}^{\infty} E_x[I_{\{X_n=x\}}] \\ &= \sum_{n=1}^{\infty} P_x(X_n = x) = \sum_{n=1}^{\infty} p^{(n)}(x,x). \end{aligned}$$

From (2.58) we read that $E_x[N_x] = \infty$ if and only if x is recurrent. Putting this statement together with the calculation of $E_x[N_x]$ above proves the theorem. □

The criterion in Theorem 2.39 has both practical and theoretical use. We apply it to the classic case of the simple random walk.

Example 2.40 (Simple random walk)**.** We investigate whether the state 0 is recurrent or transient for simple random walk. Fix the parameter $0 < p < 1$ and recall the transition probability:

$$p(x, x+1) = p, \quad p(x, x-1) = 1 - p \quad \text{for states } x \in \mathbb{Z}.$$

A return from 0 back to 0 can happen only in an even number of steps. Hence $p^{(k)}(0,0) = 0$ when k is odd. For evens we have

$$p^{(2n)}(0,0) = \binom{2n}{n} p^n(1-p)^n = \frac{(2n)!}{(n!)^2}\, p^n(1-p)^n. \tag{2.59}$$

This comes from the binomial distribution. The ±1-valued steps are independent and a return to 0 happens precisely when the number of $+1$ steps equals the number of -1 steps.

To understand the asymptotic behavior of the transition probabilities $p^{(2n)}(0,0)$, we need to understand how $n!$ behaves as n grows. This comes from Stirling's formula:

$$n! \sim e^{-n} n^n \sqrt{2\pi n}. \tag{2.60}$$

The notation

$$a_n \sim b_n \quad \text{means precisely that} \quad \lim_{n\to\infty} \frac{a_n}{b_n} = 1.$$

Applying Stirling's formula to (2.59) and cancellation give the asymptotics

$$p^{(2n)}(0,0) \sim \frac{e^{-2n}(2n)^{2n}\sqrt{4\pi n}}{e^{-2n} n^{2n} 2\pi n}\, p^n(1-p)^n \sim \frac{1}{\sqrt{\pi n}}\,\bigl(4p(1-p)\bigr)^n. \tag{2.61}$$

The analysis splits into cases that depend on the value p.

Case 1. $p = \frac12$. (2.61) simplifies to

$$p^{(2n)}(0,0) \sim \frac{1}{\sqrt{\pi n}}. \tag{2.62}$$

Recall that $\sum n^{-1/2}$ is a divergent series. This guides the last stage of our reasoning. Relation (2.62) means that

$$\lim_{n\to\infty} \sqrt{\pi n}\, p^{(2n)}(0,0) = 1.$$

We may pick n_0 so that $\sqrt{\pi n}\, p^{(2n)}(0,0) \ge \frac12$ for $n \ge n_0$. Bound the series of transition probabilities:

$$\sum_{k=1}^{\infty} p^{(k)}(0,0) = \sum_{n=1}^{\infty} p^{(2n)}(0,0) \ge \sum_{n=n_0}^{\infty} p^{(2n)}(0,0) \ge \sum_{n=n_0}^{\infty} \frac{1}{2\sqrt{\pi n}} = \infty.$$

Theorem 2.39 tells us that 0 is a recurrent state.

Case 2. $p \ne \frac12$. Calculus tells us that on the unit interval $[0,1]$ the unique maximum of the function $f(x) = x(1-x)$ is $f(\frac12) = \frac14$. Thus $p \ne \frac12$ implies that $4p(1-p) < 1$. In this case the terms in (2.61) are those of a convergent geometric series, together with the factor $1/\sqrt{\pi n}$ which cannot hurt the convergence. To prove the convergence of the series of transition probabilities, begin with the limit implicit in (2.61):

$$\lim_{n\to\infty} \frac{\sqrt{\pi n}\, p^{(2n)}(0,0)}{(4p(1-p))^n} = 1.$$

This time choose n_0 so that

$$\frac{\sqrt{\pi n}\, p^{(2n)}(0,0)}{(4p(1-p))^n} \le 2 \qquad \text{for } n \ge n_0.$$

This implies that

$$p^{(2n)}(0,0) \le 2\frac{(4p(1-p))^n}{\sqrt{\pi n}} \le 2(4p(1-p))^n \quad \text{for } n \ge n_0.$$

Bound the series of transition probabilities:

$$\begin{aligned}\sum_{k=1}^{\infty} p^{(k)}(0,0) &= \sum_{n=1}^{\infty} p^{(2n)}(0,0) = \sum_{n=1}^{n_0} p^{(2n)}(0,0) + \sum_{n=n_0+1}^{\infty} p^{(2n)}(0,0) \\ &\le \sum_{n=1}^{n_0} p^{(2n)}(0,0) + 2\sum_{n=n_0+1}^{\infty} (4p(1-p))^n < \infty.\end{aligned}$$

Since our only concern is whether the series $\sum p^{(k)}(0,0)$ converges or not, we do not need to evaluate the last line above any further. The first term is finite because it is a finite sum. The second term is a convergent geometric series because $4p(1-p) < 1$. Thus the entire line equals some finite real number. Theorem 2.39 tells us that 0 is a transient state. △

The process does not leave the set of recurrent states. A key point for understanding the structure of Markov chains is that the process *cannot* go from a recurrent state to a transient state. The next two lemmas develop this fact.

Lemma 2.41. *Let x, y be two distinct states. Suppose $\rho_{xy} > 0$ and $\rho_{yx} < 1$. Then x is transient.*

Proof. The message of the lemma is intuitively clear: if we can go from x to y but there is some chance of never returning from y to x, then x must be transient.

To write a proof, let m be the smallest integer such that $p^{(m)}(x,y) > 0$. Then there exists a sequence of intermediate states $y_1, \dots, y_{m-1}$ that can take us from x to y in m steps, in other words, such that the product of transition probabilities is strictly positive:

$$p(x,y_1)\,p(y_1,y_2)\cdots p(y_{m-1},y) > 0.$$

Since there is no shorter path from x to y, all y_i are distinct from x.

We bound from below the probability $P_x(T_x = \infty)$ by insisting that the process follow the path $y_1, \dots, y_{m-1}, y$. Then condition and apply the Markov property:

$$\begin{aligned}&P_x(T_x = \infty) \\ &\ge P_x(X_1 = y_1, \dots, X_{m-1} = y_{m-1}, X_m = y, X_n \ne x \text{ for all } n > m) \\ &= P_x(X_1 = y_1, \dots, X_{m-1} = y_{m-1}, X_m = y) \\ &\qquad \cdot P_x(X_n \ne x \text{ for all } n > m \,|\, X_1 = y_1, \dots, X_{m-1} = y_{m-1}, X_m = y) \\ &= p(x,y_1)\,p(y_1,y_2)\cdots p(y_{m-1},y)\,P_y(X_n \ne x \text{ for all } n > 0) \\ &= p(x,y_1)\,p(y_1,y_2)\cdots p(y_{m-1},y)\,(1-\rho_{yx}) > 0.\end{aligned}$$

The last quantity is strictly positive because it is a product of strictly positive numbers. Thus x is transient. □

Lemma 2.42. *Let x, y be two distinct states. Suppose x is recurrent and $\rho_{xy} > 0$. Then $\rho_{yx} = 1$ and y is recurrent.*

Proof. Lemma 2.41 forces $\rho_{yx} = 1$. Now that both ρ_{xy} and ρ_{yx} are positive, we can pick integers k, ℓ such that $p^{(k)}(y,x) > 0$ and $p^{(\ell)}(x,y) > 0$. Deduce an inequality that lets us apply Theorem 2.39:

$$\begin{aligned}\sum_{n=1}^{\infty} p^{(n)}(y,y) &\ge \sum_{n=1}^{\infty} p^{(k+n+\ell)}(y,y) \ge \sum_{n=1}^{\infty} p^{(k)}(y,x)\, p^{(n)}(x,x)\, p^{(\ell)}(x,y) \\ &= p^{(k)}(y,x)\Bigl(\sum_{n=1}^{\infty} p^{(n)}(x,x)\Bigr) p^{(\ell)}(x,y) = \infty.\end{aligned}$$

The first inequality above comes from dropping terms $p^{(n)}(y,y)$ for $j \le k+\ell$. The last equality follows (i) because x is assumed recurrent and (ii) because $c \cdot \infty = \infty$ for any positive number c. The divergence of $\sum_{n=1}^{\infty} p^{(n)}(y,y)$ gives the recurrence of y. □

Notice that the lemma implies a conclusion about the entire state space $\mathcal{S}$. For suppose $\rho_{xy} > 0$ for *all* states x and y. Then either all states are recurrent or all states are transient. The argument goes like this. Suppose there is a recurrent state x. Since every other state y satisfies $\rho_{xy} > 0$, Lemma 2.42 forces every other state y to be recurrent too. Thus a mixture of recurrent and transient is impossible.

In this situation we say that *the entire Markov chain is recurrent or transient.* But note that this makes sense only under the assumption that $\rho_{xy} > 0$ for all states x and y.

Example 2.43. By combining the observation above with earlier examples, we get the following statements.

(i) The success run chain with constant success probability $\alpha \in (0,1)$ is a recurrent Markov chain (Example 2.33).

(ii) The success run chain with varying success probabilities $\alpha_k \in (0,1)$ is recurrent if $\prod_{k=0}^{\infty} \alpha_k = 0$ and transient if $\prod_{k=0}^{\infty} \alpha_k > 0$ (Example 2.34).

(iii) Symmetric simple random walk on the one-dimensional integer lattice $\mathbb{Z}$ is recurrent (Example 2.40).

(iv) Asymmetric simple random walk on the one-dimensional integer lattice $\mathbb{Z}$ is transient (Example 2.40).

△

Canonical decomposition of the state space. We need a few more definitions for describing the overall structure of the state space.

Definition 2.44. Let x and y be two states. We say that y *is accessible from* x (abbreviated $x \longrightarrow y$) if $p^{(n)}(x,y) > 0$ for some $n \ge 0$. If both $x \longrightarrow y$ and $y \longrightarrow x$ then x and y *communicate*, abbreviated $x \longleftrightarrow y$. △

Note the distinction of the definition above with (2.49): $n \ge 0$ versus $n \ge 1$. If x and y are distinct states, then $x \longrightarrow y$ is equivalent to $\rho_{xy} > 0$. But ρ_{xx} can be zero while $x \longrightarrow x$ holds trivially due to $p^{(0)}(x,x) = 1$. The Chapman-Kolmogorov equations (2.32) imply that

$$\text{if } x \longrightarrow y \text{ and } y \longrightarrow z, \text{ then } x \longrightarrow z. \tag{2.63}$$

Exercise 2.7 asks you to prove this.

Communication is an *equivalence relation*. This means that the following three properties are satisfied:

- reflexivity: $x \longleftrightarrow x$
- symmetry: $x \longleftrightarrow y$ if and only if $y \longleftrightarrow x$
- transitivity: if $x \longleftrightarrow y$ and $y \longleftrightarrow z$, then $x \longleftrightarrow z$.

Definition 2.45. Let $\mathcal{A}$ and $\mathcal{B}$ be subsets of the state space $\mathcal{S}$. $\mathcal{A}$ is **irreducible** if $x \longleftrightarrow y$ for all $x, y \in \mathcal{A}$. $\mathcal{B}$ is **closed** if $\rho_{xy} = 0$ whenever $x \in \mathcal{B}$ and $y \notin \mathcal{B}$.

To paraphrase these definitions: All elements of an irreducible set communicate with each other. Since $x \longleftrightarrow x$ by definition, each singleton $\{x\}$ is an irreducible set. A closed set has the property that if the process starts there, it remains there for all time.

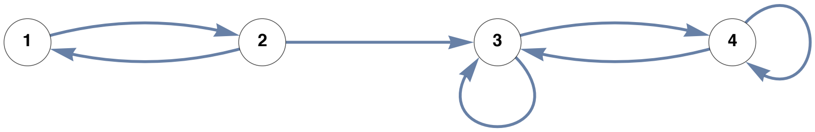

Figure 2. Transition diagram for the Markov chain of Example 2.46.

Example 2.46. Let $\mathcal{S} = \{1, 2, 3, 4\}$ and let the transition matrix be

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & x & 0 & 0 \\ x & 0 & x & 0 \\ 0 & 0 & x & x \\ 0 & 0 & x & x \end{bmatrix} \end{array}$$

where x represents a nonzero entry. Figure 2 illustrates this Markov chain.

The irreducible sets are the singletons $\{1\}$, $\{2\}$, $\{3\}$ and $\{4\}$, and also $\{1, 2\}$ and $\{3, 4\}$. The closed sets are $\{3, 4\}$ and $\mathcal{S}$. △

Define for each recurrent state x its *communicating class* $\mathcal{R}_x$ by

$$\mathcal{R}_x = \{y \in \mathcal{S} : x \longrightarrow y\}. \tag{2.64}$$

Lemma 2.47. *For each recurrent state x, $\mathcal{R}_x$ is a closed, irreducible set of recurrent states. For two recurrent states x and z, either $\mathcal{R}_x \cap \mathcal{R}_z = \varnothing$ or $\mathcal{R}_x = \mathcal{R}_z$.*

Proof. Lemma 2.42 tells us that all states in $\mathcal{R}_x$ communicate with x and are recurrent. Consequently all states in $\mathcal{R}_x$ communicate with each other. Thus $\mathcal{R}_x$ is an irreducible set of recurrent states.

Suppose $\rho_{yw} > 0$ for a state $y \in \mathcal{R}_x$ and another arbitrary state w. Then $x \longrightarrow y$ and $y \longrightarrow w$ which imply $x \longrightarrow w$, and so also $w \in \mathcal{R}_x$. In other words, the process cannot escape from the set $\mathcal{R}_x$ and thereby $\mathcal{R}_x$ is closed.

Take two distinct recurrent states x and z. Suppose $\mathcal{R}_x$ and $\mathcal{R}_z$ are not disjoint. Let $y \in \mathcal{R}_x \cap \mathcal{R}_z$. Then y communicates with both x and z. If w is any state in $\mathcal{R}_z$, the chain $x \longleftrightarrow y$, $y \longleftrightarrow z$ and $z \longleftrightarrow w$ implies $x \longleftrightarrow w$. Thus every state in $\mathcal{R}_z$ lies in $\mathcal{R}_x$, in other words $\mathcal{R}_z \subset \mathcal{R}_x$. The same argument gives the opposite inclusion $\mathcal{R}_x \subset \mathcal{R}_z$. Thus either $\mathcal{R}_x \cap \mathcal{R}_z = \varnothing$ or $\mathcal{R}_x = \mathcal{R}_z$. □

Theorem 2.48. (Canonical decomposition of the state space.) *The state space $\mathcal{S}$ of any Markov chain can be decomposed as*

$$\mathcal{S} = \mathcal{T} \cup \Big(\bigcup_i \mathcal{R}_i \Big)$$

where $\mathcal{T}$ is the set of all transient states and $\{\mathcal{R}_i\}$ is an at most countable collection of pairwise disjoint, closed, irreducible sets of recurrent states.

Proof. All the work to prove this theorem was basically done above. Put all the transient states in $\mathcal{T}$. Let $\{\mathcal{R}_i\}$ be the collection of communicating classes $\mathcal{R}_x$ of recurrent states x where each set $\mathcal{R}_x$ is listed exactly once. □

Remark 2.49 (Equivalence classes)**.** It is a general fact that an equivalence relation partitions the underlying set into mutually disjoint *equivalence classes.* Consider the communication relation $\longleftrightarrow$ on the set $\mathcal{R}$ of all recurrent states. The equivalence class of the state x is by definition $[x] = \{y \in \mathcal{R} : x \longleftrightarrow y\}$. Lemma 2.47 implies that the decomposition $\mathcal{R} = \bigcup_i \mathcal{R}_i$ is exactly the decomposition of $\mathcal{R}$ into the equivalence classes of the relation $\longleftrightarrow$.

Whether the set $\mathcal{T}$ of transient states is a single equivalence class or a union of several equivalence classes depends on the particular Markov chain. △

Remark 2.50 (Canonical form of the transition matrix)**.** The canonical decomposition of the state space gives rise to a matching canonical form of the transition matrix that illuminates the structure of the Markov chain.

To illustrate, suppose the state space consists of three closed irreducible recurrent classes $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$, and a set $\mathcal{T}$ of transient states. Order the rows and columns of the transition matrix $\mathbf{P}$ so that the states of $\mathcal{R}_1$ comes first, then those of $\mathcal{R}_2$ followed by $\mathcal{R}_3$, and the transient states last. Then the transition matrix acquires this block form:

$$\mathbf{P} = \begin{array}{c} \\ \mathcal{R}_1 \\ \mathcal{R}_2 \\ \mathcal{R}_3 \\ \mathcal{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 & \mathcal{T} \end{array} \\ \begin{bmatrix} \mathbf{P}_{\mathcal{R}_1} & 0 & 0 & 0 \\ 0 & \mathbf{P}_{\mathcal{R}_2} & 0 & 0 \\ 0 & 0 & \mathbf{P}_{\mathcal{R}_3} & 0 \\ S_1 & S_2 & S_3 & Q \end{bmatrix} \end{array} \tag{2.65}$$

The entries in the matrix above are not numbers but instead blocks of the appropriate dimensions. Example 2.51 below derives the canonical form of the transition matrix of gambler's ruin.

For $i \in \{1, 2, 3\}$, $\mathbf{P}_{\mathcal{R}_i}$ is the restriction of $\mathbf{P}$ to the states of the set $\mathcal{R}_i$. The key observation is that *each $\mathbf{P}_{\mathcal{R}_i}$ is a transition probability matrix* in its own right. This is really the content of the statement that $\mathcal{R}_i$ is a closed set. Another way to put this same idea is that each $\mathbf{P}_{\mathcal{R}_i}$ is a smaller irreducible recurrent Markov chain "inside" the larger $\mathbf{P}$.

About the blocks S_i and Q nothing very specific can be said in general. Q can be a transition matrix, the zero matrix, or anything in between. (The reader should construct small examples that illustrate different possibilities.) Section 2.4 shows how the blocks S_i and Q can be used to calculate information about how quickly and where a finite Markov chain enters one of the recurrent classes $\mathcal{R}_i$. △

Example 2.51 (Continuation of Example 2.16)**.** Consider gambler's ruin with state space $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ and transition matrix $\mathbf{P}$ given in (2.29). The absorbing states form recurrent classes $\{0\}$ and $\{5\}$ by themselves. From each state $x \in \{1, 2, 3, 4\}$ we can march straight to 0 by losing x times in a row, and so

$$P_x(T_x = \infty) \geq (1-p)^x > 0.$$

Thus the set of transient states is $\mathcal{T} = \{1, 2, 3, 4\}$. A canonical form for the transition matrix is

$$\mathbf{P} = \begin{array}{c|c|c|cccc} & 0 & 5 & 1 & 2 & 3 & 4 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 5 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 1 & 1-p & 0 & 0 & p & 0 & 0 \\ 2 & 0 & 0 & 1-p & 0 & p & 0 \\ 3 & 0 & 0 & 0 & 1-p & 0 & p \\ 4 & 0 & p & 0 & 0 & 1-p & 0 \end{array}. \tag{2.66}$$

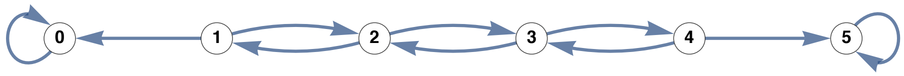

Figure 3. Transition diagram for the Markov chain of Example 2.51.

As a *matrix* this $\mathbf{P}$ is not the same as the originally given transition matrix in (2.29). They differ by a permutation of the rows and columns. But both are perfectly good representations of the same Markov chain. △

The next lemma can be helpful in identifying recurrent classes.

Lemma 2.52.

(a) *A finite closed subset of the state space has at least one recurrent state.*

(b) *In a finite closed irreducible subset of the state space all states are recurrent.*

Proof. Part (a). Let $\mathcal{A} \subset \mathcal{S}$ be a finite, closed subset of the state space. Pick a state $x \in \mathcal{A}$. By closedness, $\sum_{y \in \mathcal{A}} p^{(n)}(x, y) = 1$ for each $n \geq 1$. Thus

$$\infty = \sum_{n=1}^{\infty} 1 = \sum_{n=1}^{\infty} \sum_{y \in \mathcal{A}} p^{(n)}(x, y) = \sum_{y \in \mathcal{A}} \sum_{n=1}^{\infty} p^{(n)}(x, y) = \sum_{y \in \mathcal{A}} E_x[N_y].$$

Since $\mathcal{A}$ is finite, the last sum has finitely many terms. Hence for the last sum to be infinite, there must exist $y \in \mathcal{A}$ such that $E_x[N_y] = \infty$. By (2.58) this y must be recurrent.

Part (b). Part (a) guarantees at least one recurrent state in the set. Then by Lemma 2.42, irreducibility implies that all states in the set are recurrent. □

Example 2.53. Consider the Markov chain of Example 2.46. Find the canonical decomposition.

The set $\{3, 4\}$ is finite, closed and irreducible, hence a recurrent class (Lemma 2.52(b)). Since $p(2, 3) > 0$ but $\rho_{32} = 0$, 2 is transient (Lemma 2.41). Since $p(1, 2) > 0$ but $\rho_{21} < 1$ (from 2 the process can jump to 3 and then never see 1 again), 1 is transient.

We conclude that $\mathcal{R} = \{3, 4\}$ is a closed irreducible recurrent class and the set of transient states is $\mathcal{T} = \{1, 2\}$. A canonical form of the transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 3 \\ 4 \\ 1 \\ 2 \end{array} \!\!\begin{array}{c} \begin{array}{cccc} 3 & 4 & 1 & 2 \end{array} \\ \left[\begin{array}{cc|cc} x & x & 0 & 0 \\ x & x & 0 & 0 \\ \hline 0 & 0 & 0 & x \\ x & 0 & x & 0 \end{array}\right] \end{array}.$$

△

As mentioned before Example 2.43, a Markov chain is *recurrent* if its state space is one single irreducible set of recurrent states. A *finite Markov chain* is one whose state space is finite. Part (b) of the lemma above implies the fundamental fact that

(2.67) *every finite irreducible Markov chain is recurrent.*

Since every property of a Markov chain is really a property of its transition matrix $\mathbf{P}$, we also talk about $\mathbf{P}$ being irreducible and recurrent or transient.

We close this section with a lemma that is useful on several occasions in the sequel. In plain words, it says that no matter how an irreducible and recurrent Markov chain is started, every state will be visited eventually. And then the visits never stop. The reader should construct examples where the conclusion fails if either hypothesis is dropped.

Lemma 2.54. *Let $\mathbf{P}$ be irreducible and recurrent, μ an arbitrary initial distribution, and $x \in \mathcal{S}$ an arbitrary state. Then $P_\mu(T_x < \infty) = 1$. Furthermore,*

$$P_\mu(T_x^k < \infty \text{ for all } k \geq 1) = 1.$$

Equivalently, $P_\mu(N_x = \infty) = 1$.

Proof. The proof is a small extension of that of (2.54) in Theorem 2.36. The assumption of recurrence gives $\rho_{xx} = 1$. Then for any state $z \neq x$, $\rho_{zx} = 1$ comes from Lemma 2.42 because $x \longrightarrow z$ is assumed.

By identity (2.28) and by Theorem 2.35,

$$P_\mu(T_x^k < \infty) = \sum_z \mu(z)\, P_z(T_x^k < \infty) = \sum_z \mu(z)\, \rho_{zx}\rho_{xx}^{k-1} = \sum_z \mu(z) = 1.$$

A countable intersection of probability one events has probability one. This can be seen for example by deMorgan's law and subadditivity:

$$\begin{aligned} P_\mu(T_x^k < \infty \text{ for all } k \ge 1) &= P_\mu\Bigl(\bigcap_{k=1}^{\infty} \{T_x^k < \infty\} \Bigr) \\ &= 1 - P_\mu\Bigl(\bigcup_{k=1}^{\infty} \{T_x^k = \infty\} \Bigr) \ge 1 - \sum_{k=1}^{\infty} P_\mu(T_x^k = \infty) \\ &= 1 - \sum_{k=1}^{\infty} 0 = 1. \end{aligned}$$ □

2.4. Absorption

Start the Markov chain in the set of transient states $\mathcal{T}$. Let $\mathcal{R} = \mathcal{S} \setminus \mathcal{T}$ denote the set of recurrent states and let

$$T = \inf\{n \ge 0 : X_n \in \mathcal{R}\}$$

denote the stopping time after which the process remains $\mathcal{R}$. Assume given an initial state $x \in \mathcal{T}$. This section develops a technique for answering questions such as the following. What is the probability $P_x(X_T = y)$ that the process enters state $y \in \mathcal{R}$ as it leaves the transient set $\mathcal{T}$? What is $E_x[T]$, the expected time it takes to leave the transient set?

Example 2.22 answered exactly this type of question for gambler's ruin. The example used the method of *first-step analysis*, which proceeds by conditioning on the first step of the Markov chain. We develop that approach here more systematically with linear algebra.

In this section we assume that the state space $\mathcal{S}$ is finite, except in an example at the end. The finiteness of the state space guarantees that $E_x[T]$ is finite for all states x. Exercise 2.15 takes you step by step through the proof. The finite expectation $E_x[T] < \infty$ implies that $P_x(T < \infty) = 1$.

Decompose the transition matrix into blocks as follows.

$$\mathbf{P} = \begin{array}{c c} & \begin{array}{cc} \mathcal{R} & \mathcal{T} \end{array} \\ \begin{array}{c} \mathcal{R} \\ \mathcal{T} \end{array} & \left[\begin{array}{cc} \mathbf{P}_{\mathcal{R}} & 0 \\ S & Q \end{array} \right] \end{array} \tag{2.68}$$

$\mathbf{P}_{\mathcal{R}}$ is the $\mathcal{R} \times \mathcal{R}$ transition probability matrix for the process that lives on the state space $\mathcal{R}$. The $(\mathcal{R}, \mathcal{T})$-submatrix is zero because there are no transitions from $\mathcal{R}$ to $\mathcal{T}$. $S = \{s_{xy}\}_{x \in \mathcal{T}, y \in \mathcal{R}}$ is the submatrix of transition probabilities from $\mathcal{T}$ into $\mathcal{R}$, while $Q = \{q_{xy}\}_{x,y \in \mathcal{T}}$ is the $\mathcal{T} \times \mathcal{T}$ submatrix of transition probabilities within $\mathcal{T}$. In the finite case Q cannot be by itself a transition probability matrix because there must be some transition from $\mathcal{T}$ to $\mathcal{R}$.

Matrix multiplication shows that powers of $\mathbf{P}$ retain this same form:

$$\mathbf{P}^n = \begin{bmatrix} \mathbf{P}_{\mathcal{R}}^n & 0 \\ S_n & Q^n \end{bmatrix} \qquad \text{for } n \ge 2. \tag{2.69}$$

Above $\mathbf{P}_{\mathcal{R}}^n$ and Q^n are the powers of the blocks, but S_n is not. (Exercise 2.12 asks you to derive the equation for S_n. But we do not need it below.)

Define the $\mathcal{T} \times \mathcal{R}$ matrix $U = \{u_{xy}\}_{x\in\mathcal{T},\, y\in\mathcal{R}}$ of entrance probabilities by

$$u_{xy} = P_x(X_T = y).$$

We express U in terms of the block structure of $\mathbf{P}$. We split the probability u_{xy} according to whether entry into y happens in step one or whether the first step goes into a transient state z.

$$\begin{aligned} u_{xy} = P_x(X_T = y) &= p(x,y) + \sum_{z\in\mathcal{T}} P_x(X_1 = z, X_T = y) \\ &= p(x,y) + \sum_{z\in\mathcal{T}} p(x,z) P_z(X_T = y) \\ &= s_{xy} + \sum_{z\in\mathcal{T}} q_{xz} u_{zy}. \end{aligned} \tag{2.70}$$

In the Markovian step above the process restarts in state $z \in \mathcal{T}$ and then arrives in $\mathcal{R}$ at state y. To make the application of the Markov property explicit, we sum over the possible values of T and then switch to X_k-variables. Note that $x \in \mathcal{T}$ and the conditioning on $X_1 = z \in \mathcal{T}$ force $T \geq 2$.

$$\begin{aligned} &P_x(X_1 = z, X_T = y) = P_x(X_1 = z)\, P_x(X_T = y \,|\, X_1 = z) \\ &\quad = p(x,z) \sum_{n=2}^{\infty} P_x(T = n, X_T = y \,|\, X_1 = z) \\ &\quad = p(x,z) \sum_{n=2}^{\infty} P_x(X_2 \in \mathcal{T}, \dots, X_{n-1} \in \mathcal{T}, X_n = y \,|\, X_1 = z) \\ &\quad = p(x,z) \sum_{n=2}^{\infty} P_z(X_1 \in \mathcal{T}, \dots, X_{n-2} \in \mathcal{T}, X_{n-1} = y) \\ &\quad = p(x,z) \sum_{n=2}^{\infty} P_z(T = n-1, X_T = y) \\ &\quad = p(x,z) P_z(X_T = y). \end{aligned}$$

The last step in (2.70) replaced $p(x,y)$ with s_{xy} and $p(x,z)$ with q_{xz} because $y \in \mathcal{R}$ and $x, z \in \mathcal{T}$. In terms of matrices, (2.70) becomes

$$U = S + QU \iff U - QU = S \iff (I - Q)U = S.$$

Thus if we can invert $I - Q$ we can calculate U from

$$U = (I - Q)^{-1} S. \tag{2.71}$$

The lemma below validates this attempt.

Below we utilize a series whose terms are matrices. This is defined entry by entry, in other words, the (x,y)-entry of the matrix $\sum_{n=0}^{\infty} Q^n$ is given by $(\sum_{n=0}^{\infty} Q^n)_{xy} = \sum_{n=0}^{\infty} (Q^n)_{xy}$, as long as the second series of matrix entries converges in the usual sense of a series of real numbers.

Lemma 2.55. *Consider a finite Markov chain with a nonempty set $\mathcal{T}$ of transient states and let $q_{xy} = p(x,y)$ for transient states x, y.*

(a) *For all $x, y \in \mathcal{T}$, the series $\sum_{n=0}^{\infty} (Q^n)_{xy}$ converges.*

(b) *The matrix $I - Q$ is invertible and $(I-Q)^{-1} = \sum_{n=0}^{\infty} Q^n$.*

Proof. First we show that $(Q^n)_{xy} = p^{(n)}(x,y)$ for $x, y \in \mathcal{T}$. This is in fact already contained in (2.69), but here is a probabilistic derivation. Since y is a transient state, we cannot have $X_n = y$ unless all states up to time n are transient, and hence

$$\begin{aligned} p^{(n)}(x,y) = P_x(X_n = y) &= P_x(X_1 \in \mathcal{T}, \dots, X_{n-1} \in \mathcal{T}, X_n = y) \\ &= \sum_{x_1,\dots,x_{n-1} \in \mathcal{T}} p(x,x_1) \cdots p(x_{n-1},y) \\ &= \sum_{x_1,\dots,x_{n-1} \in \mathcal{T}} q_{x,x_1} \cdots q_{x_{n-1},y} = (Q^n)_{xy}. \end{aligned}$$

The convergence of the series follows from (2.58): for transient states x, y,

$$\begin{aligned} \sum_{n=0}^{\infty} (Q^n)_{xy} = \sum_{n=0}^{\infty} p^{(n)}(x,y) &= I_{\{x=y\}} + \sum_{n=1}^{\infty} p^{(n)}(x,y) \\ &= I_{\{x=y\}} + E_x[N_y] = I_{\{x=y\}} + \frac{\rho_{xy}}{1-\rho_{yy}}. \end{aligned}$$

The matrix $\sum_{n=0}^{\infty} Q^n$ is now well-defined. To show that $\sum_{n=0}^{\infty} Q^n$ and $I - Q$ are inverses of each other, we show that their product is the identity matrix I. The next calculation relies on the fact that a convergent matrix series can be treated exactly like a convergent series of real numbers: multiplication can be taken inside the series and terms can be separated from it.

$$\begin{aligned} (I-Q)\sum_{k=0}^{\infty} Q^k = \sum_{k=0}^{\infty} Q^k - Q\sum_{k=0}^{\infty} Q^k &= \sum_{k=0}^{\infty} Q^k - \sum_{k=0}^{\infty} Q^{k+1} \\ &= I + \sum_{k=1}^{\infty} Q^k - \sum_{k=1}^{\infty} Q^k = I. \end{aligned}$$

The other identity $(\sum_{n=0}^{\infty} Q^n)(I-Q) = I$ can be deduced similarly, although it is enough to check just one of the products (Lemma C.1 in Appendix C). □

Before examples, we use the same technique to find the expectations $m(x) = E_x[T]$ of the time till absorption in $\mathcal{R}$ from an initial state $x \in \mathcal{T}$. As above, first we derive an equation by decomposing according to the first step. Note below that if $X_1 \in \mathcal{R}$ then $T = 1$.

$$\begin{aligned}
m(x) = E_x[T] &= \sum_{n=1}^{\infty} n P_x(T = n) \\
&= \sum_{z\in\mathcal{R}} \sum_{n=1}^{\infty} n P_x(X_1 = z, T = n) + \sum_{y\in\mathcal{T}} \sum_{n=1}^{\infty} n P_x(X_1 = y, T = n) \\
&= \sum_{z\in\mathcal{R}} P_x(X_1 = z) + \sum_{y\in\mathcal{T}} P_x(X_1 = y) \sum_{n=1}^{\infty} n P_x(T = n \,|\, X_1 = y) \\
&\overset{(a)}{=} \sum_{z\in\mathcal{R}} p(x, z) + \sum_{y\in\mathcal{T}} p(x, y) \sum_{n=1}^{\infty} n\, P_y(T = n - 1) \\
&= \sum_{z\in\mathcal{R}} p(x, z) + \sum_{y\in\mathcal{T}} p(x, y) + \sum_{y\in\mathcal{T}} p(x, y) \sum_{n=1}^{\infty} (n - 1)\, P_y(T = n - 1) \\
&= 1 + \sum_{y\in\mathcal{T}} q_{xy}\, m(y).
\end{aligned}$$

In the application of the Markov property in step (a) above, once the process restarts after the first step, the entrance time T is $n-1$ time steps away. This reasoning can be made more explicit as follows. Since both $x, y \in \mathcal{T}$, the conditioning $X_1 = y$ implies that $T \geq 2$ and hence the condition $X_1 \in \mathcal{T}$ can be dropped.

$$\begin{aligned}
P_x(T = n \,|\, X_1 = y) &= P_x(X_1 \in \mathcal{T}, \dots, X_{n-1} \in \mathcal{T}, X_n \in \mathcal{R} \,|\, X_1 = y) \\
&= P_x(X_2 \in \mathcal{T}, \dots, X_{n-1} \in \mathcal{T}, X_n \in \mathcal{R} \,|\, X_1 = y) \\
&= P_y(X_1 \in \mathcal{T}, \dots, X_{n-2} \in \mathcal{T}, X_{n-1} \in \mathcal{R}) \\
&= P_y(T = n - 1).
\end{aligned}$$

We put the equation in matrix form. Let $\mathbf{m} = \{m(x)\}_{x\in\mathcal{T}}$ be the column vector of the $m(x)$-values and $\mathbf{1}$ a column vector of ones. Then from above we have $\mathbf{m} = \mathbf{1} + Q\mathbf{m}$ from which we solve for $\mathbf{m}$ by

$$\mathbf{m} = (I - Q)^{-1}\mathbf{1}. \tag{2.72}$$

Note that the equation $\mathbf{m} = \mathbf{1} + Q\mathbf{m}$ may well admit a solution where some or all values $m(x)$ are infinite. Thus we have to know ahead of time that $m(x) < \infty$. Otherwise we cannot know which solution, the finite one or the infinite one, is the correct value of $E_x[T]$. As mentioned, this is settled in Exercise 2.15.

We illustrate equations (2.71) and (2.72) with gambler's ruin.

Example 2.56. We revisit Example 2.22. Consider gambler's ruin with the state space $\mathcal{S} = \{0, 1, 2, 3, 4\}$. $\mathcal{T} = \{1, 2, 3\}$ and $\mathcal{R} = \{0, 4\}$. The transition matrix decomposes as follows

$$\mathbf{P} = \begin{array}{c|cc|ccc|} & 0 & 4 & 1 & 2 & 3 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 0 \\ \cline{2-6} 1 & 1/2 & 0 & 0 & 1/2 & 0 \\ 2 & 0 & 0 & 1/2 & 0 & 1/2 \\ 3 & 0 & 1/2 & 0 & 1/2 & 0 \end{array}$$

with

$$S = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad \text{and} \qquad Q = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}.$$

A derivation of the inverse gives

$$(I - Q)^{-1} = \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix}$$

Thus, the probabilities of winning and losing from given starting positions are

$$U = \begin{bmatrix} P_1(\text{lose}) & P_1(\text{win}) \\ P_2(\text{lose}) & P_2(\text{win}) \\ P_3(\text{lose}) & P_3(\text{win}) \end{bmatrix} = (I - Q)^{-1} S = \begin{bmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}.$$

We recovered the formula $P_x(\text{win}) = x/N$ for the special case $N = 4$ earlier derived in Example 2.22.

The vector of expected durations of the game is

$$\mathbf{m} = \begin{bmatrix} E_1[T] \\ E_2[T] \\ E_3[T] \end{bmatrix} = (I - Q)^{-1}\mathbf{1} = \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix}$$

$\triangle$

The next example shows that certain types of questions about infinite state spaces can also be answered with the approach of this section.

Example 2.57. *Reflected symmetric simple random walk* on the right half line $Z_{\geq 0} = \{0, 1, 2, 3, \dots\}$ operates like symmetric SRW on positive integers. Whenever it lands on 0, it is reflected back to 1 immediately at the next time. Thus the transition probability is

$$p(0,1) = 1, \quad p(x, x-1) = p(x, x+1) = \tfrac{1}{2} \quad \text{for} \quad x \geq 1. \tag{2.73}$$

We calculate $E_0[T_k]$, the expected time to reach state k from 0, for all $k \geq 1$.

Since we follow the process only until the first time it reaches state k, we can modify the transition probability to make k an absorbing state. The new Markov chain has state space $\{0, 1, \dots, k\}$ and transition probability

$$p(0,1) = p(k,k) = 1, \quad p(x, x-1) = p(x, x+1) = \tfrac{1}{2} \quad \text{for} \quad 1 \leq x \leq k-1.$$

We calculate the expectation $E_0[T_k]$ for this Markov chain. It is the *same* as for the original reflected symmetric SRW.

Since the process moves immediately from 0 to 1, $E_0[T_1] = 1$. We can assume $k \geq 2$ for the remaining calculations.

We develop the equations for $m(x) = E_x[T_k]$ as was done above. The first equation is $m(0) = 1 + m(1)$ which we rewrite as

$$m(0) - m(1) = 1. \tag{2.74}$$

If $k \geq 3$, then for $1 \leq x \leq k-2$ we have

$$m(x) = 1 + \tfrac{1}{2}m(x-1) + \tfrac{1}{2}m(x+1)$$

which we turn into

$$m(x) - m(x+1) = 2 + m(x-1) - m(x).$$

Iterate this down to the first increment and apply (2.74):

$$\begin{aligned} m(x) - m(x+1) &= 2 + m(x-1) - m(x) \\ &= 2 + 2 + m(x-2) - m(x-1) \\ &= \cdots = 2x + m(0) - m(1) = 2x + 1. \end{aligned} \tag{2.75}$$

The equation for the last state before k is $m(k-1) = 1 + \frac{1}{2}m(k-2)$. Combine this with (2.75) to find

$$m(k-1) = m(k-2) - m(k-1) + 2 = 2(k-2) + 1 + 2 = 2k - 1.$$

Finally,

$$\begin{aligned} m(0) &= m(k-1) + m(0) - m(k-1) \\ &= m(k-1) + \sum_{x=0}^{k-2} \bigl(m(x) - m(x+1)\bigr) \\ &= 2k - 1 + \sum_{x=0}^{k-2} (2x+1) = 2k - 1 + 2\sum_{x=0}^{k-2} x + k - 1 \\ &= 2k - 1 + (k-1)(k-2) + k - 1 = k^2. \end{aligned}$$

Thus for the reflecting SRW with transitions (2.73), $E_0[T_k] = k^2$ for $k \geq 1$. △

2.5. Invariant distributions

From following the state to following its distribution. The previous section studied the evolution of the Markov chain among the transient states and derived some quantitative information about its entrance into the set of recurrent states. What happens subsequently when the process remains in a closed, irreducible recurrent class for all time? If this class consists of a single absorbing state (as for example in gambler's ruin) nothing more happens: the process stays in this absorbing state. But if the class has more than one state, then the process never stops moving about randomly. Can anything be said about its long term evolution?

A shift in point of view becomes valuable. Instead of following the random state X_n, we follow its probability distribution $\mu^{(n)}$, defined for states $y \in \mathcal{S}$ by $\mu^{(n)}(y) = P_\mu(X_n = y) = (\mu\mathbf{P}^n)_y$ when the initial distribution is $\mu^{(0)} = \mu$. The sequence of probability distributions $\mu^{(0)}, \mu^{(1)}, \mu^{(2)}, \ldots$ is an evolution in the space

$\mathcal{M}$ of probability distributions on $\mathcal{S}$. The mathematical term for this is a *dynamical system*. If the state space $\mathcal{S}$ is finite, say with r elements, then the probability distributions $\mu^{(n)}$ are very concrete objects: they are r-vectors with nonnegative entries that sum to one.

Suppose the distributions converge: $\mu^{(n)}(x) \to \pi(x)$ as $n \to \infty$, for each state x. Can we characterize π? Since $\mu^{(n+1)} = \mu^{(n)}\mathbf{P}$, taking the limit on both sides of this identity gives $\pi = \pi\mathbf{P}$. Thus π is a *fixed point* of right multiplication by the transition matrix $\mathbf{P}$. (When $\mathcal{S}$ is infinite, taking the limit of $(\mu^{(n)}\mathbf{P})_y$ needs a little analysis. Exercise 2.25 points the way.)

In this section we study these fixed points, beginning with a precise definition. In Section 2.6 we see how these fixed points describe the long term evolution of the process.

Invariant distributions and stationary processes. As throughout, $\mathcal{S}$ is a finite or countably infinite state space.

Definition 2.58. Let $\mathbf{P} = \{p(x,y)\}_{x,y\in\mathcal{S}}$ be a transition probability and π a probability measure on $\mathcal{S}$. Then π is an **invariant distribution** for transition matrix $\mathbf{P}$ if $\pi = \pi\mathbf{P}$, or more explicitly,

$$\pi(y) = \sum_{x\in\mathcal{S}} \pi(x)\, p(x,y) \quad \text{for all } y \in \mathcal{S}. \tag{2.76}$$

△

Alternative terms for the invariant distribution are *invariant probability measure* and *stationary distribution*. These terms are used interchangeably.

Recall from (2.33) that if μ is the initial distribution of the Markov chain, then the distribution at time n is given by $P_\mu(X_n = y) = (\mu\mathbf{P}^n)(y)$. Now take an invariant distribution π as the initial distribution. The definition of invariance implies that π stays fixed under multiplication by $\mathbf{P}$ from the right:

$$\pi = \pi\mathbf{P} = (\pi\mathbf{P})\mathbf{P} = \pi\mathbf{P}^2 = \cdots = \pi\mathbf{P}^n \quad \text{for all } n \geq 0. \tag{2.77}$$

As a consequence we get this statement.

Theorem 2.59. *Suppose π is an invariant distribution for the transition probability $\mathbf{P}$ and we take π as the initial distribution of the Markov chain. Then at each time $n \geq 0$ the distribution of the state X_n is π. In other words, $P_\pi(X_n = y) = \pi(y)$ for all $n \geq 0$ and $y \in \mathcal{S}$.*

We illustrate finding an invariant distribution with finite and infinite examples. In each case the task is to solve the linear system $\pi = \pi\mathbf{P}$ for the unknown row vector π and to find a particular solution that qualifies as a probability distribution, that is, satisfies $\pi(x) \geq 0$ for all $x \in \mathcal{S}$ and $\sum_x \pi(x) = 1$.

Example 2.60. Suppose $\mathcal{S} = \{0, 1\}$ and $\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}$. The equation $\pi = \pi\mathbf{P}$ develops as follows:

$$\begin{aligned} \pi = \pi\mathbf{P} \quad &\Longleftrightarrow \quad [\pi(0)\ \pi(1)] = [\pi(0)\ \pi(1)] \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix} \\ &\Longleftrightarrow \quad \begin{cases} \pi(0) = \frac{1}{4}\pi(0) + \frac{2}{3}\pi(1) \\ \pi(1) = \frac{3}{4}\pi(0) + \frac{1}{3}\pi(1) \end{cases} \quad \Longleftrightarrow \quad \pi(1) = \tfrac{9}{8}\pi(0). \end{aligned}$$

The two equations both simplify to the same equation $\pi(1) = \frac{9}{8}\pi(0)$ that has infinitely many solutions. But once we impose the requirement $\pi(0) + \pi(1) = 1$ we get the unique invariant distribution $\pi = \left[\, \frac{8}{17} \ \ \frac{9}{17} \,\right]$. △

Example 2.61 (Success run chain)**.** The state space is $\mathcal{S} = \mathbb{Z}_{\geq 0}$ and the transition probability

$$p(k, k+1) = \alpha, \ \ p(k, 0) = 1 - \alpha \text{ for all states } k \geq 0 \tag{2.78}$$

where $0 < \alpha < 1$ is a fixed parameter of the model.

The linear equation $\pi = \pi\mathbf{P}$ gives this infinite system of equations:

$$\begin{aligned} \pi(0) &= \sum_{k=0}^{\infty} (1-\alpha)\pi(k) \\ \pi(1) &= \alpha\pi(0) \\ \pi(2) &= \alpha\pi(1) \\ &\vdots \\ \pi(k) &= \alpha\pi(k-1) \\ &\vdots \end{aligned} \tag{2.79}$$

A moment's inspection suggests one possible strategy. Take $c = \pi(0)$ as a parameter whose value we try to pin down at the end. First solve successively for $\pi(1), \pi(2), \dots$ in terms of c:

$$\begin{aligned} \pi(1) &= \alpha\pi(0) = \alpha c \\ \pi(2) &= \alpha\pi(1) = \alpha^2 c \\ &\vdots \\ \pi(k) &= \alpha\pi(k-1) = \alpha^k c \\ &\vdots \end{aligned}$$

The first equation of (2.79) now reads

$$c = \sum_{k=0}^{\infty} (1-\alpha)\alpha^k c$$

and it simplifies to $c = c$, in other words, it is satisfied for all c.

It remains to check whether we can choose c so that $\pi(k) = \alpha^k c$ is a probability distribution. The choice is $c = 1 - \alpha$ which gives $\pi(k) = \alpha^k(1-\alpha)$, a shifted geometric distribution. △

Both previous examples had a unique invariant distribution. The next two examples illustrate the possibility of a multiplicity of invariant distributions and nonexistence of an invariant distribution.

Example 2.62. Suppose $\mathcal{S} = \{0,1,2,3\}$ and

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}. \tag{2.80}$$

The reader can check that for any $0 \le \theta \le 1$ the probability distribution

$$\pi = \begin{bmatrix} \theta\frac{8}{17} & \theta\frac{9}{17} & (1-\theta)\frac{8}{17} & (1-\theta)\frac{9}{17} \end{bmatrix} \tag{2.81}$$

is invariant.

The structure of $\mathbf{P}$ shows what is going on. This Markov chain consists of two irreducible recurrent Markov chains on the closed sets $\mathcal{R}' = \{0,1\}$ and $\mathcal{R}'' = \{2,3\}$ that do not communicate with each other at all. Both are copies of the Markov chain from Example 2.60, except that the states are labeled $2,3$ in the second copy. Thus the individual two-state Markov chains have the unique invariant distribution found in Example 2.60. This gives two invariant distributions

$$\pi' = \begin{bmatrix} \frac{8}{17} & \frac{9}{17} & 0 & 0 \end{bmatrix} \quad \text{and} \quad \pi'' = \begin{bmatrix} 0 & 0 & \frac{8}{17} & \frac{9}{17} \end{bmatrix} \tag{2.82}$$

for the four-state chain of (2.80).

In general, if π' and π'' both solve $\pi = \pi\mathbf{P}$, then so does any *convex combination* $\theta\pi' + (1-\theta)\pi''$ for any choice of $\theta \in [0,1]$. Convex combinations preserve also the property of being a probability distribution. Hence convex combinations of invariant distributions of a given transition probability $\mathbf{P}$ are also invariant distributions.

This explains the formula in (2.81): the complete set of invariant distributions consists of all the convex combinations of π' and π'' given in (2.82). The invariant distributions π' and π'' are special: they themselves cannot be expressed as convex combinations of other invariant distributions. Such invariant distributions are called *extreme*. △

Example 2.63 (Deterministically monotone Markov chain)**.** Consider the Markov chain from Example 2.30 with transition probability $p(x, x+1) = 1$ for all states x. An invariant distribution will not exist. But the failure mechanism is different for the two choices of state space $\mathbb{Z}_{\ge 0}$ and $\mathbb{Z}$.

Case 1. $\mathcal{S} = \mathbb{Z}_{\ge 0}$. Now $p(x,0) = 0$ for all x. The first equation of the system $\pi = \pi\mathbf{P}$ then becomes

$$\pi(0) = (\pi\mathbf{P})_0 = \sum_x \pi(x)p(x,0) = 0.$$

The subsequent equations are $\pi(1) = \pi(0)p(0,1) = 0$, $\pi(2) = \pi(1)p(1,2) = 0$, and so on. The only solution of $\pi = \pi\mathbf{P}$ is the vector π with all zeros. There is no invariant distribution.

Case 2. $\mathcal{S} = \mathbb{Z}$. The system $\pi = \pi\mathbf{P}$ gives the equations $\pi(x) = \pi(x-1)$ for all $x \in \mathbb{Z}$. Thus for any constant c, $\pi(x) = c$ solves the linear equation $\pi = \pi\mathbf{P}$ and these are the only solutions. But none of these solutions can give a probability distribution. If $c = 0$ then the solution is identically zero. If $c < 0$ the solution cannot be a probability distribution because its values are negative. If $c > 0$ then $\sum_x \pi(x) = \infty$ and again π cannot be a probability distribution. △

The examples show that the number of invariant distributions of a Markov chain can be zero, one or infinity. The theory that clarifies this is developed in the remainder of this section.

Theorem 2.59 extends significantly: *all* finite-dimensional distributions of the Markov chain remain the same over time, if the process is started with an invariant distribution. This leads to an important definition in the theory of stochastic processes. The definition below applies to all stochastic processes and is not specific to Markov chains.

Definition 2.64. Let $\{X_k\}_{k\geq 0}$ be a stochastic process with countable state space $\mathcal{S}$, defined on a probability space $(\Omega, \mathcal{F}, P)$. Then $\{X_k\}_{k\geq 0}$ is a **stationary process** if the following is true for all integers $m, n \geq 0$ and all states $x_0, \dots, x_n \in \mathcal{S}$:

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &\qquad = P(X_m = x_0, X_{m+1} = x_1, \dots, X_{m+n} = x_n) \end{aligned} \tag{2.83}$$

△

What the definition says is that while the state X_n keeps evolving randomly, *statistically* the process looks the same for all time.

Theorem 2.65. *Suppose π is an invariant distribution for the transition probability* $\mathbf{P}$ *and we take π as the initial distribution of the Markov chain. Then the Markov chain is a stationary process: for all integers $m, n \geq 0$ and all states $x_0, \dots, x_n \in \mathcal{S}$:*

$$\begin{aligned} &P_\pi(X_m = x_0, X_{m+1} = x_1, \dots, X_{m+n} = x_n) \\ &\qquad = P_\pi(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n). \end{aligned} \tag{2.84}$$

Proof. To calculate the probability on the left-hand side of (2.84), we add over all the possible outcomes of the states $X_0, \dots, X_{m-1}$, use a multistep transition probability, and finally the definition of invariance.

$$\begin{aligned}
&P_\pi(X_m = x_0, X_{m+1} = x_1, \dots, X_{m+n} = x_n) \\
&= \sum_{z_0,\dots,z_{m-1}} P_\pi(X_0 = z_0, \dots, X_{m-1} = z_{m-1}, \\
&\qquad\qquad\qquad\qquad X_m = x_0, X_{m+1} = x_1, \dots, X_{m+n} = x_n) \\
&\overset{(2.26)}{=} \sum_{z_0,\dots,z_{m-1}} \pi(z_0)\, p(z_0, z_1) \cdots p(z_{m-1}, x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \\
&= \sum_{z_0} \pi(z_0) \Bigl(\sum_{z_1,\dots,z_{m-1}} p(z_0, z_1) \cdots p(z_{m-1}, x_0) \Bigr) p(x_0, x_1) \cdots p(x_{n-1}, x_n) \\
&= \sum_{z_0} \pi(z_0) p^{(m)}(z_0, x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \\
&= \pi\mathbf{P}^m(x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \\
&\overset{(2.77)}{=} \pi(x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \\
&\overset{(2.26)}{=} P_\pi(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n). \qquad \square
\end{aligned}$$

We turn to develop the theory of existence and uniqueness of invariant distributions.

Existence and uniqueness of invariant distributions and measures. It turns out that a convenient way to develop the theory of invariant distributions goes through a generalization to invariant measures.

A *measure* on $\mathcal{S}$ is any function $\mu : \mathcal{S} \to [0, \infty]$. In other words, the values $\mu(x)$ are nonnegative reals or infinity. Sometimes this is called a *positive measure* to distinguish it from the case that allows both positive and negative values. In general, measures are thought of as functions on sets and they behave additively, so the measure of a subset $B \subset \mathcal{S}$ is

$$\mu(B) = \sum_{x \in B} \mu(x).$$

Probability measures are special cases of a measure: a measure μ is a *probability measure* or a *probability distribution* if $\sum_{x\in\mathcal{S}} \mu(x) = 1$.

We say that a measure μ is *identically zero* or the *zero measure* if $\mu(x) = 0$ for all $x \in \mathcal{S}$. Similarly, μ is *identically infinite* if $\mu(x) = \infty$ for all $x \in \mathcal{S}$. These two are uninteresting cases.

We generalize Definition 2.58 from invariant distributions to invariant measures.

Definition 2.66. Let $\mathbf{P} = \{p(x, y)\}_{x,y\in\mathcal{S}}$ be a transition probability and μ a measure on $\mathcal{S}$. Then μ is an *invariant measure* for $\mathbf{P}$ if $\mu = \mu\mathbf{P}$ and μ is neither identically zero nor identically infinite. △

As before, the equation $\mu = \mu\mathbf{P}$ means that

$$\mu(y) = \sum_{x\in\mathcal{S}} \mu(x) p(x, y) \quad \text{for all } y \in \mathcal{S}. \tag{2.85}$$

In a linear algebra sense, a measure μ can be thought of as a row vector, and then the product $\mu\mathbf{P}$ is the standard matrix product.

The invariant distribution defined in Definition 2.58 above is the special case where the invariant measure happens to be a probability measure. Note that if an invariant measure μ is *finite*, that is, it satisfies $c = \sum_{x\in\mathcal{S}} \mu(x) < \infty$, then μ can be normalized into an invariant distribution π by defining $\pi(x) = \mu(x)/c$.

In this section we state the main theorems about existence and uniqueness of invariant measures and distributions, then present examples to illustrate the results. Proofs of the theorems come afterwards.

By the canonical decomposition, any Markov chain consists of closed irreducible recurrent classes and a set of transient states. It turns out that any invariant distribution is simply zero on transient states (see Lemma 2.84 below). Hence we focus on recurrent states. The closed irreducible classes of recurrent states are really smaller individual Markov chains. Example 2.62 illustrated how invariant distributions for the large chain can be constructed from invariant distributions of the smaller closed chains. Thus we might as well work with an irreducible chain to begin with.

Theorem 2.67. *Assume that* $\mathbf{P}$ *is irreducible and recurrent. Then there is an invariant measure* ν *such that* $0 < \nu(x) < \infty$ *for all* $x \in S$. *The invariant measure is unique up to a constant multiple: that is, if* μ *is any other invariant measure, then there is a constant* $c \in (0, \infty)$ *such that* $\mu(x) = c\nu(x)$ *for all* $x \in S$.

To go from invariant measures to invariant distributions, we need a further classification of recurrent states according to the strength of recurrence as follows.

Definition 2.68. A recurrent state x is **positive recurrent** if $E_x[T_x] < \infty$ and **null recurrent** if $E_x[T_x] = \infty$. △

Theorem 2.69. *Let* x *and* y *be two communicating states* ($x \longleftrightarrow y$). *Then if one is positive recurrent, so is the other.*

The conclusion of the theorem above is sometimes expressed by saying that positive recurrence is a *class property*.

Theorem 2.70. *Assume that* $\mathbf{P}$ *is irreducible. Then* $\mathbf{P}$ *has an invariant distribution* π *if and only if all states are positive recurrent. In this case the invariant distribution* π *is unique and its values satisfy* $\pi(x) = \frac{1}{E_x[T_x]}$ *for all* $x \in S$.

By Theorem 2.69, an irreducible Markov chain is positive recurrent as soon as there is even a single positive recurrent state. The simplest special case is captured by the next theorem.

Theorem 2.71. *An irreducible finite Markov chain is positive recurrent.*

Example 2.72 (Mean return time)**.** Calculating mean return times is often harder than finding an invariant distribution. Then the formula $\pi(x) = \frac{1}{E_x[T_x]}$ can be used to get mean return times. Here are two illustrations.

(a) In the two-state Example 2.60 we have $E_0[T_0] = 1/\pi(0) = \frac{17}{8}$ for the mean return time to 0.

(b) For success runs Example 2.33 deduced that under P_0, $T_0 \sim \text{Geom}(1-\alpha)$. From this we know immediately that $E_0[T_0] = 1/(1-\alpha)$.

Calculating $E_x[T_x]$ directly from the transition probability for even moderately higher values such as $x = 6$ seems daunting because of all the different paths that the process can take before returning to state 6. But from Example 2.61 we can find easily all the mean return times. For example

$$E_6[T_6] = \frac{1}{\pi(6)} = \frac{1}{\alpha^6(1-\alpha)}.$$

To get a feeling for this, suppose $\alpha = 1/5$ so that my average success rate is 20%. Then if I ever find myself with 6 straight successes behind me, the mean number of shots till the next such occurrence is $5^7/4 \approx 19{,}531$. △

Remark 2.73 (Positive versus null recurrence)**.** In Theorem 2.87 (Section 2.6 below) we find an explanation for the terms positive and null recurrence. The asymptotic frequency of visits to a positive recurrent state x is strictly positive, and in fact exactly $\pi(x)$. By contrast, a null recurrent state is encountered so rarely that the asymptotic frequency is zero. This brings our Markov chain theory in contact with renewal theory from Section 1.3. △

Next we go over several more substantial examples of invariant measures and distributions. The first two are entire classes of examples, namely reversible Markov chains and doubly stochastic Markov chains. After examples we turn to the proofs of the theorems above.

Reversible measures.

Definition 2.74. A measure μ on $\mathcal{S}$ is a **reversible measure** if

$$\mu(x)\, p(x,y) = \mu(y)\, p(y,x) \quad \text{for all } x, y \in \mathcal{S}. \tag{2.86}$$

If a reversible measure is a probability distribution, it is called a **reversible distribution**. △

Condition (2.86) is called *detailed balance.* A reversible measure is not unique because any constant multiple of a reversible measure is another reversible measure.

Lemma 2.75. *A reversible measure is an invariant measure. A reversible distribution is an invariant distribution.*

Proof. Assume that μ is reversible. Then we can verify invariance.

$$\sum_x \mu(x)\, p(x,y) = \sum_x \mu(y)\, p(y,x) = \mu(y) \sum_x p(y,x) = \mu(y). \qquad \square$$

Remark 2.76 (Mass transport)**.** For a concrete illustration of the distinction between detailed balance and invariance, consider the following thought experiment.

The states of the Markov chain are geographic locations. An amount $\mu(x)$ of sand is placed at location x, for each state x. Overnight trucks transport sand between locations. For each pair of locations x, y, an amount $\mu(x)p(x,y)$ of sand is moved from x to y. (In other words, location x divides its sand among all locations

y in proportions $p(x, y)$.) In the morning of the next day, the amount of sand at location y is

$$\widetilde{\mu}(y) = \sum_x \mu(x)\, p(x, y).$$

If μ is invariant, then $\widetilde{\mu} = \mu$. In other words, each location has exactly the same amount of sand as before.

If μ is reversible, then $\mu(x)p(x, y) = \mu(y)p(y, x)$. This says that exactly the same amount of sand went from x to y as went from y to x. Thus, not only is the final sand distribution the same as the initial one (invariance), but the flow of sand is balanced on each individual route between locations (detailed balance). △

A *reversible Markov chain* is a Markov chain whose initial distribution is reversible. By Theorem 2.65 a reversible Markov chain is a stationary process. But even more is true: the statistics of the process are the same *regardless of whether time is run forward or backward*. The precise statement is the following. Let π be a reversible distribution. Then for any states $x_0, \dots, x_n$,

$$\begin{aligned} &P_\pi(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &\quad = P_\pi(X_n = x_0, X_{n-1} = x_1, \dots, X_0 = x_n). \end{aligned} \tag{2.87}$$

Exercise 2.16 asks for a proof of this and gives a hint.

There are many examples of reversible Markov chains in pure and applied probability. Reversibility is useful when it is present. For example, checking detailed balance (2.86) is easier than checking invariance $\mu = \mu\mathbf{P}$.

If the Markov chain is irreducible and recurrent, then any invariant measure μ satisfies $\mu(x) > 0$ for each x. Then (2.86) cannot hold unless transitions $p(x, y)$ and $p(y, x)$ are either both positive or both zero. For some Markov chains this observation makes it easy to determine that there cannot be a reversible measure. For example, it is immediate that the success run chain cannot have a reversible measure because $p(1, 2) > 0$ but $p(2, 1) = 0$.

Example 2.77 (Random walk on a graph)**.** A *graph* is a collection of *vertices* or *nodes*, some of which are connected through pairwise *edges* or *links*. Graphs are also called *networks*. We consider here an undirected simple graph. *Undirected* means that the edges are not directed from one vertex to the other. *Simple* means that between any two distinct vertices there is at most one edge, and self-loops are not allowed, that is, there is no edge from any vertex x to itself. Two distinct vertices x and y are *adjacent*, or *neighbors*, if they are connected by an edge, abbreviated $x \sim y$.

One way to encode the structure of a graph is by its *adjacency matrix* $\mathbf{A} = \{a(x, y)\}_{x,y \in \mathcal{V}}$, indexed by the set $\mathcal{V}$ of vertices. The definition is that

$$a(x, y) = \begin{cases} 1, & \text{if } x \sim y \\ 0, & \text{if } x \not\sim y. \end{cases}$$

In other words, the entries of $\mathbf{A}$ are 0s and 1s, and the 1s mark where the edges are. Since the graph is undirected, $x \sim y$ is equivalent to $y \sim x$, and consequently $\mathbf{A}$ is a symmetric matrix.

The *degree* of a vertex is

$$\deg(x) = \sum_{y \in \mathcal{V}} a(x,y) = \text{ the number of neighbors } x \text{ has.}$$

If the graph is finite (that is, $\mathcal{V}$ is a finite set), then each vertex has finite degree. If $\mathcal{V}$ is infinite we assume that each $\deg(x)$ is a finite integer.

A *random walk on a finite-degree graph* is a Markov chain whose states are the vertices, and each time it chooses its next state uniformly at random from the neighbors of its current state. In other words, the random walk moves along the edges of the graph, each time choosing one of the available edges uniformly at random. Let us assume that there are no *isolated vertices*, in other words, that $\deg(x) > 0$ for each vertex x. Then we can express the transition probability of the random walk as

$$p(x,y) = \frac{a(x,y)}{\deg(x)}. \tag{2.88}$$

Let $\mu(x) = \deg(x)$. Then for all states x and y,

$$\begin{aligned} \mu(x)\,p(x,y) &= \deg(x) \cdot \frac{a(x,y)}{\deg(x)} = a(x,y) = a(y,x) \\ &= \deg(y) \cdot \frac{a(y,x)}{\deg(y)} = \mu(y)\,p(y,x). \end{aligned}$$

Thus

$$\mu(x) = \deg(x) \tag{2.89}$$

defines a reversible measure. If the graph is finite, we can normalize μ into a reversible distribution π by defining

$$\pi(x) = \frac{\deg(x)}{\sum_{y \in \mathcal{V}} \deg(y)}.$$

△

Example 2.78 (Symmetric simple random walk)**.** Symmetric SRW is a special case of Example 2.77. The set of vertices is $\mathbb{Z}$ and adjacency is determined by $x \sim x+1$ for each integer x. So edges connect neighboring integers and $\deg(x) = 2$ for all vertices $x \in \mathbb{Z}$. Formula (2.88) for the transition probability gives $p(x, x \pm 1) = \frac{1}{2}$ which specifies symmetric SRW.

Since the degree is constant, formula (2.89) implies that any constant measure $\mu(x) = c > 0$ is reversible. Since symmetric SRW is irreducible and recurrent (Example 2.40), by Theorem 2.67 the invariant measure of symmetric SRW is unique up to constant multiple. Thus we have found all the invariant measures of symmetric SRW.

In particular, symmetric SRW does not have an invariant probability distribution and thereby is *not* positive recurrent. We conclude that $E_0[T_0] = \infty$, that is, the mean return time is infinite. △

Example 2.79 (Simple random walk)**.** Consider random walk on $\mathbb{Z}$ with transition probability $p(x, x+1) = p = 1 - p(x, x-1)$ where $0 < p < 1$ is arbitrary. The detailed balance equation becomes

$$\mu(x)p = \mu(x+1)(1-p) \quad \Longleftrightarrow \quad \mu(x+1) = \mu(x)\,\frac{p}{1-p}.$$

This can be solved: let $\mu(0) = c$, and then iterate the equation above to find

$$\mu(x) = c\left(\frac{p}{1-p}\right)^x \qquad \text{for all states } x \in \mathbb{Z}.$$

If $p = \frac{1}{2}$ we get the constant measures found in Example 2.78. For $p \neq \frac{1}{2}$ the measure μ above is not constant. We do not get a reversible distribution because for every $c > 0$, $\sum_x \mu(x) = \infty$. △

Doubly stochastic Markov chains.

Definition 2.80. A transition probability $\mathbf{P}$ is **doubly stochastic** if

$$\sum_{x \in \mathcal{S}} p(x, y) = 1 \qquad \text{for all states } y \in \mathcal{S}. \tag{2.90}$$

A little more vividly, in a doubly stochastic matrix both rows and columns sum to one. △

Theorem 2.81. *Assume* $\mathbf{P}$ *is doubly stochastic. Then every constant measure is invariant. If the state space is finite then there is an invariant distribution.*

Proof. Suppose $\mu(x) = c$. Then

$$\sum_x \mu(x)\,p(x, y) = c\sum_x p(x, y) = c = \mu(y)$$

shows the invariance of μ. If the state space is finite and the number of states is M, then $\pi(x) = 1/M$ defines an invariant distribution. □

Example 2.82 (Asymmetric simple random walk)**.** Consider random walk on $\mathbb{Z}$ with transition probability $p(x, x+1) = p = 1 - p(x, x-1)$ where $0 < p < 1$ satisfies $p \neq 1/2$. This Markov chain is doubly stochastic:

$$\sum_x p(x, y) = p(y-1, y) + p(y+1, y) = p + 1 - p = 1.$$

Consequently every constant measure is invariant.

If we combine this with Example 2.79, we find that asymmetric simple random walk has two separate families of invariant measures that are not constant multiples of each other across families. (The nonconstant measure $\mu(x) = (\frac{p}{1-p})^x$ cannot be a constant multiple of a constant measure.) Thus the uniqueness statement of Theorem 2.67 does not hold. The theorem is not violated because asymmetric simple random walk is not recurrent. △

On the other hand, the deterministically monotone Markov chain on $\mathbb{Z}$ (Case 2 of Example 2.63) shows that a transient Markov chain can have a unique invariant measure up to constant multiples. (This example is also the asymmetric SRW with $p = 1$ and a doubly stochastic chain.)

Further examples. An especially clear illustration of the demarcation between positive recurrence and null recurrence is given by the next example, the renewal chain.

Example 2.83 (Renewal chain)**.** The state space is $\mathcal{S} = \mathbb{Z}_{\geq 0}$, the set of nonnegative integers. Assume given a probability distribution $\{f_k\}_{k\geq 1}$ on positive integers. Define the transition probability on $\mathcal{S}$ as follows:

$$\begin{aligned} & p(0,k) = f_{k+1} \quad \text{for } k \geq 0 \\ \text{and} \quad & p(k,k-1) = 1 \qquad \text{for } k \geq 1. \end{aligned} \tag{2.91}$$

To give the mathematics some concrete meaning, imagine replacing a battery in a device whenever the battery dies. Each new battery has a random lifetime: f_k is the probability that the battery has to be replaced after k time units. The Markov chain X_n with transition probability (2.91) represents the remaining lifetime of the battery currently in use.

Let us assume that $f_k > 0$ for arbitrarily large k. Otherwise there is an integer k_0 such that $f_k = 0$ for $k > k_0$. Then the recurrent states are $\{0, 1, \dots, k_0\}$ which form a finite Markov chain. Then this example cannot demonstrate null recurrence.

The first observation is that, for any $k > 0$,

$$\begin{aligned} P_0(T_0 = k) &= P_0(X_1 = k-1, X_2 = k-2, \dots, X_k = 0) \\ &= p(0, k-1)\, p(k-1, k-2)\, p(k-2, k-3) \cdots p(1,0) \\ &= f_k \cdot 1 \cdot 1 \cdots 1 = f_k. \end{aligned}$$

In other words, $\{f_k\}$ is the probability distribution of T_0 when the Markov chain starts at 0.

From this we see that

$$P_0(T_0 < \infty) = \sum_{1\leq k<\infty} P_0(T_0 = k) = \sum_{1\leq k<\infty} f_k = 1$$

the last equality being simply the assumption that $\{f_k\}$ is a probability distribution. We have established that 0 is a recurrent state. Since the Markov chain is irreducible, it follows that all states are recurrent.

Next we turn to solving $\mu = \mu\mathbf{P}$. Theorem 2.67 guarantees the existence of an invariant measure. The system (2.85) now specializes to the equations

$$\mu(k) = \mu(0) f_{k+1} + \mu(k+1) \quad \forall\, k \geq 0.$$

We start by manipulating the first couple equations to see if a pattern emerges.

$$\mu(0) = \mu(0) f_1 + \mu(1) \quad \Longrightarrow \quad \mu(1) = \mu(0)(1 - f_1) = \mu(0) P_0(T_0 \geq 2).$$

$$\begin{aligned} \mu(1) = \mu(0) f_2 + \mu(2) \quad \Longrightarrow \quad \mu(2) &= \mu(1) - \mu(0) f_2 = \mu(0) P_0(T_0 \geq 2) - \mu(0) f_2 \\ &= \mu(0) P_0(T_0 \geq 2) - \mu(0) P_0(T_0 = 2) \\ &= \mu(0) P_0(T_0 \geq 3). \end{aligned}$$

The pattern $\mu(k) = \mu(0) P_0(T_0 \geq k+1)$ works for $k = 0, 1, 2$. Let us prove inductively that it holds for all k. So assume that $\mu(k) = \mu(0) P_0(T_0 \geq k+1)$ is true. (This is the *induction assumption.*) We check whether it is true for $k+1$:

$$\mu(k) = \mu(0) f_{k+1} + \mu(k+1)$$

implies that

$$\begin{aligned}\mu(k+1) = \mu(k) - \mu(0)f_{k+1} &= \mu(0)P_0(T_0 \ge k+1) - \mu(0)P_0(T_0 = k+1) \\ &= \mu(0)P_0(T_0 \ge k+2).\end{aligned}$$

Thus the pattern holds also for $k+1$. We have now proved that the system $\mu = \mu\mathbf{P}$ is solved by

$$\mu(k) = c\,P_0(T_0 \ge k+1) = c \sum_{j\ge k+1} f_j \quad \text{for } k \ge 0 \tag{2.92}$$

where $c > 0$ is any positive real number. These are the invariant measures whose existence was given by Theorem 2.67.

Next we ask under what condition we can have an invariant distribution. This requires that we can choose an invariant measure whose values sum to one:

$$\sum_{k=0}^{\infty} c\,P_0(T_0 \ge k+1) = 1$$

or equivalently, by a shift of the summation index and by taking the constant c outside the series,

$$c\sum_{j=1}^{\infty} P_0(T_0 \ge j) = 1. \tag{2.93}$$

This is possible if and only if

$$\sum_{j=1}^{\infty} P_0(T_0 \ge j) < \infty$$

because then we can choose c so that the value of the sum is exactly one. By Lemma B.4, $E_0[T_0] = \sum_{j=1}^{\infty} P_0(T_0 \ge j)$, so the condition above is exactly that $E_0[T_0] < \infty$, in other words, that 0 is a positive recurrent state. Then by irreducibility and Theorem 2.69 all states are positive recurrent.

Finally, under this positive recurrence assumption $E_0[T_0] < \infty$ which is the same as $\sum_{j=1}^{\infty} jf_j < \infty$, condition (2.93) is the same as $cE_0[T_0] = 1$ from which we solve $c = 1/E_0[T_0]$. Putting this c value back into (2.92) gives the invariant distribution

$$\pi(k) = \frac{P_0(T_0 \ge k+1)}{E_0[T_0]} \quad \text{for } k \ge 0. \tag{2.94}$$

We can also express this solution in terms of the given data $\{f_k\}$:

$$\pi(k) = \frac{\sum_{j\ge k+1} f_j}{\sum_{j\ge 1} jf_j} \quad \text{for } k \ge 0. \tag{2.95}$$

△

Proofs of the theorems for invariant distributions and measures. This subsection is devoted to the proofs. We start with a lemma that will be useful on a couple occasions.

Lemma 2.84. *Suppose* $\mathbf{P}$ *has an invariant distribution* π *and* x *is a state that satisfies* $\pi(x) > 0$*. Then* x *is a recurrent state.*

Proof. The proof begins by adding up a positive number $\pi(x)$ infinitely many times. Then we use invariance and switch the order of summation.

$$\begin{aligned}\infty &= \sum_{n=1}^{\infty} \pi(x) = \sum_{n=1}^{\infty} \sum_{z} \pi(z) p^{(n)}(z,x) = \sum_{z} \pi(z) \sum_{n=1}^{\infty} p^{(n)}(z,x) \\ &= \sum_{z} \pi(z) E_z[N(x)].\end{aligned}$$

Switching the order of the summation symbols is legitimate because all the terms are nonnegative (see Lemma A.2).

Suppose x were transient. Then by Theorem 2.38,

$$E_z[N(x)] = \frac{\rho_{zx}}{1-\rho_{xx}} \le \frac{1}{1-\rho_{xx}} < \infty.$$

Putting this back into the calculation above gives

$$\infty = \sum_{z} \pi(z) E_z[N(x)] \le \frac{1}{1-\rho_{xx}} \sum_{z} \pi(z) = \frac{1}{1-\rho_{xx}} < \infty.$$

We derived a contradiction from the assumption that x is transient. Hence x must be recurrent. □

Let x be a recurrent state so that $P_x(T_x < \infty) = 1$. Then for all states y, the expectation below makes sense, though a priori it could be infinite:

$$\lambda_x(y) = E_x\Big[\sum_{k=1}^{T_x} I_{\{X_k = y\}}\Big]. \tag{2.96}$$

We make two preliminary observations before turning to study λ_x. The total mass of the measure λ_x is the expected return time $E_x[T_x]$:

$$\begin{aligned}\lambda_x(\mathcal{S}) &= \sum_{y\in\mathcal{S}} \lambda_x(y) = \sum_{y\in\mathcal{S}} E_x\Big[\sum_{k=1}^{T_x} I_{\{X_k=y\}}\Big] \\ &= E_x\Big[\sum_{k=1}^{T_x}\sum_{y\in\mathcal{S}} I_{\{X_k=y\}}\Big] = E_x[T_x].\end{aligned} \tag{2.97}$$

The last equality follows from $\sum_{y\in\mathcal{S}} I_{\{X_k=y\}} = 1$. Sums and expectations can be rearranged above because all the terms are nonnegative.

Note also that

$$\lambda_x(x) = 1. \tag{2.98}$$

This follows from

$$P_x\Big(\sum_{k=1}^{T_x} I_{\{X_k=x\}} = 1\Big) = 1$$

which itself follows from (i) that T_x is finite with probability one under P_x by the assumption of recurrence, and (ii) there is then exactly one nonzero term in the sum, namely $I_{\{X_{T_x}=x\}} = 1$. Thus for each x, λ_x is a positive measure on the state space $\mathcal{S}$ that is neither identically zero nor identically infinite.

Theorem 2.85. *Let* $\mathbf{P}$ *be the transition matrix of an irreducible and recurrent Markov chain on the state space* $\mathcal{S}$*. Then for all* $x, y \in \mathcal{S}$*,* $0 < \lambda_x(y) < \infty$ *and* $\lambda_x(x) = 1$*. For all* $x \in \mathcal{S}$*, the measure* λ_x *is invariant:* $\lambda_x = \lambda_x\mathbf{P}$*.*

Proof. We prove the invariance $\lambda_x = \lambda_x\mathbf{P}$ first because this makes the proof of $0 < \lambda_x(y) < \infty$ easy. The proof of the invariance is a calculation that takes a few steps but each individual step requires nothing more than care with the symbols. The beginning of the calculation below uses a common trick: we rid ourselves of the random upper summation limit by introducing an indicator into the sum, and then we can take the sum outside the expectation. At this point we do not know if the sums below converge, but since all the terms are nonnegative, the sums can be rearranged safely.

$$
\begin{aligned}
\lambda_x(y) &= E_x\Big[\sum_{k=1}^{T_x} I_{\{X_k=y\}}\Big] = E_x\Big[\sum_{k=1}^{\infty} I_{\{X_k=y, T_x\ge k\}}\Big] \\
&= \sum_{k=1}^{\infty} P_x(X_k = y, T_x \ge k) \\
&= \sum_{k=1}^{\infty} P_x(X_1 \ne x, \dots, X_{k-1} \ne x, X_k = y) \\
&\overset{(a)}{=} p(x,y) + \sum_{k=2}^{\infty} P_x(X_1 \ne x, \dots, X_{k-1} \ne x, X_k = y) \\
&\overset{(b)}{=} p(x,y) + \sum_{k=2}^{\infty}\sum_{z\ne x} P_x(X_1 \ne x, \dots, X_{k-2} \ne x, X_{k-1} = z, X_k = y) \\
&\overset{(c)}{=} p(x,y) + \sum_{k=2}^{\infty}\sum_{z\ne x} P_x(X_1 \ne x, \dots, X_{k-2} \ne x, X_{k-1} = z) p(z,y) \\
&= p(x,y) + \sum_{z\ne x} p(z,y) \sum_{k=2}^{\infty} P_x(X_1 \ne x, \dots, X_{k-2} \ne x, X_{k-1} = z) \\
&\overset{(d)}{=} p(x,y) + \sum_{z\ne x} p(z,y) \sum_{k=2}^{\infty} P_x(X_{k-1} = z, T_x \ge k-1) \\
&\overset{(e)}{=} p(x,y) + \sum_{z\ne x} p(z,y) \sum_{k=1}^{\infty} P_x(X_k = z, T_x \ge k) \\
&\overset{(f)}{=} p(x,y)\lambda_x(x) + \sum_{z\ne x} p(z,y)\lambda_x(z) = \sum_{z} \lambda_x(z) p(z,y).
\end{aligned}
$$

This proves invariance. Explanation of the steps taken above:

Step (a) separated the term $k = 1$ from the sum.

Step (b) decomposed the probability according to the state $z \neq x$ at time $k-1$.

Step (c) used the Markov property to restart the process at time $k-1$.

Step (d) converted the event $X_1 \neq x, \ldots, X_{k-2} \neq x$ into the equivalent statement $T_x \geq k-1$. It would have been equally correct to write $T_x \geq k$ since z runs over states other than x. But writing $k-1$ is more convenient for the last steps of the proof.

Step (e) shifted the summation index.

Step (f) multiplied $p(x, y)$ by $\lambda_x(x) = 1$, and identified the sum over k as $\lambda_x(z)$, exactly as was done in the beginning of the calculation.

From the invariance and the assumed irreducibility we can deduce that $0 < \lambda_x(y) < \infty$ for all states x, y. Pick m and n so that $p^{(m)}(x, y) > 0$ and $p^{(n)}(y, x) > 0$. Note that $\lambda_x = \lambda_x \mathbf{P}$ implies that $\lambda_x = \lambda_x \mathbf{P}^k$ for all $k \geq 0$. Then

$$\lambda_x(y) = \sum_z \lambda_x(z) p^{(m)}(z, y) \geq \lambda_x(x) p^{(m)}(x, y) = p^{(m)}(x, y) > 0. \tag{2.99}$$

Next,

$$1 = \lambda_x(x) = \sum_z \lambda_x(z) p^{(m)}(z, x) \geq \lambda_x(y) p^{(n)}(y, x),$$

from which $\lambda_x(y) \leq \frac{1}{p^{(n)}(y,x)} < \infty$. □

Proof of Theorem 2.67. We begin by showing that any nontrivial invariant measure, that is, neither the zero measure nor identically infinite, is a constant multiple of some λ_x.

Suppose that $\widetilde{\mu}$ is an invariant measure. As in (2.99),

$$\widetilde{\mu}(y) \geq \widetilde{\mu}(x) p^{(n)}(x, y) \qquad \text{for all } x, y \text{ and } n \geq 1.$$

By irreducibility, for any x and y we can pick n so that $p^{(n)}(x, y) > 0$. Then if there is any y such that $\widetilde{\mu}(y) = 0$ we can conclude that $\widetilde{\mu}(x) = 0$ for all x. Utilizing the inequality in the opposite direction shows that if there is any x such that $\widetilde{\mu}(x) = \infty$ then $\widetilde{\mu}(y) = \infty$ for all y. Thus if $\widetilde{\mu}$ is neither the zero measure nor identically infinite, then $0 < \widetilde{\mu}(x) < \infty$ for all x.

Fix any x and replace $\widetilde{\mu}$ by $\mu(z) = \widetilde{\mu}(x)^{-1} \widetilde{\mu}(z)$ which is also invariant, is a constant multiple of $\widetilde{\mu}$ and satisfies $\mu(x) = 1$. Our first step is to prove the following statement:

$$\mu = \lambda_x. \tag{2.100}$$

The road to (2.100) goes through this claim that we prove by induction on n:

$$\text{for all } y \in \mathcal{S}\text{:} \;\; \mu(y) \geq \sum_{k=1}^{n} P_x(X_k = y, T_x \geq k). \tag{2.101}$$

The base case $n = 1$ comes from

$$\mu(y) = \sum_z \mu(z) p(z, y) \geq p(x, y) = P_x(X_1 = y) = P_x(X_1 = y, T_x \geq 1). \tag{2.102}$$

The last step used that $T_x \geq 1$ by definition.

Assume that (2.101) is true for n. We prove it for $n+1$. The first inequality below uses the induction assumption.

$$\begin{aligned}
\mu(y) &= \sum_z \mu(z)p(z,y) = p(x,y) + \sum_{z\neq x} \mu(z)\,p(z,y) \\
&\geq p(x,y) + \sum_{z\neq x}\sum_{k=1}^{n} P_x(X_k = z, T_x \geq k)\,p(z,y) \\
&\overset{(a)}{=} p(x,y) + \sum_{k=1}^{n}\sum_{z\neq x} P_x(X_k = z, T_x \geq k, X_{k+1} = y) \\
&\overset{(b)}{=} p(x,y) + \sum_{k=1}^{n} P_x(T_x \geq k+1, X_{k+1} = y) \\
&\overset{(c)}{=} P_x(T_x \geq 1, X_1 = y) + \sum_{k=2}^{n+1} P_x(T_x \geq k, X_k = y) \\
&= \sum_{k=1}^{n+1} P_x(T_x \geq k, X_k = y)
\end{aligned}$$

Step (a) used the Markov property in the following way, relying on the fact that the event $T_x \geq k$ depends only on $X_0, \dots, X_{k-1}$:

$$\begin{aligned}
P_x(X_k = z, T_x \geq k)\,p(z,y) &= P_x(X_k = z, T_x \geq k)\,P_x(X_{k+1} = y \,|\, X_k = z, T_x \geq k) \\
&= P_x(X_k = z, T_x \geq k, X_{k+1} = y).
\end{aligned}$$

Step (b) combined the events $X_k \neq x$ and $T_x \geq k$ into $T_x \geq k+1$. Step (c) applied (2.102) to $p(x,y)$ and shifted the index in the k-sum.

Inequality (2.101) has been verified for $n+1$, and now it holds for all $n \geq 1$.

Let $n \to \infty$ in (2.101) to obtain

$$\mu(y) \geq \sum_{k=1}^{\infty} P_x(X_k = y, T_x \geq k) = \lambda_x(y). \tag{2.103}$$

We have now verified $\mu \geq \lambda_x$. By this inequality, we can define yet another positive measure $\sigma(y) = \mu(y) - \lambda_x(y)$. This difference is well-defined for all y because both $\mu(y)$ and $\lambda_x(y)$ are finite quantities. We check that σ is another invariant measure:

$$\begin{aligned}
\sum_z \sigma(z)p(z,y) &= \sum_z (\mu(z) - \lambda_x(z))p(z,y) \overset{(a)}{=} \sum_z \mu(z)p(z,y) - \sum_z \lambda_x(z)p(z,y) \\
&= \mu(y) - \lambda_x(y) = \sigma(y).
\end{aligned}$$

Step (a) above is nontrivial: it is valid because the series $\sum_z \mu(z)p(z,y) = \mu(y)$ and $\sum_z \lambda_x(z)p(z,y) = \lambda_x(y)$ converge. If both series were to diverge, we would not be able to split the series in two as we did in step (a) because $\infty - \infty$ is not defined.

To show that σ vanishes, let $y \in \mathcal{S}$ and use irreducibility to pick n so that $p^{(n)}(y,x) > 0$. Then

$$0 = 1 - 1 = \mu(x) - \lambda_x(x) = \sigma(x) = \sum_z \sigma(z)p^{(n)}(z,x) \geq \sigma(y)p^{(n)}(y,x).$$

Since $p^{(n)}(y,x) > 0$, this forces $\sigma(y) = 0$. Thus σ is the zero measure, and we conclude that $\mu = \lambda_x$, in other words, we have established (2.100). Since x was arbitrary, it follows that the original nontrivial invariant measure $\widetilde{\mu}$ is a constant multiple of λ_x for every x.

To finish the proof, we can choose any λ_w as the invariant measure ν whose existence is claimed in Theorem 2.67. □

We prove that positive recurrence is a class property.

Proof of Theorem 2.69. Suppose one of the states x and y is positive recurrent. Then by Lemma 2.42 they are both recurrent. Apply Theorems 2.85 and 2.67 to the irreducible, recurrent Markov chain whose state space is the closed, recurrent class $\mathcal{R}$ that contains x and y. Since the process can never leave $\mathcal{R}$, calculation of expectations E_x and E_y are the same, regardless of whether we take the original (possibly larger) state space $\mathcal{S}$ or $\mathcal{R}$.

By Theorem 2.67 there is a finite, positive constant c such that $\lambda_y = c\lambda_x$. Hence by (2.97),

$$E_y[T_y] = \sum_{z\in\mathcal{S}} \lambda_y(z) = c\sum_{z\in\mathcal{S}} \lambda_x(z) = cE_x[T_x].$$

Thus $E_x[T_x]$ and $E_y[T_y]$ are finite together, or infinite together. □

We prove that irreducibility and positive recurrence together are equivalent to the existence of a unique invariant distribution.

Proof of Theorem 2.70. Assume irreducibility and positive recurrence. Then any of the λ_x measures gives an invariant measure that now has finite total mass $\lambda_x(\mathcal{S}) = E_x[T_x] < \infty$. Thus we can take any particular $x \in \mathcal{S}$ and define an invariant probability measure by $\pi = \frac{1}{E_x[T_x]}\lambda_x$.

Conversely, suppose $\mathbf{P}$ is irreducible and has an invariant distribution π. By Lemma 2.84 there must be at least one recurrent state. Then by irreducibility the entire state space is recurrent. Theorem 2.67 applies and says that for each $x \in \mathcal{S}$ there is a constant c_x such that $\lambda_x = c_x\pi$. Recalling now (2.97), we get

$$E_x[T_x] = \lambda_x(\mathcal{S}) = c_x\pi(\mathcal{S}) = c_x < \infty. \tag{2.104}$$

In particular, each expectation $E_x[T_x]$ is finite, and so the Markov chain is positive recurrent.

We have proved the equivalence of positive recurrence and existence of an invariant distribution for irreducible Markov chains. It remains to prove that an invariant distribution necessarily satisfies $\pi(x) = \frac{1}{E_x[T_x]}$ and hence in particular is unique. In (2.104) above we deduced that $c_x = E_x[T_x]$ is the constant that satisfies $\lambda_x = c_x\pi$. In other words,

$$\lambda_x(y) = c_x\pi(y) \quad \text{for all states } y. \tag{2.105}$$

Recalling that $\lambda_x(x) = 1$, we get

$$\pi(x) = \frac{\lambda_x(x)}{c_x} = \frac{1}{E_x[T_x]}. \qquad \square$$

We prove that an irreducible Markov chain with finite state space is always positive recurrent.

Proof of Theorem 2.71. By Lemma 2.52 an irreducible Markov chain with finite state space is recurrent. Thus we can apply Theorem 2.85. The quantity

$$E_x[T_x] = \sum_{z\in\mathcal{S}} \lambda_x(z)$$

must now be finite because it is a sum of finitely many terms and by Theorem 2.85 each term $\lambda_x(z)$ is finite. □

2.6. Limit theorems

In this section we prove two limit theorems for Markov chains as time tends to infinity: a version of the strong law of large numbers for Markov chains and then the convergence of the probability distributions.

Strong law of large numbers. Laws of large numbers are concerned with convergence of averages. Let $f : \mathcal{S} \to \mathbb{R}$ be a real-valued function on the state space $\mathcal{S}$. The idea is that while the random quantities $f(X_0), f(X_1), f(X_2), \dots$ may fluctuate randomly without ever settling down, under suitable conditions their temporal averages $n^{-1}\sum_{k=0}^{n-1} f(X_k)$ do converge as $n \to \infty$ to the nonrandom limit $\sum_x f(x)\pi(x)$ where π is the invariant distribution. A succinct way to put the result is that *the time average of a function converges to the space average given by the invariant distribution.*

The proof we give of the Markov chain strong law of large numbers relies on the SLLN for i.i.d. random variables that was stated as Theorem 1.13 in Section 1.2. To apply the i.i.d. SLLN we uncover an i.i.d. structure in the Markov chain. Fix a recurrent state z and define the stopping times T_z^j as before in (2.52): $T_z^0 = 0$ and for $j \geq 1$,

$$T_z^j = \inf\{n > T_z^{j-1} : X_n = z\}.$$

Assuming irreducibility and recurrence guarantees that each $T_z^j < \infty$ with probability one (Lemma 2.54). Every time the process returns to z it forgets its past, so the subsequent evolution has the same distribution as a Markov chain started from z, independently of what happened before. This gives the i.i.d. structure.

The finite Markov chain case is simpler than the general countable Markov chain case. Consequently we present first the SLLN of the finite case. After that we go through the technically more involved general countable state space case.

The treatments are separate, except for the frequency limit of Theorem 2.87 that is used for both. The reader who is interested in the stronger result can skip ahead.

Strong law of large numbers for a finite Markov chain. The lemma below establishes the i.i.d. structure of the inter-arrival times. The distribution of the first cycle length T_z^1 can differ from the subsequent ones, unless the process is started from the state z.

Lemma 2.86. *Suppose* $\mathbf{P}$ *is irreducible and recurrent. Let* μ *be an arbitrary initial distribution. Then the random variables* $\{T_z^j - T_z^{j-1} : j \geq 1\}$ *are independent and* $\{T_z^j - T_z^{j-1} : j \geq 2\}$ *is an i.i.d. process. Specifically, they have the following joint distribution: for integers* $k \geq 1$ *and* $m_1, \ldots, m_k \geq 1$,

$$\begin{aligned} &P_\mu\{T_z^j - T_z^{j-1} = m_j \textit{ for } j = 1, \ldots, k\} \\ &\qquad = P_\mu\{T_z^1 = m_1\} \cdot \prod_{j=2}^{k} P_z\{T_z^1 = m_j\}. \end{aligned} \tag{2.106}$$

Proof. The case $k = 1$ is obvious since $T_z^1 - T_z^0 = T_z^1$. We assume (2.106) and prove that it holds for $k + 1$. The argument is a routine application of the strong Markov property: restart the process at time T_z^k and use induction.

$$\begin{aligned} &P_\mu\{T_z^j - T_z^{j-1} = m_j \text{ for } j = 1, \ldots, k+1\} \\ &= P_\mu\bigl\{T_z^j - T_z^{j-1} = m_j \text{ for } j = 1, \ldots, k,\ T_z^k < \infty,\ X_{T_z^k} = z, \\ &\qquad\qquad X_{T_z^k+1} \neq z, \ldots, X_{T_z^k+m_{k+1}-1} \neq z,\ X_{T_z^k+m_{k+1}} = z\bigr\} \\ &= P_\mu\bigl\{T_z^j - T_z^{j-1} = m_j \text{ for } j = 1, \ldots, k\bigr\} \\ &\qquad \cdot P_z\bigl\{X_1 \neq z, \ldots, X_{m_{k+1}-1} \neq z,\ X_{m_{k+1}} = z\bigr\} \\ &\quad = P_\mu\{T_z^1 = m_1\} \cdot \Bigl(\prod_{j=2}^{k} P_z\{T_z^1 = m_j\}\Bigr) \cdot P_z\{T_z^1 = m_{k+1}\}. \end{aligned}$$

Identity (2.106) has been verified. □

As a preliminary step towards the SLLN, the next theorem below gives the long term frequency of visits to a given state. Finite state space is *not* assumed, only irreducibility and recurrence. Let

$$J_z^n = \sum_{k=1}^{n} I_{\{X_k = z\}} \tag{2.107}$$

denote the number of visits to z between times 1 and n (inclusive). The normalized quantity J_z^n/n is the *frequency* of visits to z up to time n. The relationship with the random variable N_z introduced in (2.56) is that

$$\lim_{n\to\infty} J_z^n = N_z. \tag{2.108}$$

Theorem 2.87. *Suppose* $\mathbf{P}$ *is irreducible and recurrent. Let* μ *be an arbitrary initial distribution. Let* z *be any state. Then the following limits hold with probability one:*

$$\lim_{k\to\infty} \frac{T_z^k}{k} = E_z[T_z] \tag{2.109}$$

and

$$\lim_{n\to\infty} \frac{J_z^n}{n} = \frac{1}{E_z[T_z]}. \tag{2.110}$$

In the second limit above, the convention is that $\frac{1}{\infty} = 0$. This illuminates the terms positive recurrent and null recurrent: the long term frequency of visits to a positive recurrent state is strictly positive, but zero for null recurrent states.

The reader can anticipate the limits (2.109) and (2.110) from the renewal process discussion of Section 1.3. The inter-arrival times $T_z^j - T_z^{j-1}$ for $j \geq 2$ are i.i.d. with expectation $E_z[T_z]$. The process $\{J_z^n : n \geq 0\}$ is essentially a discrete-time renewal process that counts the visits to z. The first cycle of length T_z^1 may be different from the rest, but one cycle does not influence asymptotic averages. Consequently (2.109) comes from the i.i.d. SLLN in Theorem 1.15 and (2.110) is basically the renewal limit of Theorem 1.19. A detailed proof follows. The proof of (2.110) is essentially a repeat of the argument given earlier for Theorem 1.19.

Proof of Theorem 2.87. By Lemma 2.54, each $T_z^k < \infty$ with probability one. Rewrite the ratio T_z^k/k in terms of a telescoping sum:

$$\frac{T_z^k}{k} = \frac{T_z^1}{k} + \frac{1}{k}\sum_{j=2}^{k}(T_z^j - T_z^{j-1})$$

In the first term T_z^1 is a random quantity that does not change with k. Hence as $k \to \infty$, $T_z^1/k \to 0$.

In the telescopic sum we have, by Theorem 2.93, a sequence $\{T_z^j - T_z^{j-1}\}_{j\geq 2}$ of nonnegative i.i.d. random variables whose mean is

$$E_\mu[T_z^j - T_z^{j-1}] = E_z[T_z^1 - T_z^0] = E_z[T_z^1].$$

Note in particular that the form of the expectation above follows from the common distribution in (2.116): for $j \geq 2$,

$$P_\mu(T_z^j - T_z^{j-1} = m) = P_z(T_z^1 = m).$$

By Corollary 1.15, regardless of whether the expectation is finite or infinite, the limit below holds with probability one:

$$\lim_{k\to\infty} \frac{1}{k}\sum_{j=2}^{k}(T_z^j - T_z^{j-1}) = E_z[T_z^1].$$

We have proved the limit in (2.109).

The random variables T_z^k and J_z^n are connected by the inequalities

$$T_z^{J_z^n} \leq n < T_z^{J_z^n+1}. \tag{2.111}$$

This holds because $J_z^n = k$ means that there are exactly k visits to z during times $1, \dots, n$ and $T_z^k \le n < T_z^{k+1}$ is an equivalent way of saying the same.

From the assumptions of irreducibility and recurrence it follows that $P_\mu(N_z = \infty) = 1$ (Lemma 2.54). Hence as $n \to \infty$, $J_z^n \to \infty$ with probability one. In particular, it is strictly positive for large enough n, and we can divide by it in (2.111) to get

$$\frac{T_z^{J_z^n}}{J_z^n} \le \frac{n}{J_z^n} < \frac{T_z^{J_z^n+1}}{J_z^n+1} \cdot \frac{J_z^n+1}{J_z^n}. \tag{2.112}$$

Let $n \to \infty$. As the random index J_z^n tends to infinity, two things happen: the limit in (2.109) takes place on the left and right in (2.112), and on the right also $\frac{J_z^n+1}{J_z^n} \to 1$. Since n/J_z^n is sandwiched between two sides that converge to $E_z[T_z]$, we must have $n/J_z^n \to E_z[T_z]$. By taking reciprocals we get the claim (2.110). □

Theorem 2.88. (Strong law of large numbers for finite Markov chains.) *Suppose the state space $\mathcal{S}$ is finite and $\mathbf{P}$ is irreducible. Let π be the unique invariant distribution. Let $f : \mathcal{S} \to \mathbb{R}$ be a real function on the state space. Let μ be an arbitrary initial distribution. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{x\in\mathcal{S}} f(x)\,\pi(x) \quad \textit{with probability one.} \tag{2.113}$$

Proof. This proof rests on the observation that $\sum_{k=1}^n f(X_k)$ is a sum of terms $f(x)$ where each $f(x)$ is repeated J_x^n times. Here is the formal derivation:

$$\sum_{k=1}^{n} f(X_k) = \sum_{k=1}^{n} \Bigl(\sum_{x\in\mathcal{S}} f(x) I_{\{X_k=x\}} \Bigr) = \sum_{x\in\mathcal{S}} f(x) \sum_{k=1}^{n} I_{\{X_k=x\}} = \sum_{x\in\mathcal{S}} f(x) J_x^n.$$

The limit comes from Theorem 2.87 and Theorem 2.70 that identified $\pi(x)$ as $\frac{1}{E_x[T_x]}$:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) &= \lim_{n\to\infty} \frac{1}{n} \sum_{x\in\mathcal{S}} f(x) J_x^n \overset{(a)}{=} \sum_{x\in\mathcal{S}} f(x) \lim_{n\to\infty} \frac{J_x^n}{n} \\ &\overset{(2.110)}{=} \sum_{x\in\mathcal{S}} f(x) \frac{1}{E_x[T_x]} = \sum_{x\in\mathcal{S}} f(x)\,\pi(x).\end{aligned}$$

The assumption of finite $\mathcal{S}$ was used in step (a) above to switch the limit and the summation. For an infinite state space this would be a nontrivial step. □

Example 2.89. Suppose $\mathcal{S} = \{0, 1\}$ and $\mathbf{P} = \begin{bmatrix} \frac14 & \frac34 \\ \frac23 & \frac13 \end{bmatrix}$. Given a real-valued function f on $\mathcal{S}$, find the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

This is a finite irreducible Markov chain so Theorem 2.88 applies directly. Example 2.60 found the invariant distribution $\pi = \bigl[\, \frac{8}{17} \;\; \frac{9}{17} \,\bigr]$. Thus

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \pi(0) f(0) + \pi(1) f(1) = \tfrac{8}{17} f(0) + \tfrac{9}{17} f(1)$$

with probability one, no matter how the process is started. △

Example 2.90. Suppose $\mathcal{S} = \{0, 1, 2, 3\}$ and

$$\mathbf{P} = \begin{bmatrix} \frac14 & \frac14 & \frac12 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac14 & \frac34 \\ 0 & 0 & \frac23 & \frac13 \end{bmatrix}. \tag{2.114}$$

Start the Markov chain at 0. Let $f : \mathcal{S} \to \mathbb{R}$ and find the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

This time 0 is a transient state, and there are two closed irreducible recurrent classes: $\{1\}$ and $\{2, 3\}$. We cannot predict with certainty which class we enter and consequently at best we can give the probabilities of different limits.

If the process is absorbed in state 1, then $T_1 < \infty$ and $T_2 = \infty$, and we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) &= \lim_{n\to\infty} \Bigl(\frac{1}{n} \sum_{k=1}^{T_1 - 1} f(X_k) + \frac{1}{n} \sum_{k=T_1}^{n} f(X_k) \Bigr) \\ &= \lim_{n\to\infty} \Bigl(\frac{T_1 - 1}{n} f(0) + \frac{n - T_1 + 1}{n} f(1) \Bigr) = f(1). \end{aligned}$$

If the process is absorbed in the class $\{2, 3\}$, then this happens at time $T_2 < \infty$. After that the process forgets its past and behaves like a Markov chain started in the set $\{2, 3\}$. This 2-state chain is the one in Example 2.89 above and hence the limit is the same one. Thus on the event $T_2 < \infty$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) &= \lim_{n\to\infty} \Bigl(\frac{1}{n} \sum_{k=1}^{T_2 - 1} f(X_k) + \frac{1}{n} \sum_{k=T_2}^{n} f(X_k) \Bigr) \\ &= \lim_{n\to\infty} \Bigl(\frac{T_1 - 1}{n} f(0) + \frac{1}{n} \sum_{k=T_2}^{n} f(X_k) \Bigr) = \tfrac{8}{17} f(2) + \tfrac{9}{17} f(3). \end{aligned}$$

For a full description we find the probabilities of the two eventualities.

$$\begin{aligned} P_0(T_1 < \infty) &= \sum_{n=1}^{\infty} P_0(T_1 = n) = \sum_{n=1}^{\infty} P_0(X_1 = \cdots = X_{n-1} = 0, X_n = 1) \\ &= \sum_{n=1}^{\infty} \bigl(\tfrac14\bigr)^n = \tfrac13. \end{aligned}$$

The final answer:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \begin{cases} f(1) & \text{with probability } \frac13, \\ \frac{8}{17} f(2) + \frac{9}{17} f(3) & \text{with probability } \frac23. \end{cases}$$

△

When the state space is countably infinite, the assumption of irreducibility alone does not guarantee the existence of an invariant distribution, and hence this assumption has to be made explicitly. Then the conclusion is again the same limit (2.113). This is proved as Theorem 2.94 in the next section. Here is an example of its application.

Example 2.91. Consider the success run chain with transition probability

$$p(k, k+1) = \alpha, \quad p(k, 0) = 1-\alpha, \quad \text{and} \quad p(k, \ell) = 0 \ \text{ for } \ell \notin \{0, k+1\}$$

for nonnegative integers k and with $0 < \alpha < 1$. This Markov chain is irreducible and from Example 2.61 we know that its unique invariant distribution is $\pi(k) = \alpha^k(1-\alpha)$. Thus SLLN applies and gives the conclusion that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{k=0}^{\infty} f(k)\alpha^k(1-\alpha)$$

for any function $f : \mathbb{Z}_{\geq 0} \to \mathbb{R}$ that satisfies $\sum_{k\geq 0} |f(k)|\alpha^k(1-\alpha) < \infty$.

Here are some particular questions that can be answered by this theorem.

(a) Suppose I receive a reward of 1 dollar for each success, but only when the success run has length at least two. What is the long-term rate of rewards, relative to the number of trials? Define f by

$$f(k) = I_{\{k\geq 3\}} = \begin{cases} 0, & k \leq 2 \\ 1, & k \geq 3. \end{cases}$$

The long-term reward rate is

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{k=0}^{\infty} f(k)\alpha^k(1-\alpha) = \sum_{k=3}^{\infty} \alpha^k(1-\alpha) = \alpha^3.$$

(b) Calculate the average size of the success run over the long term.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \sum_{k=0}^{\infty} k\alpha^k(1-\alpha) = \frac{\alpha}{1-\alpha}.$$

On the line above we applied SLLN to the function $f(k) = k$. △

Strong law of large numbers for a countable Markov chain. This part starts by strengthening the i.i.d. structure to include the evolution of the Markov chain between successive visits to the recurrent state z. This leads to the *dissection principle* that decomposes the entire infinite path of the Markov chain into independent blocks. The random objects that capture the inter-arrival evolution are defined by

$$\eta_j = \big(T_z^j - T_z^{j-1}, X_{T_z^{j-1}}, X_{T_z^{j-1}+1}, \dots, X_{T_z^j - 1}\big), \quad j \geq 1. \tag{2.115}$$

The next example illustrates how the notation works.

Example 2.92. Suppose the state space of the Markov chain is some set of letters and we pick $z = a$. Suppose the first 15 states of the process evolve as follows:

$$(X_0, X_1, \dots, X_{14}, \dots) = (c, d, a, a, b, b, d, a, c, e, a, a, e, f, f, \dots).$$

First observe that

$$T_a^0 = 0,\ T_a^1 = 2,\ T_a^2 = 3,\ T_a^3 = 7,\ T_a^4 = 10,\ T_a^5 = 11.$$

Separate the evolution into blocks that start at the times T_z^j:

$$\boxed{c, d}\ \boxed{a}\ \boxed{a, b, b, d}\ \boxed{a, c, e}\ \boxed{a}\ \boxed{a, e, f, f, \dots}$$

Note that the first block can in principle start with any state, controlled by the initial distribution, but the subsequent blocks start with the state $z = a$. Then the inter-arrival evolution variables are given by

$$\eta_1 = (2, c, d),\ \ \eta_2 = (1, a),\ \ \eta_3 = (4, a, b, b, d),\ \ \eta_4 = (3, a, c, e),\ \ \eta_5 = (1, a).$$

$\triangle$

The random vector η_j is discrete and its length is random. The first component of η_j is a positive integer, and the other components are the coordinates of an $\mathcal{S}$-valued vector whose length is determined by the first component. Thus the state space of the process $\{\eta_j\}_{j\ge1}$ is the set

$$\mathcal{U} = \bigcup_{1\le m<\infty} \{m\} \times \mathcal{S}^m.$$

Below we write elements of the space $\mathcal{U}$ as $\mathbf{u} = (m, x_0, \dots, x_{m-1})$ where $m \in \mathbb{Z}_{>0}$ and $x_0, \dots, x_{m-1} \in \mathcal{S}$.

Theorem 2.93. (Dissection principle.) *Suppose* $\mathbf{P}$ *is irreducible and recurrent. Let* μ *be an arbitrary initial distribution. Then the random variables* $\{\eta_j\}_{j\ge1}$ *are independent and* $\{\eta_j\}_{j\ge2}$ *are i.i.d. Specifically, they have the following joint distribution:*

$$\begin{aligned} &P_\mu\{\eta_j = (m(j), x_0^j, \dots, x_{m(j)-1}^j) \text{ for } j = 1, \dots, k\} \\ &\quad = P_\mu\{T_z^1 = m(1), (X_0, \dots, X_{T_z^1-1}) = (x_0^1, \dots, x_{m(1)-1}^1)\} \\ &\qquad \cdot \prod_{j=2}^k P_z\{T_z^1 = m(j), (X_0, \dots, X_{T_z^1-1}) = (x_0^j, \dots, x_{m(j)-1}^j)\} \end{aligned} \tag{2.116}$$

where $\{(m(j), x_0^j, \dots, x_{m(j)-1}^j) : 1 \le j \le k\}$ *are arbitrary elements of* $\mathcal{U}$*.*

This theorem implies Lemma 2.86 since $T_z^j - T_z^{j-1}$ appears as the first component of η_j.

Before turning to the proof which is a straightforward application of the strong Markov property, let us highlight the salient features of equation (2.116). If we abbreviate the elements of $\mathcal{U}$ above as $\mathbf{u}_j = (m(j), x_0^j, \dots, x_{m(j)-1}^j)$ then (2.116) can be expressed succinctly as

$$P_\mu\biggl(\bigcap_{j=1}^k \{\eta_j = \mathbf{u}_j\}\biggr) = P_\mu(\eta_1 = \mathbf{u}_1) \cdot \prod_{j=2}^k P_z(\eta_1 = \mathbf{u}_j). \tag{2.117}$$

For $k \ge 2$, sum over all $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$. This leaves

$$P_\mu(\eta_k = \mathbf{u}_k) = P_z(\eta_1 = \mathbf{u}_k).$$

This shows that, for $k \ge 2$, the distribution of η_k under P_μ is the same as the distribution of η_1 under P_z. Then the product on the right of (2.117) shows that $\{\eta_j\}_{j\ge1}$ are independent. The choice of initial distribution μ influences the distribution of η_1. After that the distributions of $\eta_2, \eta_3, \eta_4, \dots$ are the same because each marks a new cycle started at z.

Proof of Theorem 2.93. The proof of (2.116) (equivalently, of (2.117)) goes by induction on k. The case $k = 1$ is true by the definition of η_1. Assume that (2.117) is true. We prove that it holds also for $k + 1$.

We can assume that $\mathbf{u}_{k+1}$ is of the form $\mathbf{u}_{k+1} = (m(k{+}1), z, x_1^{k+1} \dots, x_{m(k+1)-1}^{k+1})$ where the entries $x_1^{k+1} \dots, x_{m(k+1)-1}^{k+1}$ are distinct from z. For otherwise the probability that $\eta_{k+1} = \mathbf{u}_{k+1}$ is zero and (2.116) holds trivially for $k + 1$ because both sides are zero.

With this case out of the way, we calculate. In the first equality below we (i) add some superfluous events inside the probability to connect explicitly with form (2.43) of the strong Markov property, and (ii) express the event $\eta_{k+1} = \mathbf{u}_{k+1}$ entirely in terms of the random variables $\{X_{T_z^k+n}\}_{0\le n\le m(k+1)}$.

$$\begin{aligned}
&P_\mu\{\eta_j = \mathbf{u}_j \text{ for } j = 1, \dots, k+1\}\\
&= P_\mu\{\eta_j = \mathbf{u}_j \text{ for } j = 1, \dots, k,\ T_z^k < \infty,\ X_{T_z^k} = z,\\
&\qquad (X_{T_z^k}, X_{T_z^k+1}, \dots, X_{T_z^k+m(k+1)-1}) = (z, x_1^{k+1}, \dots, x_{m(k+1)-1}^{k+1}),\\
&\qquad X_{T_z^k+m(k+1)} = z\}\\
&\overset{(a)}{=} P_\mu\{\eta_j = \mathbf{u}_j \text{ for } j = 1, \dots, k,\ T_z^k < \infty,\ X_{T_z^k} = z\}\\
&\qquad \cdot P_z\{(X_0, X_1, \dots, X_{m(k+1)-1}) = (z, x_1^{k+1}, \dots, x_{m(k+1)-1}^{k+1}),\ X_{m(k+1)} = z\}\\
&\overset{(b)}{=} P_\mu(\,\eta_j = \mathbf{u}_j \text{ for } j = 1, \dots, k\,) \cdot P_z(\eta_{k+1} = \mathbf{u}_{k+1})\\
&\overset{(c)}{=} P_\mu(\eta_1 = \mathbf{u}_1) \cdot \prod_{j=2}^{k+1} P_z(\eta_j = \mathbf{u}_j).
\end{aligned}$$

Step (a) applies (2.43) where the event $A = \{\eta_j = \mathbf{u}_j \text{ for } j = 1, \dots, k\}$ depends on the process only up to time T_z^k. Step (b) tidies things up and step (c) applies the induction assumption. □

Next comes the main theorem, the SLLN for a countable Markov chain. Note the small difference in the summation limits between the earlier statement (2.113) and (2.118) below: one uses $\sum_{k=1}^n$ and the other $\sum_{k=0}^{n-1}$. This is just for notational convenience. The two versions are equivalent because $f(X_0)/n \to 0$ and $\frac{n+1}{n} \to 1$ as $n \to \infty$.

Theorem 2.94. (Strong law of large numbers for countable Markov chains.) *Suppose* $\mathbf{P}$ *is irreducible and positive recurrent with invariant distribution* π. *Let* $f : \mathcal{S} \to \mathbb{R}$ *be a real function on the state space such that the absolute mean* $\sum_x |f(x)|\, \pi(x)$ *is a finite number. Let* μ *be an arbitrary initial distribution. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \sum_{x\in\mathcal{S}} f(x)\, \pi(x) \quad \textit{with probability one.} \tag{2.118}$$

Proof. The proof uses the dissection principle to rewrite the sum in (2.113) as a sum of i.i.d. terms plus some error terms that vanish in the limit. Define the function $F : \mathcal{U} \to \mathbb{R}$ by

$$F(m, x_0, \dots, x_{m-1}) = \sum_{i=0}^{m-1} f(x_i).$$

Fix a state z. Applying F to the random vector η_j of (2.115) gives

$$F(\eta_j) = \sum_{k=T_z^{j-1}}^{T_z^j - 1} f(X_k).$$

Break up the sum in (2.113), utilizing (2.111) for $n-1$ instead of n. We adopt the notational convention $T_z(n) = T_z^n$ to avoid overly complicated superscripts on superscripts. Take n large enough so that $J_z^{n-1} > 0$ to justify writing the identity below:

$$\begin{aligned} \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \;&=\; \frac{F(\eta_1)}{n} \\ &+\; \frac{J_z^{n-1}}{n} \cdot \frac{1}{J_z^{n-1}} \sum_{j=2}^{J_z^{n-1}} F(\eta_j) \;+\; \frac{1}{n} \sum_{k=T_z(J_z^{n-1})}^{n-1} f(X_k). \end{aligned} \tag{2.119}$$

On the right in (2.119), the middle sum is the important term. By Theorem 2.93 it is a sum of i.i.d. terms. To apply the strong law to it, we first verify that the expectation of the absolute value $|F(\eta_j)|$ is finite for $j \ge 2$. The calculation may seem a little circuitous, but it brings us directly to definition (2.96) of λ_z and then to π.

$$\begin{aligned} &E_\mu[\,|F(\eta_j)|\,] = E_z[\,|F(\eta_1)|\,] = E_z\Bigl[\,\Bigl|\sum_{k=0}^{T_z-1} f(X_k)\Bigr|\,\Bigr] \le E_z\Bigl[\sum_{k=0}^{T_z-1} |f(X_k)|\Bigr] \\ &= E_z\Bigl[\sum_{k=0}^{T_z-1}\sum_x |f(x)| I_{\{X_k=x\}}\Bigr] = \sum_x |f(x)|\, E_z\Bigl[\sum_{k=0}^{T_z-1} I_{\{X_k=x\}}\Bigr] \\ &\overset{(2.96)}{=} \sum_x |f(x)|\,\lambda_z(x) \overset{(2.105)}{=} E_z[T_z]\sum_x |f(x)|\,\pi(x) < \infty. \end{aligned} \tag{2.120}$$

At the end we used the assumption on f. Following the same calculation without the absolute values results in an equality instead of an inequality and gives the value of the expectation:

$$E_\mu[F(\eta_j)] \;=\; E_z[T_z]\sum_x f(x)\,\pi(x). \tag{2.121}$$

Since $J_z^n \to \infty$ with probability one as $n \to \infty$, we combine the SLLN for i.i.d. variables (Theorem 1.13) and limit (2.109) to get

$$\begin{aligned}
&\lim_{n\to\infty} \frac{J_z^{n-1}}{n} \cdot \frac{1}{J_z^{n-1}} \sum_{j=2}^{J_z^{n-1}} F(\eta_j) \\
&\qquad = \lim_{n\to\infty} \frac{n-1}{n} \cdot \frac{J_z^{n-1}}{n-1} \;\cdot\; \lim_{n\to\infty} \frac{1}{J_z^{n-1}} \sum_{j=2}^{J_z^{n-1}} F(\eta_j) \\
&\qquad = \frac{1}{E_z[T_z]} \cdot E_z[F(\eta_1)] = \sum_x f(x)\,\pi(x).
\end{aligned} \tag{2.122}$$

The last equality came from the expectation calculation in (2.121).

So the middle sum on the right in (2.119) gives the limit on the right in (2.113). To finish the proof we must show that the first and third terms on the right in (2.119) converge to zero as $n \to \infty$.

For the first term

$$\frac{F(\eta_1)}{n} \to 0 \quad \text{as } n \to \infty$$

is immediate because the numerator is a finite random quantity that stays constant as $n \to \infty$.

Treating the last term of (2.119) is not so obvious because it moves as n grows. But we can handle it with another judicious application of the strong law. The goal is to show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=T_z(J_z^{n-1})}^{n-1} f(X_k) = 0. \tag{2.123}$$

To take advantage again of the i.i.d. cycle evolutions, we wish to extend the sum above to the full cycle from $T_z(J_z^{n-1})$ to $T_z(J_z^{n-1}+1)-1$. We can do this safely if we put absolute values around everything and apply the triangle inequality to get an upper bound.

$$\begin{aligned}
\left| \frac{1}{n} \sum_{k=T_z(J_z^{n-1})}^{n-1} f(X_k) \right| &\le \frac{1}{n} \sum_{k=T_z(J_z^{n-1})}^{n-1} |f(X_k)| \\
&\le \frac{1}{n} \sum_{k=T_z(J_z^{n-1})}^{T_z(J_z^{n-1}+1)-1} |f(X_k)|.
\end{aligned} \tag{2.124}$$

Define another function $H : \mathcal{U} \to \mathbb{R}$ by

$$H(m, x_0, \dots, x_{m-1}) = \sum_{i=0}^{m-1} |f(x_i)|.$$

Applying H to the random vector η_j of (2.115) gives

$$H(\eta_j) = \sum_{k=T_z^{j-1}}^{T_z^j - 1} |f(X_k)|. \tag{2.125}$$

Continue from (2.124).

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=T_z(J_z^{n-1})}^{n-1} f(X_k)\right| &\le \frac{1}{n}\sum_{k=T_z(J_z^{n-1})}^{T_z(J_z^{n-1}+1)-1} |f(X_k)| \\
&= \frac{1}{n}H(\eta_{J_z^{n-1}+1}) = \frac{1}{n}\sum_{j=2}^{J_z^{n-1}+1} H(\eta_j) \;-\; \frac{1}{n}\sum_{j=2}^{J_z^{n-1}} H(\eta_j) \\
&= \frac{J_z^{n-1}+1}{n}\cdot\frac{1}{J_z^{n-1}+1}\sum_{j=2}^{J_z^{n-1}+1} H(\eta_j) \;-\; \frac{J_z^{n-1}}{n}\cdot\frac{1}{J_z^{n-1}}\sum_{j=2}^{J_z^{n-1}} H(\eta_j) \\
&\underset{n\to\infty}{\longrightarrow} \frac{1}{E_z[T_z]}E_z[H(\eta_1)] \;-\; \frac{1}{E_z[T_z]}E_z[H(\eta_1)] = 0.
\end{aligned}$$

The limit above comes from (2.110) and the strong law applied to the i.i.d. random variables $\{H(\eta_j)\}_{j\ge 2}$. This verifies (2.123).

To summarize, we have shown that all the terms on the right-hand side of (2.119) converge: the first one to zero, the second one to $\sum_x f(x)\pi(x)$, and the third one to zero. We have proved (2.113). □

Markov chain convergence theorem. In this section we address the convergence of transition probabilities $p^{(n)}(x,y)$ as $n\to\infty$. Equivalently, we look at the asymptotics of high powers $\mathbf{P}^n$ of the transition matrix. An immediate obstruction appears, illustrated by this simplest of Markov chains.

Example 2.95. With state space $\mathcal{S}=\{0,1\}$, let the transition matrix be

$$\mathbf{P}=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

The powers of $\mathbf{P}$ alternate:

$$\mathbf{P}=\mathbf{P}^3=\mathbf{P}^5=\cdots=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \quad\text{and}\quad \mathbf{P}^2=\mathbf{P}^4=\mathbf{P}^6=\cdots=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}.$$

Thus no limit of $\mathbf{P}^n$ is possible as $n\to\infty$. △

The phenomenon in the example is *periodicity*: the transition matrix cycles forever among a fixed finite set of matrices and hence cannot converge. We create the right tools to take care of this shortcoming. Recall that the *greatest common divisor* $\gcd A$ of a nonempty set A of positive integers is the largest integer n such that each $k\in A$ is a multiple of n. For example, $\gcd\{15,20,30\}=5$ and $\gcd\{15,8\}=1$.

Definition 2.96. For each recurrent state, let

$$I_x=\{n\in\mathbb{Z}_{>0} : p^{(n)}(x,x)>0\} \tag{2.126}$$

denote the set of integers $n\ge 1$ such that return from x to x in n steps is possible. Define the **period** $d(x)$ of a recurrent state as $d(x)=\gcd I_x$. △

Recurrence of x guarantees that I_x is nonempty. A useful property I_x has is *additivity*: if $m,n\in I_x$ then also $m+n\in I_x$. This follows from $p^{(m+n)}(x,x)\ge p^{(m)}(x,x)\,p^{(n)}(x,x)$.

Example 2.97. Here are several simple examples of the period.

(a) In Example 2.95, $d(0) = d(1) = 2$.

(b) Any state x that satisfies $p(x,x) > 0$ is an example of $d(x) = 1$.

(c) Consider the Markov chain on the state space $\{1, 2, 3, \}$ with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix}$$

The set I_1 contains the integers 2 and 3. This already tells us that $d(1) = \gcd I_1 = 1$. Similarly $d(2) = d(3) = 1$. △

The next lemma establishes that communicating recurrent states share their period.

Lemma 2.98. *Let $x \neq y$ be two communicating ($x \longleftrightarrow y$) recurrent states. Then $d(x) = d(y)$.*

Proof. Pick integers $k, m \geq 1$ such that $p^{(k)}(x,y) > 0$ and $p^{(m)}(y,x) > 0$. Chapman-Kolmogorov equations (2.32) give us the inequality

$$p^{(k+m)}(x,x) \geq p^{(k)}(x,y)\, p^{(m)}(y,x) \; > \; 0.$$

This says that $k + m \in I_x$ and hence $k + m = a \cdot d(x)$ for some integer a.

Let ℓ be an arbitrary element of I_y. Then $p^{(\ell)}(y,y) > 0$, and we also have

$$p^{(k+\ell+m)}(x,x) \geq p^{(k)}(x,y)\, p^{(\ell)}(y,y)\, p^{(m)}(y,x) \; > \; 0.$$

Thus $k + \ell + m = b \cdot d(x)$ for some integer b. Together we have

$$\ell = (k + \ell + m) - (k + m) = (b - a)d(x).$$

We have shown that every element of I_y is a multiple of $d(x)$. Thus $d(x)$ is a common divisor of I_y, and hence $d(y) = \gcd I_y \geq d(x)$. Switching x and y around in the argument gives the opposite inequality $d(x) \geq d(y)$. The proof is complete. □

A consequence of the lemma is that all the states of an irreducible, recurrent Markov have the same period. This makes the next definition sensible.

Definition 2.99. Define the period of an irreducible, recurrent Markov chain as the common period of all its states. If that period is one, call the Markov chain **aperiodic**. △

Example 2.100. Consider the success run chain with transition probability

$$p(k, k+1) = \alpha, \quad p(k, 0) = 1 - \alpha, \quad \text{and} \quad p(k, \ell) = 0 \; \text{ for } \ell \notin \{0, k+1\}$$

for all nonnegative integers k. Assume $0 < \alpha < 1$. This is an irreducible, recurrent Markov chain, as deduced in Example 2.33. Hence to find the period of every state, it is enough to find the period of a single state. Since $p(0,0) > 0$, state 0 has period 1. Consequently this is an aperiodic, irreducible, recurrent Markov chain.

Note in particular that the periods are 1 for all states, even though the number of steps required for return depends dramatically on the state. For example, from state 100 it is not possible to return to state 100 in fewer than 101 steps. △

We have the terminology in place for stating the Markov chain convergence theorem.

Theorem 2.101. *Let* $\mathbf{P}$ *be aperiodic and irreducible, and assume that it has an invariant distribution* π. *Let* μ *be an arbitrary initial distribution. Then*

$$\lim_{n\to\infty} \sum_{y\in\mathcal{S}} |\, P_\mu(X_n = y) - \pi(y)| = 0. \tag{2.127}$$

We prove Theorem 2.101 after some discussion and examples.

By specializing μ to a point mass δ_x and recalling that $P_x(X_n = y) = p^{(n)}(x, y)$, we get the statement that for all states $x \in \mathcal{S}$,

$$\lim_{n\to\infty} \sum_{y\in\mathcal{S}} |\, p^{(n)}(x, y) - \pi(y)| = 0. \tag{2.128}$$

Since the sum above bounds any particular term in it, we also get a limit for each entry of the transition matrix $\mathbf{P}^n$:

$$\lim_{n\to\infty} p^{(n)}(x, y) = \pi(y) \quad \text{for all states } x, y \in \mathcal{S}. \tag{2.129}$$

The last statement has a concrete linear algebraic version: as $n \to \infty$, the matrix $\mathbf{P}^n$ converges entry by entry to the matrix whose each row equals π.

We revisit earlier examples.

Example 2.102. Return to the success run chain with transition probability

$$p(k, k+1) = \alpha, \quad p(k, 0) = 1 - \alpha, \quad \text{and} \quad p(k, \ell) = 0 \ \text{ for } \ell \notin \{0, k+1\}$$

for all nonnegative integers k with $0 < \alpha < 1$. From Examples 2.33, 2.61, and 2.100 we know that this is an aperiodic, irreducible, recurrent Markov chain with invariant distribution $\pi(k) = \alpha^k(1-\alpha)$. Thus Theorem 2.101 applies and gives the conclusion that

$$\lim_{n\to\infty} P_\mu(X_n = \ell) = \alpha^\ell(1-\alpha) \quad \text{for all nonnegative integers } k, \ell.$$

In plain English, this says that if you come and observe the process after a large time n, the probability that you see the process in state ℓ is approximately $\alpha^\ell(1-\alpha)$. Furthermore, this statement holds regardless of how the process was started.

For example, suppose the success probability is $\alpha = 0.7$. Suppose we begin at state 0. Then the probability that after 5 steps we find ourselves in state 5 is $p^{(5)}(0, 5) = 0.7^5 \approx 0.168$. After 6 steps we cannot be in state 5 so $p^{(6)}(0, 5) = 0$. But for very large n, $p^{(n)}(0, 5) \approx 0.7^5 \cdot 0.3 \approx 0.050$. △

Example 2.103. Suppose $\mathcal{S} = \{0, 1\}$ and $\mathbf{P} = \begin{bmatrix} \frac14 & \frac34 \\ \frac23 & \frac13 \end{bmatrix}$. This is a finite, aperiodic, irreducible Markov chain so Theorem 2.101 applies directly. Example 2.60 found the invariant distribution $\pi = \left[\, \frac{8}{17} \ \ \frac{9}{17} \,\right]$. Thus, for any initial state x,

$$\lim_{n\to\infty} p^{(n)}(x, y) = \begin{cases} \pi(0) = \frac{8}{17} & \text{if } y = 0, \\ \pi(1) = \frac{9}{17} & \text{if } y = 1. \end{cases}$$

△

Example 2.104. Suppose $\mathcal{S} = \{0, 1, 2, 3\}$ and

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

Start the Markov chain at 0. Find the limit $\lim_{n\to\infty} p^{(n)}(0, y)$ for each state y.

Theorem 2.101 does not apply immediately because this Markov chain is not irreducible.

Starting from 0 and looking at time n, there are three scenarios: either the process has already been absorbed to 1, or the process has entered the recurrent class $\{2, 3\}$, or the process is still at 0. Thus we decompose the transition probability:

$$\begin{aligned} p^{(n)}(0, y) = P_0(X_n = y) &= P_0(T_1 \le n, X_n = y) \\ &\quad + P_0(T_2 \le n, X_n = y) + P_0(\min\{T_1, T_2\} > n, X_n = y). \end{aligned}$$

From here the analysis splits into cases for different y. If the chain ever leaves 0 it cannot return. Hence

$$p^{(n)}(0, 0) = P_0(X_n = 0) = P_1(X_0 = \cdots = X_n = 0) = \left(\tfrac{1}{4}\right)^n \underset{n\to\infty}{\longrightarrow} 0.$$

Since the state $y = 1$ is absorbing,

$$\begin{aligned} p^{(n)}(0, 1) = P_0(T_1 \le n) = \sum_{m=1}^{n} P_0(T_1 = m) = \sum_{m=1}^{n} \left(\tfrac{1}{4}\right)^n \\ \underset{n\to\infty}{\longrightarrow} \sum_{n=1}^{\infty} \left(\tfrac{1}{4}\right)^n = \tfrac{1}{3}. \end{aligned}$$

For $y \in \{2, 3\}$, utilizing the Markov property:

$$p^{(n)}(0, y) = P_0(T_2 \le n, X_n = y) = \sum_{m=1}^{n} P_0(T_2 = m)\, p^{(n-m)}(2, y). \tag{2.130}$$

Taking the limit is not so obvious now. Separately we see that

$$\sum_{m=1}^{n} P_0(T_2 = m) = \sum_{m=1}^{n} \left(\tfrac{1}{4}\right)^{m-1} \tfrac{1}{2} \underset{n\to\infty}{\longrightarrow} \sum_{m=1}^{\infty} \left(\tfrac{1}{4}\right)^{m-1} \tfrac{1}{2} = \tfrac{2}{3}$$

and from Example 2.103, by applying the convergence theorem to the smaller Markov chain on $\{2, 3\}$,

$$\lim_{k\to\infty} p^{(k)}(2, y) = \pi(y) = \begin{cases} \frac{8}{17}, & y = 2 \\ \frac{9}{17}, & y = 3. \end{cases}$$

Above π is now the invariant distribution of the smaller Markov chain on $\{2, 3\}$.

The limit on line (2.130) can be reasoned nonrigorously as follows. After some n_0, the terms of the convergent series do not matter much and $p^{(n)}(2,y)$ is approximately $\pi(y)$. Hence for $n > 2n_0$,

$$\begin{aligned}\sum_{m=1}^{n} P_0(T_2 = m)\, p^{(n-m)}(2,y) &\approx \sum_{m=1}^{n_0} P_0(T_2 = m)\, p^{(n-m)}(2,y)\\ &\approx \sum_{m=1}^{n_0} P_0(T_2 = m)\pi(y) \approx \tfrac{2}{3}\pi(y).\end{aligned}$$

Exercise 2.26 asks you to derive rigorously $p^{(n)}(0,y) \to \frac{2}{3}\pi(y)$ for $y \in \{2,3\}$.

The full answer:

$$\lim_{n\to\infty} p^{(n)}(0,y) = \begin{cases} 0, & y = 0\\ \frac{1}{3}, & y = 1\\ \frac{16}{51}, & y = 2\\ \frac{18}{51}, & y = 3.\end{cases}$$

△

The next corollary carries significant meaning for applications. Suppose that a Markovian system has been running for a long time and the assumptions of aperiodicity, irreducibility and positive recurrence are reasonable. Then it is safe to assume that, no matter how the system started in the distant past, it has become approximately stationary. In other words, an observer can assume that the system behaves as a Markov chain that starts from its invariant distribution.

Corollary 2.105. *Let* $\mathbf{P}$ *be aperiodic and irreducible, and assume that it has an invariant distribution* π. *Let* μ *be an arbitrary initial distribution. Then for any sequence of states* $x_0, \dots, x_m$,

$$\begin{aligned}&\lim_{n\to\infty} P_\mu(X_n = x_0, X_{n+1} = x_1, \dots, X_{n+m} = x_m)\\ &\qquad = P_\pi(X_0 = x_0, X_1 = x_1, \dots, X_m = x_m).\end{aligned} \tag{2.131}$$

Proof of Corollary 2.105 assuming Theorem 2.101. Use the Markov property to restart the process at time n and apply the limit (2.127) to the probability $P_\mu(X_n = x_0)$:

$$\begin{aligned}&P_\mu(X_n = x_0, X_{n+1} = x_1, \dots, X_{n+m} = x_m)\\ &= P_\mu(X_n = x_0)\, P_{x_0}(X_1 = x_1, \dots, X_{n+m} = x_m)\\ &\xrightarrow[n\to\infty]{} \pi(x_0)\, P_{x_0}(X_1 = x_1, \dots, X_{n+m} = x_m)\\ &= P_\pi(X_0 = x_0, X_1 = x_1, \dots, X_{n+m} = x_m).\end{aligned}$$

□

Example 2.106. Suppose a Markov chain on the state space $\{0,1\}$ with transition matrix

$$\mathbf{P} = \begin{array}{c|cc} & 0 & 1\\ \hline 0 & 1-a & a\\ 1 & b & 1-b\end{array}$$

has been running for a long time. You come to observe the process for three time units. What is the probability that the sequence of states you see is $(0,0,0)$?

There are three cases.

(i) If $a = b = 0$ then $\mathbf{P} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and then also $\mathbf{P}^n = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ for all $n \geq 0$. $\mathbf{P}$ is not irreducible so the convergence theorem does not apply. The answer depends on the initial distribution. Namely, for all n,

$$\begin{aligned} &P_\mu(X_n = 0, X_{n+1} = 0, X_{n+2} = 0) \\ &\quad = \mu(0)p^{(n)}(0,0)p(0,0)^2 + \mu(1)p^{(n)}(1,0)p(0,0)^2 \\ &\quad = \mu(0) \cdot 1 + \mu(1) \cdot 0 = \mu(0). \end{aligned}$$

(ii) If $a = b = 1$ then $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. $\mathbf{P}$ is irreducible but not aperiodic so the convergence theorem does not apply. However, under this $\mathbf{P}$ the state flips every time so no matter what the initial distribution, $P_\mu(X_n = 0, X_{n+1} = 0, X_{n+2} = 0) = 0$.

(iii) Suppose $0 < a + b < 2$. Then $\mathbf{P}$ is irreducible, aperiodic, and has the unique invariant distribution $\pi = [\pi(0), \pi(1)] = [\frac{b}{a+b}, \frac{a}{a+b}]$. By Corollary 2.105,

$$\lim_{n\to\infty} P_\mu(X_n = 0, X_{n+1} = 0, X_{n+2} = 0) = \pi(0)\, p(0,0)^2 = \frac{b(1-a)^2}{a+b}.$$

△

We need some more preparatory work before we can prove Theorem 2.101.

Lemma 2.107. *Let x be a recurrent state such that $d(x) = 1$. Then there exists a finite integer $m(x)$ such that $p^{(n)}(x,x) > 0$ for all $n \geq m(x)$.*

Proof. Though the statement is probabilistic, this lemma is really a fact about integers. The purely number-theoretic statement is the following.

(2.132) Suppose A is a nonempty subset of positive integers such that $\gcd A = 1$ and $m, n \in A$ implies $m + n \in A$. Then there exists an integer N such that A contains all integers $n \geq N$.

This statement applies to I_x and gives the claim of the lemma. We prove (2.132) in two steps.

Step 1. We show that A contains two consecutive integers. The assumptions guarantee that A is infinite, so we start by picking two integers $m, m + k$ in A where $k \geq 1$.

If $k = 1$ we have found two consecutive integers m and $m + 1$ in A.

Suppose $k > 1$. Then k cannot be a common divisor of A, and we can find $a \in A$ such that k does not divide a. Write $a = bk + r$ where $b \geq 0$ is the maximal number of times k goes into a and r is the remainder. Since k does not divide a evenly, we know that $r \in \{1, 2, \dots, k-1\}$. Additivity of A implies that $(b+1)(m+k)$ and $(b + 1)m + a$ are elements of A. Their difference is

$$(b + 1)(m + k) - (b + 1)m - a = bk + k - a = k - r \in \{1, 2, \dots, k - 1\}.$$

Thus we have found two elements of A whose difference is positive but strictly less than k.

We can repeat this argument as many times as needed until the original difference k has been reduced to 1. Then we have found two consecutive integers in A.

Step 2. Let $m, m+1 \in A$ from Step 1. Given $n \geq m^2$, there are $k \geq 0$ and $0 \leq r < m$ such that $n = m^2 + km + r$. Rewrite this as $n = (m - r + k)m + r(m+1)$ to show that $n \in A$. We have proved that (2.132) is true for $N = m^2$. □

We are ready for the proof of Theorem 2.101. In addition to the results of this section, the proof appeals to Lemmas 2.54 and 2.84 from previous sections.

Proof of Theorem 2.101. The idea of the proof is the following. We run two versions of the same Markov chain, called X_n and Y_n. X_n starts with distribution μ, while Y_n starts with distribution π. By invariance, Y_n is π-distributed at all times $n \geq 0$. We can control the discrepancy between the distribution of X_n and π by controlling the discrepancy between X_n and Y_n.

We let X_n and Y_n run independently of each other, but despite this, we can prove that at some point in time they come together. Since they obey the same transition probabilities, after this random meeting they are statistically indistinguishable. The farther we go in time, the likelier it is that X_n and Y_n met in the past, and hence the closer their distributions. In the limit their distributions come together.

Technically speaking this is a *coupling* of the Markov chains X_n and Y_n. In general, a coupling is any joint construction of different random variables or processes for the purpose of comparing them. Coupling is a ubiquitous tool in probability theory.

To turn this idea into a precise proof, we introduce the pair Markov chain $\overline{X}_n = (X_n, Y_n)$ on the Cartesian product state space $\overline{\mathcal{S}} = \mathcal{S} \times \mathcal{S}$. The transition probability of $\overline{X}_n$ is

$$\overline{p}((x,y),(z,w)) = p(x,z)\,p(y,w) \qquad \text{for states } (x,y),(z,w) \in \overline{\mathcal{S}}.$$

The logic of the construction is evident from the product formula above: $\overline{X}_n = (X_n, Y_n)$ moves from (x,y) to (z,w) if X_n moves from x to z and Y_n moves from y to w, and the latter two moves are made independently of each other.

Let $\overline{P}_\alpha$ denote the path probability measure associated with the pair process $\overline{X}_n$ when the initial distribution is α. We present the proof of Theorem 2.101 as a sequence of small individual steps.

Step 1. Take the initial distribution α to be of product form

$$\alpha(x,y) = \gamma(x)\eta(y) \tag{2.133}$$

where γ and η are two probability distributions on $\mathcal{S}$. We show that then the two components X_n and Y_n of $\overline{X}_n$ are independent versions of the original Markov chain with transition probability p and initial distributions γ and η. For this, we show that the probability of the intersection of an event of the X-process and an event of the Y-process is the product of the probabilities. Let $x_0, \dots, x_n, y_0, \dots, y_n$ be states in $\mathcal{S}$.

$$\begin{aligned}
&\overline{P}_\alpha(X_0 = x_0, \dots, X_n = x_n, Y_0 = y_0, \dots, Y_n = y_n) \\
&= \overline{P}_\alpha\big(\overline{X}_0 = (x_0, y_0), \dots, \overline{X}_n = (x_n, y_n)\big) \\
&= \alpha(x_0, y_0)\, \overline{p}((x_0, y_0), (x_1, y_1)) \cdots \overline{p}((x_{n-1}, y_{n-1}), (x_n, y_n)) \\
&= \gamma(x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n) \cdot \eta(y_0)\, p(y_0, y_1) \cdots p(y_{n-1}, y_n) \\
&= P_\gamma(X_0 = x_0, \dots, X_n = x_n) \cdot P_\eta(X_0 = y_0, \dots, X_n = y_n).
\end{aligned}$$

Above we used the definitions and applied (2.26) first to $\overline{P}_\alpha$ and then in the last equality to the original Markov chains. Summing above over $x_0, \dots, x_{n-1}, y_0, \dots, y_{n-1}$ and renaming x_n as x and y_n as y gives the identity

$$\overline{P}_\alpha(X_n = x, Y_n = y) = P_\gamma(X_n = x) \cdot P_\eta(X_n = y). \tag{2.134}$$

In (2.133) take $\gamma = \delta_z$ and $\eta = \delta_w$, point masses on states z and w. Then $\alpha = \delta_{(z,w)}$ and (2.134) specializes to

$$\overline{p}^{\,(n)}((z, w), (x, y)) = p^{(n)}(z, x) \cdot p^{(n)}(w, y). \tag{2.135}$$

Summing in (2.134) over y and then again separately over x gives,

$$\overline{P}_\alpha(X_n = x) = P_\gamma(X_n = x) \quad \text{and} \quad \overline{P}_\alpha(Y_n = y) = P_\eta(X_n = y). \tag{2.136}$$

Step 2. Transition probability $\overline{p}$ *has the invariant distribution* $\overline{\pi}(x, y) = \pi(x)\pi(y)$. We check this by a calculation: using the definitions of $\overline{\pi}$ and $\overline{p}$ and the invariance of π,

$$\begin{aligned}
\sum_{(z,w)\,\in\,\overline{\mathcal{S}}} \overline{\pi}(z, w)\, \overline{p}((z, w), (x, y)) &= \sum_{z,w\,\in\,\mathcal{S}} \pi(z)\, \pi(w)\, p(z, x)\, p(w, y) \\
&= \sum_{z\,\in\,\mathcal{S}} \pi(z)\, p(z, x) \sum_{w\,\in\,\mathcal{S}} \pi(w)\, p(w, y) = \pi(x)\pi(y) = \overline{\pi}(x, y).
\end{aligned}$$

Step 3. Transition probability $\overline{p}$ *is irreducible.* This step uses the assumption that the original transition probability p is irreducible, recurrent, and aperiodic. Given $(x, y), (z, w) \in \overline{\mathcal{S}}$, we need to show that for some n,

$$\overline{p}^{\,(n)}((x, y), (z, w)) > 0.$$

Use first the irreducibility of p to choose k and ℓ so that $p^{(k)}(x, z) > 0$ and $p^{(\ell)}(y, w) > 0$. Then use recurrence, aperiodicity and Lemma 2.107 to choose m large enough so that $p^{(\ell+m)}(z, z) > 0$ and $p^{(k+m)}(w, w) > 0$. Using (2.135) combine these to derive

$$\begin{aligned}
\overline{p}^{\,(k+\ell+m)}((x, y), (z, w)) &= p^{(k+\ell+m)}(x, z)\, p^{(k+\ell+m)}(y, w) \\
&\ge p^{(k)}(x, z)\, p^{(\ell+m)}(z, z)\, p^{(\ell)}(y, w)\, p^{(k+m)}(w, w) > 0.
\end{aligned}$$

Step 4. Transition probability $\overline{p}$ *is irreducible and recurrent.* Lemma 2.84 and the invariant distribution (Step 2) imply that there is at least one recurrent state. Then by irreducibility (Step 3) all states are recurrent.

Step 5. Fix a state $u \in \mathcal{S}$ and let

$$\tau = \inf\{n \ge 1 : \overline{X}_n = (u, u)\} \tag{2.137}$$

be the first visit of $\overline{X}_n$ to state (u, u) after time 0. Then for any initial distribution α, $\overline{P}_\alpha(\tau < \infty) = 1$. This follows from Lemma 2.54 applied to the pair Markov chain $\overline{X}_n$.

Step 6. With τ as in (2.137) and y an arbitrary state in $\mathcal{S}$,

$$\overline{P}_\alpha(\tau \le n, X_n = y) = \overline{P}_\alpha(\tau \le n, Y_n = y). \tag{2.138}$$

This is the precise expression of the notion that after X_n and Y_n have met, they are statistically identical. Step (a) of the calculation below uses the following symmetry of the transition probability:

$$\overline{p}((u, u), (y, z)) = p(u, y)p(u, z) = p(u, z)p(u, y) = \overline{p}((u, u), (z, y)).$$

To use the Markov property, we sum over the values of the random variables τ and $\overline{X}_n$.

$$\begin{aligned}
\overline{P}_\alpha(\tau \le n, X_n = y) &= \sum_{m=1}^{n} \sum_{z} \overline{P}_\alpha\bigl(\tau = m, \overline{X}_m = (u, u), \overline{X}_n = (y, z)\bigr) \\
&= \sum_{m=1}^{n} \sum_{z} \overline{P}_\alpha\bigl(\tau = m, \overline{X}_m = (u, u)\bigr)\, \overline{p}^{\,(n-m)}((u, u), (y, z)) \\
&\overset{(a)}{=} \sum_{m=1}^{n} \sum_{z} \overline{P}_\alpha\bigl(\tau = m, \overline{X}_m = (u, u)\bigr)\, \overline{p}^{\,(n-m)}((u, u), (z, y)) \\
&= \sum_{m=1}^{n} \sum_{z} \overline{P}_\alpha\bigl(\tau = m, \overline{X}_m = (u, u), \overline{X}_n = (z, y)\bigr) \\
&= \overline{P}_\alpha(\tau \le n, Y_n = y).
\end{aligned}$$

Step 7. The final calculation. Take the initial distribution in (2.133) to be $\alpha(x, y) = \mu(x)\pi(y)$.

$$\begin{aligned}
\sum_y \bigl| P_\mu(X_n = y) - \pi(y) \bigr| &= \sum_y \bigl| P_\mu(X_n = y) - P_\pi(X_n = y) \bigr| \\
&\overset{(2.136)}{=} \sum_y \bigl| \overline{P}_\alpha(X_n = y) - \overline{P}_\alpha(Y_n = y) \bigr| \\
&= \sum_y \Bigl| \overline{P}_\alpha(\tau \le n, X_n = y) - \overline{P}_\alpha(\tau \le n, Y_n = y) \\
&\qquad\qquad + \overline{P}_\alpha(\tau > n, X_n = y) - \overline{P}_\alpha(\tau > n, Y_n = y) \Bigr| \\
&\overset{(2.138)}{=} \sum_y \Bigl| \overline{P}_\alpha(\tau > n, X_n = y) - \overline{P}_\alpha(\tau > n, Y_n = y) \Bigr| \\
&\le \sum_y \overline{P}_\alpha(\tau > n, X_n = y) \; + \; \sum_y \overline{P}_\alpha(\tau > n, Y_n = y) \\
&= 2\overline{P}_\alpha(\tau > n).
\end{aligned}$$

The last inequality used the triangle inequality $|a+b| \le |a|+|b|$. The last probability $\overline{P}_\alpha(\tau > n)$ converges to zero as $n \to \infty$ because $\overline{P}_\alpha(\tau < \infty) = 1$ as verified in Step

5. The zero limit is justified by Lemma B.2 from Appendix B. Limit (2.127) has been proved. □

2.7. Generating function methods

This section introduces the probability generating function as a tool for studying random processes and then applies this tool to the branching process, another central model of both pure and applied probability.

Probability generating function.

Definition 2.108. Let X be a random variable whose values are nonnegative integers, possibly including infinity. The **probability generating function** (p.g.f.) g of X is defined by

$$g(s) = \sum_{k=0}^{\infty} s^k P(X = k) \tag{2.139}$$

for real s such that $|s| \le 1$. We write g_X when the notation needs to specify the p.g.f. of the random variable X. △

The series in (2.139) is absolutely convergent for $|s| \le 1$:

$$\sum_{k=0}^{\infty} |s|^k P(X = k) \le \sum_{k=0}^{\infty} P(X = k) \le 1.$$

The last sum is always ≤ 1 but is < 1 if $P(X = \infty) > 0$. Note that the series in (2.139) does not include the term $P(X = \infty)$, even if this probability were positive, because the series is defined as the limit $\lim_{n\to\infty} \sum_{k=0}^{n} s^k P(X = k)$. We can express the p.g.f. as an expectation: $g(s) = E[s^X\, I_{\{X<\infty\}}]$ where the role of the indicator is to restrict the expectation to the event $X < \infty$. The formula simplifies if X is a finite random variable:

$$\text{if } X \text{ is finite, then } g(s) = E[s^X].$$

Thus g is a power series with radius of convergence at least 1. A function given by such a series determines uniquely its coefficients. In other words, the probability mass function $P(X = k)$ and the p.g.f. g determine each other uniquely. We state this as a theorem.

Theorem 2.109. *The probability distribution of a $\mathbb{Z}_{\ge 0} \cup \{\infty\}$-valued random variable is uniquely determined by its probability generating function.*

Example 2.110. Suppose $X \sim \text{Bin}(n, p)$. Then

$$g_X(s) = \sum_{k=0}^{n} s^k \binom{n}{k} p^k (1-p)^{n-k} = (ps + 1 - p)^n,$$

in this case defined for all real s.

Suppose $Y \sim \text{Geom}(p)$. This means that Y is the number of trials required for the first success in a sequence of independent trials with success probability p.

Then

$$g_Y(s) = \sum_{k=1}^{\infty} s^k (1-p)^{k-1} p = \frac{ps}{1-s(1-p)} \quad \text{for } |s| < \frac{1}{1-p},$$

where we observed that the series converges if and only if $|s(1-p)| < 1$. △

We record various properties of the p.g.f.

The p.g.f. is nondecreasing for nonnegative arguments: that is,

$$0 \le s < t \text{ implies that } g(s) \le g(t). \tag{2.140}$$

The reason is that the coefficients $P(X = k)$ are nonnegative, so each (nonzero) term in (2.139) increases as s increases.

The values of $g(s)$ are not in general probabilities, but these two are:

$$g(0) = P(X = 0) \tag{2.141}$$

and

$$g(1) = \sum_{k=0}^{\infty} P(X = k) = P(0 \le X < \infty). \tag{2.142}$$

In particular, $P(X = \infty) = 1 - g(1)$. If X is finite, $g(1) = 1$.

Next we observe that the generating function of a sum of independent random variables is the product of the generating functions. Assume that X and Y are independent and finite. Then

$$g_{X+Y}(t) = E[t^{X+Y}] = E[t^X \cdot t^Y] = E[t^X] \cdot E[t^Y] = g_X(t)\, g_Y(t).$$

Consequently, if $\{X_k\}$ are i.i.d. and $S_n = X_1 + \cdots + X_n$ for $n \ge 1$,

$$g_{S_n}(t) = E[t^{X_1+\cdots+X_n}] = E[t^{X_1}] \cdots E[t^{X_n}] = g_{X_1}(t)^n.$$

Let N be a nonnegative integer-valued random variable, independent of the i.i.d. terms $\{X_k\}$. The *random sum* S_N is defined by setting it equal to S_n on the event $\{N = n\}$. We calculate its p.g.f. by conditioning on the value of N. Step (a) below drops the conditioning, by the independence of N and $\{X_k\}$.

$$\begin{aligned} g_{S_N}(t) &= E[t^{S_N}] = \sum_{n=0}^{\infty} P(N=n)\, E[t^{S_N} \,|\, N = n\,] \\ &= \sum_{n=0}^{\infty} P(N=n)\, E[t^{S_n} \,|\, N = n\,] \overset{(a)}{=} \sum_{n=0}^{\infty} P(N=n)\, E[t^{S_n}] \\ &= \sum_{n=0}^{\infty} P(N=n)\, g_{X_1}(t)^n = g_N\big(g_{X_1}(t)\big). \end{aligned} \tag{2.143}$$

The derivative of the generating function. It turns out that the slope of the generating function $g(s)$ at $s = 1$ gives the mean EX. However, g fails to be differentiable at 1 if it blows up at all values $s > 1$ and the random variable may take infinite values. These complications make the matter somewhat technical. The next theorem states the simplest case and the previous examples revisited. After this we turn to the serious technical work.

Theorem 2.111. *Suppose X takes only nonnegative integer values and that its generating function $g(s)$ is finite in an open interval around $s = 1$. Then*

$$g'(1) = EX.$$

Example 2.112. For $X \sim \text{Bin}(n, p)$ we calculated $g_X(s) = (ps + 1 - p)^n$. From this, $g'_X(s) = np(ps + 1 - p)^{n-1}$ and then $g'_X(1) = np$, which we know agrees with EX.

For $Y \sim \text{Geom}(p)$, if $|s| < \frac{1}{1-p}$, we have

$$g_Y(s) = \frac{ps}{1 - s(1-p)} \quad \text{and} \quad g'_Y(s) = \frac{p}{(1 - s(1-p))^2}.$$

Then $g'_Y(1) = 1/p$, which we know agrees with EY. △

As a power series with radius of convergence 1, g can be differentiated term by term for $-1 < s < 1$. At $s = 1$ the derivative fails to exist if $g(s) = \infty$ for all $s > 1$. However, there does always exist a *left derivative* $g'(1-)$ defined as the limit

$$g'(1-) = \lim_{s \to 1-} \frac{g(1) - g(s)}{1 - s} \tag{2.144}$$

where $s \to 1-$ means that s approaches 1 from the left. We prove this in the next theorem and give the probabilistic meaning of this derivative.

Theorem 2.113. *The left derivative of g at 1 exists as a $[0, \infty]$-valued limit in (2.144) and equals the expectation of X restricted to the event where it is finite: $g'(1-) = E[X\, I_{\{X<\infty\}}]$. In particular, if $P(X < \infty) = 1$, then $g'(1-) = EX$. Furthermore, the derivatives converge: $\lim\limits_{s \to 1-} g'(s) = g'(1-)$.*

Proof. Let $s < 1$. We start by computing the difference quotient in terms of probabilities. Note that the term $k = 0$ vanishes in the second term below and can be dropped. Use the identity $\sum_{j=0}^{n} s^j = \frac{1-s^{n+1}}{1-s}$ for finite geometric series when $s \neq 1$. Switch the order of summation.

$$\begin{aligned} \frac{g(1) - g(s)}{1 - s} &= \sum_{k=1}^{\infty} \frac{1 - s^k}{1 - s} P(X = k) = \sum_{k=1}^{\infty} \sum_{j=0}^{k-1} s^j P(X = k) \\ &= \sum_{j=0}^{\infty} s^j \sum_{k=j+1}^{\infty} P(X = k) = \sum_{j=0}^{\infty} s^j P(j + 1 \le X < \infty). \end{aligned} \tag{2.145}$$

At this point we have to appeal to the monotone convergence theorem (Theorem A.4) from analysis that lies beyond the level of this book. This theorem guarantees that as $s \to 1-$, the last series above converges to $\sum_{j=0}^{\infty} P(j + 1 \le X < \infty)$. Since $P(j + 1 \le X < \infty) = P(X\, I_{\{X<\infty\}} \ge j + 1)$ for $j \ge 0$, we can write this series as

$$\sum_{j=0}^{\infty} P\big(X\, I_{\{X<\infty\}} \ge j + 1\big) = E\big[X\, I_{\{X<\infty\}}\big]$$

where the equality comes from Lemma B.4. Consequently also the leftmost member of (2.145) converges to this same value as $s \to 1-$.

For the last statement of the theorem, compute the derivative at $0 < s < 1$ and let $s \to 1-$.

$$g'(s) = \sum_{k=1}^{\infty} k s^{k-1} P(X = k) \underset{s \to 1-}{\longrightarrow} \sum_{k=1}^{\infty} k\, P(X = k) = E[X\, I_{\{X<\infty\}}] = g'(1-).$$

The limit above utilized again the monotone convergence theorem. □

The next two examples illustrate the theorem. The first one is designed to have $P(X < \infty) = 1$ but $g'(1-) = EX = \infty$, while the second one has $P(X = \infty) > 0$ but $g'(1-) = E[XI_{\{X<\infty\}}] < \infty$.

Example 2.114. Let X have probability mass function $P(X = k) = \frac{1}{k(k-1)}$ for integers $k \ge 2$. This is a genuine probability mass function because

$$\sum_{k=2}^{\infty} \frac{1}{k(k-1)} = \lim_{n\to\infty} \sum_{k=2}^{n} \frac{1}{k(k-1)} = \lim_{n\to\infty} \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = \lim_{n\to\infty} \left(1 - \frac{1}{n} \right) = 1.$$

Its mean is the harmonic series

$$EX = \sum_{k=2}^{\infty} k\, P(X = k) = \sum_{k=2}^{\infty} \frac{1}{k-1} = \infty.$$

The p.g.f. $g(s) = \sum_{k=2}^{\infty} \frac{s^k}{k(k-1)}$ satisfies

$$g'(s) = \sum_{k=2}^{\infty} \frac{s^{k-1}}{k-1} = -\log(1-s)$$

and

$$g'(1-) = -\lim_{s\to 1-} \log(1-s) = \infty.$$

△

Example 2.115. Suppose $P(X = 1) = \frac{1}{2}$, $P(X = 2) = \frac{1}{3}$ and $P(X = \infty) = \frac{1}{6}$. Then $g(s) = \frac{1}{2}s + \frac{1}{3}s^2$ and $g'(s) = \frac{1}{2} + \frac{2}{3}s$.

$$E\big[X\, I_{\{X<\infty\}}\big] = \tfrac{1}{2} \cdot 1 + \tfrac{1}{3} \cdot 2 = \tfrac{7}{6}$$

which agrees with $g'(1) = \frac{7}{6}$. The unrestricted expectation satisfies $EX = \infty$ because $P(X = \infty) > 0$. △

The above properties of the p.g.f. are sufficient for our purposes and we turn to the branching process.

Extinction in the branching process. The purpose of this section is to introduce the *Galton-Watson* or *branching process* and to compute its extinction probability with generating functions. The branching process is a simple population model. Along with random walk, it is among the most important discrete stochastic processes for applications.

The state of the branching process at time n is the number X_n of individuals alive in the nth generation of the population. To deduce the next value X_{n+1}, each member of generation n gives birth to a random number of offspring independently of all the other individuals. These offspring form generation $n + 1$ and their total number is X_{n+1}. The process continues this way forever unless there is *extinction*, which means that $X_m = 0$ for some m, and then $X_n = 0$ for all $n \geq m$ thereafter. The only parameter needed to completely define the process is the *offspring distribution* $\beta = \{\beta_k\}_{0 \leq k < \infty}$ whose meaning is that

$$\beta_k = \text{ the probability that an individual leaves } k \text{ offspring in the next generation.}$$

To be specific, $\beta_\infty = 0$ so we do not admit the possibility of infinitely many offspring.

For a concrete picture, imagine a petri dish of bacteria. The number of bacteria alive at time n is X_n. Between time n and $n+1$, each bacterium dies and leaves k descendants with probability β_k for $k = 0, 1, 2, \ldots$

We begin the mathematical treatment of the branching process as we did with random walk. First we construct the process from i.i.d. random variables and then deduce its transition probability.

Let $\{Z_{n,j} : n, j \geq 1\}$ be i.i.d. random variables with probability mass function $P(Z_{n,j} = k) = \beta_k$ for $k \geq 0$. In the construction $Z_{n,j}$ represents the number of offspring that individual j from generation $n-1$ contributes to generation n. Since we do not know the sizes of the generations ahead of time, for each n we take infinitely many offspring variables $Z_{n,j}$ for $j = 1, 2, 3, \ldots$ so that we have enough to work with.

Here is the construction of the process started from a single live individual in generation 0. So first

$$X_0 = 1.$$

The lone individual of generation 0 gives rise to $Z_{1,1}$ offspring that form generation 1:

$$X_1 = Z_{1,1}.$$

Then unless $X_1 = 0$, the X_1 members of generation 1 give rise to $Z_{2,1}, Z_{2,2}, \ldots, Z_{2,X_1}$ offspring that make up generation 2:

$$X_2 = \begin{cases} 0, & \text{if } X_1 = 0 \\ Z_{2,1} + Z_{2,2} + \cdots + Z_{2,X_1}, & \text{if } X_1 \geq 1. \end{cases}$$

The pattern goes on. The general rule for going from generation n to generation $n+1$ is that

$$X_{n+1} = \begin{cases} 0, & \text{if } X_n = 0 \\ Z_{n+1,1} + Z_{n+1,2} + \cdots + Z_{n+1,X_n}, & \text{if } X_n \geq 1. \end{cases} \tag{2.146}$$

From the discussion and the formula in (2.146), the Markov property of the process is intuitively clear. Given the present state X_n, computing the next state X_{n+1} requires only new independent inputs $\{Z_{n+1,j} : j \geq 1\}$ and there is no dependence on the past. We leave the rigorous proof of the Markov property as Exercise 2.20 and turn to deriving a formula for the transition probability $p(k, \ell)$ for $k, \ell \in \mathbb{Z}_{\geq 0}$.

The first case of equation (2.146) gives $p(0,0)=1$. We deduce $p(k,\ell)$ for $k\ge 1$ from the second case of equation (2.146).

$$\begin{aligned}
p(k,\ell) &= P(X_{n+1}=\ell\,|\,X_n=k) = P\Bigl(\sum_{j=1}^{X_n} Z_{n+1,j}=\ell \Bigm| X_n=k\Bigr)\\
&\overset{(a)}{=} P\Bigl(\sum_{j=1}^{k} Z_{n+1,j}=\ell \Bigm| X_n=k\Bigr) \overset{(b)}{=} P\Bigl(\sum_{j=1}^{k} Z_{n+1,j}=\ell\Bigr)\\
&\overset{(c)}{=} \sum_{\substack{\ell_1,\dots,\ell_k\ge 0:\\ \ell_1+\dots+\ell_k=\ell}} P(Z_{n+1,1}=\ell_1)P(Z_{n+1,2}=\ell_2)\cdots P(Z_{n+1,k}=\ell_k)\\
&\overset{(d)}{=} \sum_{\substack{\ell_1,\dots,\ell_k\ge 0:\\ \ell_1+\dots+\ell_k=\ell}} \beta_{\ell_1}\beta_{\ell_2}\cdots\beta_{\ell_k} \overset{(e)}{=} \beta_\ell^{\star k}.
\end{aligned}$$

- In step (a) the conditioning fixes the upper summation limit $X_n=k$.
- In step (b) we can drop the conditioning because X_n is independent of $\{Z_{n+1,j}: j\ge 1\}$. This is because X_n is a function of the previously used offspring variables $\{Z_{m,j}: 1\le m\le n, j\ge 1\}$ and by assumption all the offspring variables are independent.
- Step (c) decomposes the probability of the sum into the different mutually exclusive cases and uses independence of $Z_{n+1,j}$s.
- Step (d) expressed probabilities in terms of the distribution β of each $Z_{n+1,j}$.
- Step (e) identified the sum of products as the k-fold convolution power of the probability mass function β.

Since $p(0,\ell)=\delta_0(\ell)=\beta_\ell^{\star 0}$ we can write one simple formula for the transition probability:

$$p(k,\ell)=\beta_\ell^{\star k} \quad \text{for } k,\ell\ge 0. \tag{2.147}$$

We have established the branching process as a Markov chain on the state space $\mathcal{S}=\mathbb{Z}_{\ge 0}$ of nonnegative integers. In our description above we started it from state $X_0=1$, but like any Markov chain, it can be started from any initial state, deterministic or random. When we compute in terms of the transition probability, we use the earlier notation P_k for path probabilities when the initial state is k.

Our goal is to understand the long term behavior, in particular, the possibility of extinction. Let

$$D=\{X_n=0 \text{ for some } n\ge 1\}$$

denote the extinction event and let

$$\pi=P_1(D)$$

denote the extinction probability of a process started from a single individual. The complement of D is the event of infinite survival.

There are two trivial cases. If $\beta_0=0$ then each individual always leaves at least one offspring, extinction never happens, and $\pi=0$. If $\beta_0=1$ then the process dies out immediately and $\pi = 1$. Hence let us assume in the sequel that

$$0 < \beta_0 < 1.$$

From any state k immediate extinction is possible:

$$p(k, 0) = \beta_0^k > 0.$$

This implies that *all states* $k > 0$ *are transient.* The only recurrent state is the absorbing state 0. Recall that each transient state is visited at most finitely many times. Thus for any k, there is a random time N such that $X_n \notin \{1, 2, \dots, k\}$ for all $n \geq N$. This implies a dichotomy: either extinction happens so that $X_m = X_{m+1} = X_{m+2} = \cdots = 0$ for some m, or $X_n \to \infty$. In probability terms, for any initial state k,

$$P_k(\text{extinction}) + P_k\big(\lim_{n\to\infty} X_n = \infty\big) = 1. \tag{2.148}$$

This already gives a nontrivial qualitative conclusion: infinite survival is not possible unless the population size tends to infinity. But qualitative considerations alone cannot decide the probabilities of the two alternatives in (2.148). So we compute.

Let

$$g(s) = \sum_{k=0}^{\infty} \beta_k s^k$$

denote the p.g.f. of the offspring distribution. Apply a first-step decomposition.

$$\begin{aligned}\pi = P_1(D) &= \sum_{k=0}^{\infty} P_1(X_1 = k, D) \overset{(a)}{=} p(1,0) + \sum_{k=1}^{\infty} p(1,k) P_k(D)\\ &\overset{(b)}{=} \beta_0 + \sum_{k=1}^{\infty} \beta_k \pi^k = \sum_{k=0}^{\infty} \beta_k \pi^k = g(\pi).\end{aligned} \tag{2.149}$$

- In step (a), either extinction happens immediately or there are $k \geq 1$ offspring and we restart the process in state k and wait for the extinction. A careful application of the Markov property goes as follows for $k \geq 1$:

$$\begin{aligned}&P_1(X_1 = k, D) = P_1(X_1 = k, \text{process } X_2, X_3, X_4, \dots \text{ goes extinct})\\ &= P_1(X_1 = k)\, P_1(\text{process } X_2, X_3, X_4, \dots \text{ goes extinct} \,|\, X_1 = k)\\ &= P_1(X_1 = k)\, P_k(\text{process } X_1, X_2, X_3, \dots \text{ goes extinct})\\ &= p(1,k) P_k(D).\end{aligned}$$

- In step (b) we expressed the transition probabilities in terms of the offspring distribution β. A less obvious observation is $P_k(D) = \pi^k$. A branching process started with k individuals consists of k independently evolving lines of descent, simply because all reproduction events happen independently of each other. The entire process goes extinct if and only if all k lines go extinct, and these extinction events happen independently of each other.

Equation (2.149) tells us that $\pi = g(\pi)$, in other words that π is a *fixed point* of the generating function. Unfortunately this does not always identify π uniquely, as the next example illustrates.

Example 2.116. Let $0 < p < 1$. Imagine a colony of cells. In one time step, each cell either splits into two new cells with probability p or dies without leaving descendants with probability $1-p$. Thus the offspring distribution is

$$\beta_0 = 1-p,\ \beta_2 = p, \text{ and } \beta_k = 0 \text{ for } k \notin \{0,2\}.$$

The p.g.f. is $g(s) = ps^2 + 1 - p$. The quadratic equation $g(s) = s$ has the following solutions:

- If $p = \frac{1}{2}$ then the only solution is $s = 1$.
- If $p \neq \frac{1}{2}$ then there are two solutions $s = \frac{1-p}{p}$ and $s = 1$.

Only solutions in $[0,1]$ can be probabilities. We get the following conclusions.

(i) If $0 < p \le \frac{1}{2}$ then extinction is certain ($\pi = 1$) because $s = 1$ is the only solution in $[0,1]$. In other words, if $p \le \frac{1}{2}$, the cell population is certain to eventually die out.

(ii) For $\frac{1}{2} < p < 1$ there are two solutions $\frac{1-p}{p}$ and 1 in the unit interval $[0,1]$. We need a sharper criterion to decide which one is π. △

The conclusive answer comes in the next theorem. Its proof utilizes properties of the p.g.f. established above.

Theorem 2.117. *The extinction probability π is the smallest solution of the equation $g(s) = s$ in the unit interval* $[0,1]$.

Proof. Since we already know that $g(\pi) = \pi$, we just need to show that if $g(\sigma) = \sigma$ and $0 \le \sigma \le 1$, then $\pi \le \sigma$.

The first observation is that

$$\pi = \lim_{n\to\infty} P_1(X_n = 0). \tag{2.150}$$

This follows from a decomposition according to the first extinct generation. The events $\{X_{m-1} > 0, X_m = 0\}$ are pairwise disjoint and their union is D. Hence

$$\begin{aligned}\pi = P_1(D) &= \sum_{m=1}^{\infty} P_1(X_{m-1} > 0, X_m = 0)\\ &= \lim_{n\to\infty} \sum_{m=1}^{n} P_1(X_{m-1} > 0, X_m = 0) = \lim_{n\to\infty} P_1(X_n = 0).\end{aligned}$$

The last equality comes from the disjoint union $\{X_n = 0\} = \bigcup_{m=1}^n \{X_{m-1} > 0, X_m = 0\}$.

The next point is that formula (2.143) for the p.g.f. of a random sums applies to (2.146) and gives the equation $g_{X_{n+1}}(s) = g_{X_n}(g(s))$. From this we can conclude inductively that

$$g_{X_n}(\sigma) = \sigma \quad \text{for all } n \ge 1. \tag{2.151}$$

Namely, since we start with a single individual, $g_{X_1}(\sigma) = g(\sigma) = \sigma$. Induction step: assuming that $g_{X_n}(\sigma) = \sigma$, we extend the identity to $n+1$:

$$g_{X_{n+1}}(\sigma) = g_{X_n}(g(\sigma)) = g_{X_n}(\sigma) = \sigma.$$

We are ready to complete the argument. By (2.141) and the monotonicity (2.140), for each n,

$$P_1(X_n = 0) = g_{X_n}(0) \le g_{X_n}(\sigma) = \sigma.$$

The inequality is preserved to the limit:

$$\pi = \lim_{n\to\infty} P_1(X_n = 0) \le \sigma.$$

The proof is complete. □

Remark 2.118 (The trivial cases)**.** Theorem 2.117 is in fact valid without the assumption $0 < \beta_0 < 1$. These are the two other cases.

- If $\beta_0 = 0$ then $g(0) = 0$ and $\pi = 0$ in agreement with the characterization of π as the smallest root of $g(s) = s$ in $[0, 1]$.
- If $\beta_0 = 1$ then $g(s) = 1$ for all s, the only solution of $g(s) = s$ is $s = 1$ which is the value of π. △

Example 2.119 (Continuation of Example 2.116)**.** Return to the branching process with offspring distribution $\beta_2 = p$ and $\beta_0 = 1 - p$ for some $\frac{1}{2} < p < 1$. Together with the calculation of Example 2.116, Theorem 2.117 implies that $\pi = \frac{1-p}{p}$. From (2.148) we conclude that

$$P_1\bigl(\lim_{n\to\infty} X_n = \infty\bigr) = 1 - \pi = \frac{2p-1}{1-p}.$$

Casting this in terms of the story of the cell colony, if the environment of the cells can be altered to favor dividing into two over dying ever so slightly, the colony has a chance to live forever, in which case it grows without bound. △

Beautiful as Theorem 2.117 is, it is somewhat impractical in that it requires knowledge of the entire offspring distribution in order to determine π. If instead of the exact value of π we settle for the cruder information of whether extinction is certain or not, we can extract from Theorem 2.117 a simpler criterion. Let

$$\mu = \sum_{k=1}^{\infty} k\beta_k \tag{2.152}$$

be the mean number of offspring. It is a well-defined (possible extended) real in $[0, \infty]$.

Infinite survival is certain if $\beta_0 = 0$ but not if $\beta_0 > 0$. The next theorem gives us a precise criterion for the possibility of extinction versus infinite survival: infinite survival can happen if and only if the mean number of offspring is strictly above 1. This leads to the following terminology: a branching process is

- *subcriticial* if $\mu < 1$,
- *criticial* if $\mu = 1$, and
- *supercriticial* if $\mu > 1$.

Theorem 2.120. *Suppose* $\beta_0 > 0$.

(a) *If* $\mu \le 1$ *then extinction is certain:* $\pi = 1$.

(b) *If* $\mu > 1$ *then* $0 < \pi < 1$.

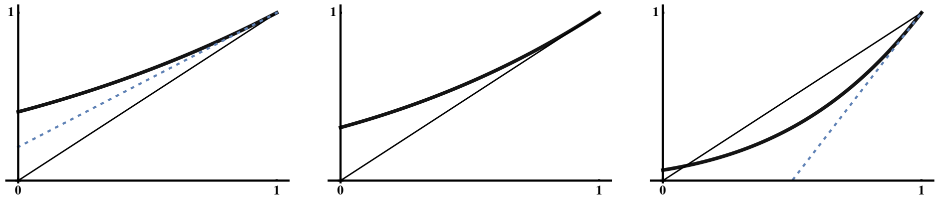

Figure 4. Examples of the generating function $g(s)$ in the cases $\mu < 1$, $\mu = 1$ and $\mu > 1$. The dotted line in the first and third graph is the tangent line to $g(s)$ at $s = 1-$, whose slope is $\mu = g'(1-)$.

Proof. We separate the case when $\beta_0 + \beta_1 = 1$. Then $g(s) = \beta_0 + (1 - \beta_0)s$ with $\beta_0 > 0$ by assumption, and hence $\mu < 1$. Solving $g(s) = s$ yields $s = 1$ as the only solution.

For the remainder of the proof assume that $\beta_0 + \beta_1 < 1$. Then $\beta_j > 0$ for some $j \geq 2$, and hence $g''(s) = \sum_{k=1}^{\infty} k(k-1)\beta_k s^k \geq j(j-1)\beta_j s^j > 0$ for $s \in (0, 1]$. (Still, $g''(1-)$ can be ∞.)

We use the following properties of a twice differentiable function f. Denote its tangent line at point c by $\ell_c(x) = f'(c)(x - c) + f(c)$. If $f''(x) > 0$ on the interval (a, b) then f is *strictly convex* in (a, b). This means that for any point $c \in (a, b)$, the tangent line at c lies *strictly below* $f(x)$ at all $x \in (a, b)$ except at $x = c$ where $\ell_c(c) = f(c)$. Furthermore, *any line* $\ell(x) = c_1 x + c_2$ can intersect the graph of f at most twice in (a, b).

The proof of this theorem can now be read off from Figure 4. The roots of $g(s) = s$ appear at points where the graph of g crosses the diagonal line. Since $g(0) = \beta_0 > 0$, $g(1) = 1$ and g is strictly convex, g crosses the diagonal at a point in $(0, 1)$ if and only if $g(s)$ approaches $g(1)$ from strictly *below* the diagonal. This happens if and only if the slope $g'(1-) > 1$, which is the same as $\mu > 1$. □

2.8. Technical appendix

Constructions of a Markov chain. Here is the construction of a Markov chain $\{X_k\}_{k \in \mathbb{Z}_{\geq 0}}$ with countable state space $\mathcal{S}$, transition probability p, and initial distribution μ.

(i) The sample space Ω is the space of $\mathcal{S}$-valued sequences:

$$\Omega = \{\omega = (\omega_i)_{i \in \mathbb{Z}_{\geq 0}} : \text{ each } \omega_i \in \mathcal{S}\}. \tag{2.153}$$

(ii) The probability measure P_μ does not have a formula for all possible events in Ω, but it can be uniquely determined by giving the probabilities of events that restrict finitely many coordinates: for any states $x_0, \dots, x_n \in \mathcal{S}$,

$$P_\mu\{\omega : \omega_0 = x_0, \dots, \omega_n = x_n\} = \mu(x_0)p(x_0, x_1) \cdots p(x_{n-1}, x_n). \tag{2.154}$$

The event $\{\omega : \omega_0 = x_0, \dots, \omega_n = x_n\}$ is the set of those sequences $\omega \in \Omega$ that are of the form $\omega = (x_0, x_1, \dots, x_n, \omega_{n+1}, \omega_{n+2}, \dots)$, where the coordinates $\omega_{n+1}, \omega_{n+2}, \dots$ that come after x_n are unrestricted.

(iii) Finally, the random variables X_k on Ω are defined as the *coordinate functions* $X_k : \Omega \to \mathcal{S}$ by

$$X_k(\omega) = \omega_k \quad \text{for the sample point } \omega = (\omega_0, \omega_1, \omega_2, \omega_3, \dots). \tag{2.155}$$

(iv) The class of events $\mathcal{F}$ that is part of the definition of the probability space $(\Omega, \mathcal{F}, P_\mu)$ cannot be described without measure theory. Precisely speaking, $\mathcal{F}$ is the *product σ-algebra* on Ω. It can be ignored for the purposes of this course.

The construction above is called the *canonical* one. There can be other constructions too. For example, the construction of the random walk in Example 2.1 first took a construction of the i.i.d. process $\{Y_k\}$ (which could have been the canonical one) and then constructed the random walk $\{S_n\}$ as a function of $\{Y_k\}$ via formula (2.7).

Derivations of extensions of the Markov property. The most general form of the Markov property, stated earlier as Theorem 2.21, appears as (2.160) below. The lemmas lead up to it by establishing successively more general statements.

Lemma 2.121. *Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$ and initial distribution μ. For any states $y_0, \dots, y_n$, $x_1, \dots, x_m$ from $\mathcal{S}$, and any initial distribution μ,*

$$\begin{aligned} &P_\mu\big[X_{n+m} = x_m, \dots, X_{n+2} = x_2, X_{n+1} = x_1 \\ &\qquad\qquad \,\big|\, X_n = y_n, X_{n-1} = y_{n-1}, \dots, X_0 = y_0\big] \\ &\quad = p(y_n, x_1)p(x_1, x_2) \cdots p(x_{m-1}, x_m) \\ &\quad = P_{y_n}\big[X_1 = x_1, X_2 = x_2, \dots, X_m = x_m\big]. \end{aligned} \tag{2.156}$$

provided the conditioning event has positive probability.

Proof. We use the definition of conditional probability and then equation (2.26):

$$\begin{aligned} &P_\mu\big[X_{n+m} = x_m, \dots, X_{n+2} = x_2, X_{n+1} = x_1 \,|\, X_n = y_n, X_{n-1} = y_{n-1}, \dots, X_0 = y_0\big] \\ &\quad = \frac{P_\mu\big[X_{n+m} = x_m, \dots, X_{n+1} = x_1, X_n = y_n, \dots, X_0 = y_0\big]}{P_\mu\big[X_n = y_n, \dots, X_0 = y_0\big]} \\ &\quad = \frac{\mu(y_0)p(y_0, y_1) \cdots p(y_{n-1}, y_n)p(y_n, x_1) \cdots p(x_{m-1}, x_m)}{\mu(y_0)p(y_0, y_1) \cdots p(y_{n-1}, y_n)} \\ &\quad = p(y_n, x_1)p(x_1, x_2) \cdots p(x_{m-1}, x_m) \\ &\quad = P_{y_n}\big[X_1 = x_1, X_2 = x_2, \dots, X_m = x_m\big]. \end{aligned}$$

The second last equality follows from cancelling, the last equality by (2.27). □

Eq. (2.156) remains valid if we add a condition on X_n inside the probability:

$$\begin{aligned} &P_\mu\big[X_{n+m} = x_m, \dots, X_{n+1} = x_1, X_n = x_0 \,|\, X_n = y_n, X_{n-1} = y_{n-1}, \dots, X_0 = y_0\big] \\ &\quad = P_{y_n}\big[X_0 = x_0, X_1 = x_1, \dots, X_m = x_m\big]. \end{aligned} \tag{2.157}$$

If $x_0 = y_n$ then adding or removing the condition $X_n = x_0$ does not affect the value of the first probability above, and adding or removing the condition $X_0 = x_0$ does not affect the value of the second probability. Then (2.157) reduces to (2.156). On the other hand, if $x_0 \neq y_n$, then both probabilities above are zero because X_n (and X_0) cannot simultaneously equal both x_0 and y_n.

Lemma 2.122. *Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$ and initial distribution μ. For any states $y_0, \dots, y_n$ from $\mathcal{S}$, any subset $U \subseteq \mathcal{S}^{m+1}$, and any initial distribution μ,*

$$\begin{aligned} &P_\mu\big[(X_n, X_{n+1}, \dots, X_{n+m}) \in U \,|\, X_n = y_n, X_{n-1} = y_{n-1}, \dots, X_0 = y_0\big] \\ &\qquad = P_{y_n}\big[(X_0, X_1, \dots, X_m) \in U\big] \end{aligned} \tag{2.158}$$

provided the conditioning event has positive probability.

Proof. This follows from property (2.157) by addition:

$$\begin{aligned} &P_\mu\big[(X_n, \dots, X_{n+m}) \in U \,|\, X_n = y_n, \dots, X_0 = y_0\big] \\ &\qquad = \sum_{(x_0,\dots,x_m)\in U} P_\mu\big[X_{n+m} = x_m, \dots, X_n = x_0 \,|\, X_n = y_n, \dots, X_0 = y_0\big] \\ &\qquad = \sum_{(x_0,\dots,x_m)\in U} P_{y_n}\big[X_0 = x_0, \dots, X_m = x_m\big] \\ &\qquad = P_{y_n}\big[(X_0, \dots, X_m) \in U\big]. \end{aligned}$$

□

Lemma 2.123. *Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$ and initial distribution μ. For any states $y_0, \dots, y_n$ from $\mathcal{S}$, any event $U \subseteq \mathcal{S}^{\mathbb{Z}_{\geq 0}}$, and any initial distribution μ,*

$$\begin{aligned} &P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U \,|\, X_n = y_n, X_{n-1} = y_{n-1}, \dots, X_0 = y_0\big] \\ &\qquad = P_{y_n}\big[(X_0, X_1, X_2, \dots) \in U\big] \end{aligned} \tag{2.159}$$

provided the conditioning event has positive probability.

Proof. This follows from (2.158) with a little leap of faith (let $m \to \infty$). Rigorous justification needs measure theory. □

Finally the most general formula where we add an arbitrary event to the conditioning side.

Theorem 2.124. *Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$ and initial distribution μ. Let $x \in \mathcal{S}$, $B \subseteq \mathcal{S}^{n+1}$ and $U \subseteq \mathcal{S}^{\mathbb{Z}_{\geq 0}}$. Then for any initial distribution μ,*

$$\begin{aligned} &P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U \;\big|\; X_n = x, (X_0, \dots, X_n) \in B\big] \\ &\qquad = P_x\big[(X_0, X_1, X_2, \dots) \in U\big] \end{aligned} \tag{2.160}$$

provided the conditioning event has positive probability.

Proof. Let B_0 be the set of $(n+1)$-tuples $(y_0, \dots, y_n)$ in B that satisfy $y_n = x$. Then

$$\big\{X_n = x, (X_0, \dots, X_n) \in B\big\} = \big\{(X_0, \dots, X_n) \in B_0\big\}.$$

The next calculation uses first the definition of conditional probability, then breaks up the probability into a sum, then applies the product rule. These steps lead to conditional probabilities to which we can directly apply (2.159).

$$
\begin{aligned}
&P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U \;\big|\; X_n = x, (X_0, \dots, X_n) \in B\big] \\
&\quad = \frac{P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U, X_n = x, (X_0, \dots, X_n) \in B\big]}{P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]} \\
&\quad = \frac{\sum_{(y_0,\dots,y_n)\in B_0} P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U, (X_0, \dots, X_n) = (y_0, \dots, y_n)\big]}{P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]} \\
&\quad = \Big\{ \sum_{(y_0,\dots,y_n)\in B_0} P_\mu\big[(X_n, X_{n+1}, X_{n+2}, \dots) \in U \;\big|\; (X_0, \dots, X_n) = (y_0, \dots, y_n)\big] \\
&\qquad\qquad \cdot P_\mu\big[(X_0, \dots, X_n) = (y_0, \dots, y_n)\big] \Big\} \\
&\qquad\qquad\qquad \cdot \frac{1}{P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]} \\
&\quad \overset{(2.159)}{=} \Big\{ \sum_{(y_0,\dots,y_n)\in B_0} P_x\big[(X_0, X_1, X_2, \dots) \in U\big] \\
&\qquad\qquad \cdot P_\mu\big[(X_0, \dots, X_n) = (y_0, \dots, y_n)\big] \Big\} \\
&\qquad\qquad\qquad \cdot \frac{1}{P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]} \\
&\quad = \frac{P_x\big[(X_0, X_1, X_2, \dots) \in U\big] \cdot P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]}{P_\mu\big[X_n = x, (X_0, \dots, X_n) \in B\big]}.
\end{aligned}
$$

Cancelling above leaves the right-hand side of (2.36). □

Exercises

Exercise 2.1 (General random walk on the integers)**.** Let $\{Y_k\}_{k\in\mathbb{Z}_{>0}}$ be an i.i.d. process with state space $\mathcal{S} = \mathbb{Z}$ (the set of integers) with common probability mass function $p_Y(y) = P(Y_k = y)$ for all indices k and integer points y. Define the *random walk with step distribution* p_Y as the following process: $S_0 = 0$, and for $n \geq 1$, $S_n = Y_1 + \cdots + Y_n$. Show that S_n is a Markov chain and identify its transition probability.

Exercise 2.2. Let us say that a a stochastic process $\{X_n\}_{n\geq 0}$ is periodic with period k if the process satisfies $P(X_n = X_{n+k}\ \forall\, n \geq 0) = 1$. In both parts below verify what you claim.

(a) Create an example of a stochastic process $\{X_n\}_{n\geq 0}$ that is periodic with period 3 and is a Markov chain.

(b) Create an example of a stochastic process $\{X_n\}_{n\geq 0}$ that is periodic with period 4 and is not a Markov chain.

Exercise 2.3. Prove the Chapman-Kolmogorov equations (2.32) from the definition of $p^{(k)}(x,y)$ given in (2.30)–(2.31).

Exercise 2.4. Prove Theorem 2.35 by using the basic Markov property in the form (2.36), *without* using the strong Markov property.

Hint. You might start by decomposing the event $\{T_y^{k+1} < \infty\}$ according to the value of T_y^1 or T_y^k, depending on how you wish to organize the induction proof.

Exercise 2.5. Repeat a sequence of independent trials with success probability p and use the encoding $0 =$ failure and $1 =$ success. Calculate the probability that immediately after the second success we see the sequence 00110.

The immediate intuitive answer is $p^2(1-p)^3$ because the past trials do not influence the future ones. The point of the exercise is to perform a rigorous derivation. There are at least three ways to do the calculation. There are two different Markov chains to which you can apply the strong Markov property. Or you can do an elementary calculation without Markov techniques utilizing the independence of the trials.

Exercise 2.6. Prove that if y is recurrent, then for any initial state x,

$$P_x(T_y^1 < \infty) = P_x(T_y^k < \infty \text{ for all } k \geq 1).$$

Exercise 2.7. Prove that $\rho_{xz} \geq \rho_{xy}\rho_{yz}$ for all states x, y, z.

Exercise 2.8. Referring to Definition 2.44, prove that if $x \longrightarrow y$ and $y \longrightarrow z$, then $x \longrightarrow z$.

Exercise 2.9. Consider the transition matrix $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ on the state space $\{0, 1\}$. Let μ be an arbitrary initial distribution. Under what condition does the limit

$$\lim_{n\to\infty} P_\mu(X_n = 0, X_{n+1} = 1, X_{n+2} = 0)$$

exist and what is then its value?

Exercise 2.10. Let a, b be two states of an irreducible, recurrent Markov chain such that $p(a,b) > 0$. Show that the probability that the process never makes the jump from a to b is zero.

Exercise 2.11. Consider the success run chain with transitions $p(x, x+1) = \alpha$ and $p(x, 0) = 1 - \alpha$ for all states $x \in \{0, 1, 2, \dots\}$, where $0 < \alpha < 1$ is a fixed parameter.

(a) Calculate $E_0[T_k]$, that is, the expected time to get to a success run of length k when you start from zero. For an easier version of the exercise, do $k = 2$ or $k = 3$.
Hint. You can change the rules so that k is an absorbing state.

(b) Suppose you are at k. What is the expected time until you reach $k + 1$?

Exercise 2.12. Find an equation for S_n in (2.69) in terms of the blocks $\mathbf{P}_{\mathcal{R}}$, S and Q.

Exercise 2.13. Let Y_n be a Markov chain with transition probability $\mathbf{P}$. Assume there are no absorbing states. Let $0 = \sigma_0 < \sigma_1 < \sigma_2 < \cdots$ be the random times at which Y_n moves to a different state. That is, for each $k \geq 1$, $\sigma_k = \min\{n > \sigma_{k-1} : Y_n \neq Y_{n-1}\}$. Let $Z_k = Y_{\sigma_k}$ for $k \geq 0$. In other words, the process $\{Z_k\}$ records the moves of $\{Y_n\}$ to a state different from the previous one.

(a) Derive the finite-dimensional probability

$$P(Z_0 = x_0, Z_1 = x_1, \ldots, Z_n = x_n)$$

for any states $x_0, \ldots, x_n$.

(b) Show that Z_k is a Markov chain and find its transition probability $R = \{r(x, y) : x, y \in \mathcal{S}\}$.

Exercise 2.14. Suppose that p is a transition probability function on a countable state space $\mathcal{S}$. Suppose that on some probability space we have independent random variables $X_0, Y_1, Y_2, \ldots$ such that $X_0 \in \mathcal{S}$ and $Y_k \sim \text{Unif}[0, 1]$. Find a function $F : \mathcal{S} \times [0, 1] \to \mathcal{S}$ so that the recursion

$$X_{n+1} = F(X_n, Y_n) \qquad n \geq 0$$

defines a stochastic process X_n that is a Markov chain with transition probability p.

Hint. By Lemma 2.14 you have to construct a function F that satisfies $p(x, y) = P(F(x, Y_1) = y)$ for $x, y \in \mathcal{S}$.

Exercise 2.15. Assume that the set $\mathcal{T}$ of transient states is finite. Let

$$T = \inf\{n \geq 0 : X_n \in \mathcal{R}\}$$

denote the time of the first entrance into the set $\mathcal{R}$ of recurrent states. This exercise proves that $E_x[T] < \infty$ for all $x \in \mathcal{T}$.

(a) Explain why the set $\mathcal{R}$ must be nonempty.

(b) Derive the identity

$$P_x(T > n) = \sum_{y \in \mathcal{T}} (Q^n)_{x,y}$$

for $x \in \mathcal{T}$ and $n \geq 0$.

(c) The formula

$$E_x[T] = \sum_{n=0}^{\infty} P_x(T > n)$$

is valid for all nonnegative random variables, including those with infinite expectation or even infinite values. (See Lemma B.4 in Appendix B.) Use this formula and Lemma 2.55 to prove that $E_x[T] < \infty$ and to give an alternative derivation of equation (2.72): $\mathbf{m} = (I - Q)^{-1}\mathbf{1}$.

Exercise 2.16. Let π be a reversible distribution for a transition probability $\mathbf{P} = \{p(x, y)\}_{x,y \in \mathcal{S}}$. Show that for any states $x_0, \ldots, x_n$,

$$\begin{aligned} &P_\pi(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ &\qquad = P_\pi(X_n = x_0, X_{n-1} = x_1, \ldots, X_0 = x_n). \end{aligned}$$

Hint. Express the right-hand side as

$$\begin{aligned}&P_\pi(X_n = x_0, X_{n-1} = x_1, \dots, X_0 = x_n)\\&\qquad = \pi(x_n)\, p(x_n, x_{n-1})\, p(x_{n-1}, x_{n-2}) \cdots p(x_1, x_0).\end{aligned}$$

Beginning from the left, apply detailed balance to each successive pair of states.

Exercise 2.17. Assume $\mathbf{P}$ is doubly stochastic.

(a) Give an example where the state space is infinite and there is an invariant probability distribution.
Hint. Irreducibility is not assumed.

(b) Show that if $\mathbf{P}$ is irreducible and doubly stochastic, then the existence of an invariant probability distribution implies that the state space must be finite.

Exercise 2.18. Marvin's employer has a program to help employees quit smoking. After an employee has been smoke-free for three consecutive days, every subsequent consecutive smoke-free day earns one point towards a prize. Points are cumulative and are never lost. After a smoking day, another three consecutive smoke-free days have to come before more points are collected.

For Marvin, if today is smoke-free then so is tomorrow with probability $\frac{3}{4}$. If Marvin smokes today, he smokes tomorrow with probability $\frac{1}{2}$.

Let A_n be the number of points Marvin has collected in the first n days. Find the limiting rate $\lim_{n\to\infty} A_n/n$ in two different ways.

(a) Model the situation with a Markov chain.

(b) Model the situation with a renewal process.

(The numerical answer is $\frac{9}{32}$.)

Random walk.

Exercise 2.19. Fix $0 < p < \frac{1}{2}$. Let $S_0 = 0$, $S_n = Y_1 + \cdots + Y_n$ be the asymmetric simple random walk with i.i.d. steps $\{Y_k\}$ with marginal distribution $P(Y_k = 1) = p = 1 - P(Y_n = -1)$. Let $Z = \max_{n\ge 0} S_n$ be the maximum level that the walk reaches over all time.

(a) Show that Z is a finite, nonnegative random variable. **Hint.** The SLLN informs us about the behavior of S_n as $n \to \infty$.

(b) Find the probability mass function of Z. **Hint.** Condition on the first step Y_1. Note that $S'_n = Y_2 + \cdots + Y_{n+1}$ is a random walk with the same distribution as S_n but independent of Y_1.

Branching processes.

Exercise 2.20. Let X_n be the branching process constructed through formula (2.146) from the initial state $X_0 = 1$. Prove that X_n is a Markov chain with transition probability given by equation (2.147).

Exercise 2.21. Consider a branching process whose offspring distribution satisfies $\beta(0) = 0$ and $\beta(1) < 1$. Show that all states $k \ge 1$ are transient.

Exercise 2.22. Let $0 < p < 1$. Consider a branching process whose offspring distribution satisfies $\beta_0 = 1 - p$ and $\beta_3 = p$. Find the extinction probability.

Exercise 2.23. Let $0 < p < 1$. Consider a population of organisms whose lifecycle goes as follows. A newborn individual has probability p of reaching adulthood. Once an adult, the individual gives birth to exactly two offspring, and then dies. Start with a single individual. Find the probability that this population eventually goes extinct.

Exercise 2.24. Consider a branching process with offspring distribution $\{\beta_k\}_{k\geq 0}$ and assume $0 < \beta_0 < 1$. Suppose the evolution is altered as follows. Start with $X_0 = 1$, a single individual. If the next generation is not empty (that is, if $X_1 > 0$), they invite an immigrant to join them (that is, X_1 is replaced by X_1+1). After that the evolution continues according to the usual rule. Find the extinction probability $\widehat{\pi}$ of this new process. Express it in terms of β_0 and the original extinction probability π. Do we have $\widehat{\pi} < \pi$ for all $0 < \pi < 1$, in other words, does the immigrant strictly improve chances of survival for all values $0 < \pi < 1$? Is it possible that $\widehat{\pi} < \pi = 1$?

Analytic exercises.

Exercise 2.25. Let $\mathcal{S}$ be an at most countable state space. Let $\mu^{(n)}$ and π be probability measures on $\mathcal{S}$ and assume that $\mu^{(n)}(x) \to \pi(x)$ as $n \to \infty$, for each $x \in \mathcal{S}$. Show that for each $y \in \mathcal{S}$, $(\mu^{(n)}\mathbf{P})_y \to (\pi\mathbf{P})_y$ as $n \to \infty$.

Hint. The pointwise convergence $\mu^{(n)}(x) \to \pi(x)$ implies the stronger convergence $\sum_x |\mu^{(n)}(x) - \pi(x)| \to 0$, because $\mu^{(n)}$ and π are probability measures.

Exercise 2.26. Derive rigorously the limit $p^{(n)}(0,y) \to \frac{2}{3}\pi(y)$ in Example 2.104 for $y \in \{2,3\}$, where $\pi = [\pi(2)\ \pi(3)] = \left[\, \frac{8}{17}\ \ \frac{9}{17} \,\right]$. This means satisfying the definition of a limit: given $\varepsilon > 0$, show that there exists an integer n_0 such that

$$|\, p^{(n)}(0,y) - \tfrac{2}{3}\pi(y) \,| \leq \varepsilon \quad \text{for } n \geq n_0.$$

Chapter 3

Martingales

3.1. Introduction

Imagine that you are placing bets on fair coin flips. Before each coin flip you get to decide the amount of money that you bet on that coin flip. Your decision can be influenced by the outcomes of the previous coin flips. In other words, you are allowed to employ a *betting strategy* that utilizes the information that you have collected about the previous games.

To turn this into a stochastic process, denote your net winnings after the nth round by M_n. M_n is a random variable that can depend on the outcomes of the first n coin flips, because the bets were allowed to depend on the previous flips. The question is whether there can be a betting strategy that gives you a strictly positive expected profit.

If you simply bet \$1 on each flip then we know the outcome: M_n is a symmetric simple random walk and the expected profit is always zero. Can you do better with a clever strategy? Since the coin flips are fair and independent, it is reasonable to believe that even when you use the information of all the previous outcomes, the expected change in your wealth is zero after each coin flip.

Mathematically this means that, given the outcomes of the first n coin flips, the conditional expectation of M_{n+1} is equal to M_n. We give a precise proof of this below. This is called the *martingale property.* In this section we show how the martingale property allows us to compute various interesting quantities related to the process.

From another perspective, a martingale is a generalization of symmetric simple random walk (or more generally, of any random walk with mean zero steps). In a mean zero random walk each step is a mean zero random variable, independent of the previous steps. In a martingale this property is generalized so that the next step always has zero conditional expectation, given everything that happened in the past. Although independence is lost, the resulting process still has a number of interesting and tractable properties.

The mathematics of martingales originates in gambling. Nowadays it is a central part of the theory of probability and important for a variety of applications, in particular financial mathematics.

To prepare for martingales we start with a brief review of conditional expectation in the discrete setting. In this chapter all random variables are discrete, unless otherwise stated. For a more thorough treatment of conditional expectations at the level of introductory probability, consult the textbook [ASV18].

3.2. Review of conditional expectation

Let X be a discrete random variable and $\mathbf{Y} = (Y_1, \dots, Y_k)$ a discrete random vector, defined together on some probability space. The events $\{\mathbf{Y} = \mathbf{y}\} = \{Y_1 = y_1, \dots, Y_k = y_k\}$ form a *partition* of the sample space, as the vector $\mathbf{y} = (y_1, \dots, y_k)$ ranges over the possible values of $\mathbf{Y}$. This means that the events $\{\mathbf{Y} = \mathbf{y}\}$ are pairwise disjoint and their union is the entire sample space:

$$\Omega = \bigcup_{\mathbf{y}} \{\mathbf{Y} = \mathbf{y}\}.$$

We can condition X on those events $\{\mathbf{Y} = \mathbf{y}\}$ of this partition that have positive probability. This gives the conditional probability mass function and conditional expectation of X given $\mathbf{Y} = \mathbf{y}$, whenever $P(\mathbf{Y} = \mathbf{y}) > 0$:

$$P(X = x | \mathbf{Y} = \mathbf{y}) = \frac{P(X = x, \mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} \tag{3.1}$$

$$\text{and} \qquad E[X | \mathbf{Y} = \mathbf{y}] = \sum_x x P(X = x | \mathbf{Y} = \mathbf{y}). \tag{3.2}$$

The conditional expectation is a well-defined finite number if the expectation of X is finite. We make this assumption throughout the section. (Exercise 3.1 gives a hint for verifying this.)

When we calculate the conditional expectation of a function of a random variable, or of several random variables, formula (3.2) extends in a familiar way. For example, suppose $g : \mathbb{R}^m \to \mathbb{R}$ is a function of m variables and $\mathbf{Z} = (Z_1, \dots, Z_m)$ is a discrete random vector. Then

$$E[g(\mathbf{Z}) | \mathbf{Y} = \mathbf{y}] = \sum_{\mathbf{z}} g(\mathbf{z}) P(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}), \tag{3.3}$$

where $\mathbf{z} = (z_1, \dots, z_m)$ ranges over the possible values of $\mathbf{Z}$ and we assume that $g(\mathbf{Z})$ has a finite expectation.

We can regard the the conditional expectation $E[X | \mathbf{Y} = \mathbf{y}]$ as a function of the variable $\mathbf{y}$. Denote it temporarily by $H(\mathbf{y}) = E[X | \mathbf{Y} = \mathbf{y}]$. At first this function is defined only for those $\mathbf{y}$ that satisfy $P(\mathbf{Y} = \mathbf{y}) > 0$. This does not hamper its utility, but if necessary, one can extend the definition by setting $H(\mathbf{y}) = 0$ for those $\mathbf{y}$ that satisfy $P(\mathbf{Y} = \mathbf{y}) = 0$.

Substituting the random variable $\mathbf{Y}$ into $H(\cdot)$ then defines a random variable $H(\mathbf{Y})$ that is called the *conditional expectation of* X *given* Y. The notation for it is $E[X | \mathbf{Y}]$. Here is the precise definition restated.

Definition 3.1. The **conditional expectation of X given Y** is the random variable defined by

$$E[X|\mathbf{Y}] = H(\mathbf{Y}) \tag{3.4}$$

where the function H is defined by $H(\mathbf{y}) = E[X|\mathbf{Y} = \mathbf{y}]$ for those values $\mathbf{y}$ that satisfy $P(\mathbf{Y} = \mathbf{y}) > 0$. △

Formula (3.4) reveals why it does not matter whether $H(\mathbf{y})$ is defined when $P(\mathbf{Y} = \mathbf{y}) = 0$: these values do not appear on the right-hand side of (3.4).

The conditional expectation $E[X|\mathbf{Y}]$ is a random variable which is a function of $\mathbf{Y}$. Given knowledge of $\mathbf{Y}$, $E[X|\mathbf{Y}]$ is the *best predictor* of X in the sense that it minimizes the mean square error. For this, see Theorem 10.34 in Section 10.4 of the textbook [**ASV18**].

Example 3.2. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent and $S = X + Y$. Find the conditional expectation $E[\,X \,|\, S\,]$.

Start by finding the necessary probability mass functions. For $n \geq 0$, decomposing the event $\{S = n\}$ into the different cases for the values of X and Y:

$$\begin{aligned} P(S = n) &= \sum_{k=0}^{n} P(X = k, Y = n - k) = \sum_{k=0}^{n} P(X = k)P(Y = n - k) \\ &= \sum_{k=0}^{n} e^{-\lambda}\frac{\lambda^k}{k!}e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} = \frac{e^{-(\lambda+\mu)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda^k\mu^{n-k} \\ &= e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}. \end{aligned}$$

We (re)proved the (possibly familiar) fact that $S \sim \text{Poisson}(\lambda + \mu)$.

Since $0 \leq X \leq S$, we only need to consider $0 \leq k \leq \ell$ in the probability below:

$$\begin{aligned} P(X = k, S = \ell) &= P(X = k, X + Y = \ell) = P(X = k, Y = \ell - k) \\ &= P(X = k)P(Y = \ell - k) = \tfrac{\lambda^k}{k!}e^{-\lambda}\tfrac{\mu^{\ell-k}}{(\ell-k)!}e^{-\mu}. \end{aligned}$$

Last, the conditional probability mass function: for $0 \leq k \leq \ell$,

$$\begin{aligned} P(X = k \,|\, S = \ell) &= \frac{P(X = k, S = \ell)}{P(S = \ell)} = \frac{\frac{\lambda^k}{k!}e^{-\lambda}\frac{\mu^{\ell-k}}{(\ell-k)!}e^{-\mu}}{\frac{(\lambda+\mu)^\ell}{\ell!}e^{-(\lambda+\mu)}} = \frac{\lambda^k\mu^{\ell-k}}{(\lambda+\mu)^\ell}\frac{\ell!}{k!(\ell-k)!} \\ &= \binom{\ell}{k}\left(\tfrac{\lambda}{\lambda+\mu}\right)^k\left(\tfrac{\mu}{\lambda+\mu}\right)^{\ell-k}. \end{aligned}$$

This says that, given $S = \ell$, the conditional distribution of X is $\text{Bin}(\ell, \frac{\lambda}{\lambda+\mu})$. Recalling that the mean of $\text{Bin}(n, p)$ equals np, we can write down the conditional expectations without further computation: $E[X|S = \ell] = \ell\frac{\lambda}{\lambda+\mu}$ and thus $E[X|S] = \frac{\lambda}{\lambda+\mu}S$. △

Example 3.3 (Markov chain)**.** Consider a Markov chain on a countable state space $\mathcal{S}$ and a function $f : \mathcal{S} \to \mathbb{R}$. The initial distribution μ is arbitrary. Assume that f is bounded so that the expectations below are well-defined. (This means that for some constant c, $|f(x)| \leq c$ for all states $x \in \mathcal{S}$.) Let $n \geq 0$. We take $f(X_{n+1})$

as the random variable to be conditioned, and the conditioning vector is the past evolution $\mathbf{Y} = (X_0, \dots, X_n)$. The Markov property gives us

$$\begin{aligned}
&E_\mu[f(X_{n+1}) \,|\, (X_0, \dots, X_n) = (x_0, \dots, x_n)] \\
&= \sum_y f(y) P_\mu(X_{n+1} = y \,|\, (X_0, \dots, X_n) = (x_0, \dots, x_n)) \\
&= \sum_y p(x_n, y) f(y) = E_{x_n}[f(X_1)].
\end{aligned}$$

Hence

$$E_\mu[f(X_{n+1}) \,|\, X_0, \dots, X_n] = E_{X_n}[f(X_1)]. \tag{3.5}$$

In the last formula we substituted the random variable X_n for x in the formula $E_x[f(X_1)]$. △

The following key property of the conditional expectation is sometimes called the law of total expectation.

Theorem 3.4. *The expectation of the conditional expectation gives back the original expectation:*

$$E[E[X|\mathbf{Y}]] = E[X]. \tag{3.6}$$

Proof. For the proof, it is convenient to use the functional notation $E[X|\mathbf{Y}] = H(\mathbf{Y})$.

$$\begin{aligned}
E[E[X|\mathbf{Y}]] = E[H(\mathbf{Y})] &= \sum_{\mathbf{y}} H(\mathbf{y}) P(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{y}} E[X|\mathbf{Y} = \mathbf{y}] P(\mathbf{Y} = \mathbf{y}) \\
&= \sum_{\mathbf{y}} \sum_x x\, P(X = x|\mathbf{Y} = \mathbf{y}) P(\mathbf{Y} = \mathbf{y}) \\
&= \sum_{\mathbf{y}} \sum_x x\, \frac{P(X = x, \mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} P(\mathbf{Y} = \mathbf{y}) \\
&= \sum_x x \sum_{\mathbf{y}} P(X = x, \mathbf{Y} = \mathbf{y}) = \sum_x x P(X = x) = E[X].
\end{aligned}$$

From the beginning, the summation can be restricted to those $\mathbf{y}$ that satisfy $P(\mathbf{Y} = \mathbf{y}) > 0$. Then the denominator does not cause any trouble. □

We illustrate equation (3.6) in the two examples.

Example 3.5. Remembering that the mean of a Poisson(λ) random variable is λ, we can check that (3.6) holds:

$$E\big[\, E[X|S] \,\big] = \tfrac{\lambda}{\lambda+\mu} E[S] = \tfrac{\lambda}{\lambda+\mu} \cdot (\lambda + \mu) = \lambda = EX.$$

△

Example 3.6 (Markov chain)**.** In the calculation below, identify $p(x, y)$ as the conditional probability $P_\mu(X_{n+1} = y \,|\, X_n = x)$, as stipulated by the Markov property.

The first step is the outcome from (3.5).

$$\begin{aligned}
&E_\mu\big[\,E_\mu[f(X_{n+1})\,|\,X_0,\dots,X_n]\,\big] = E_\mu\big[\,E_{X_n}[f(X_1)]\,\big]\\
&\quad= \sum_x P_\mu(X_n = x)E_x[f(X_1)]\\
&\quad= \sum_x P_\mu(X_n = x)\sum_y p(x,y)f(y)\\
&\quad= \sum_y f(y)\sum_x P_\mu(X_n = x)P_\mu(X_{n+1} = y\,|\,X_n = x)\\
&\quad= \sum_y f(y)\sum_x P_\mu(X_n = x, X_{n+1} = y) = \sum_y f(y)P_\mu(X_{n+1} = y)\\
&\quad= E_\mu[f(X_{n+1})].
\end{aligned}$$

As expected, equation (3.6) checks out again. △

The conditional expectation (whether we mean $E[\cdot\,|\mathbf{Y} = \mathbf{y}]$ or $E[\cdot\,|\mathbf{Y}]$) possesses many of the properties of the regular expectation. In particular it is linear: for any deterministic values a_1, a_2 and random variables X_1, X_2 we have

$$E[a_1X_1 + a_2X_2|\mathbf{Y}] = a_1E[X_1|\mathbf{Y}] + a_2E[X_2|\mathbf{Y}]. \tag{3.7}$$

The next theorem collects several properties of the conditional expectation that are useful for calculations. In part(a) the independence of X and $\mathbf{Y}$ implies that the value of $\mathbf{Y}$ does not carry any information about X, and consequently the conditional expectation equals the ordinary expectation. Parts (b) and (c) are manifestations of the fact that when we condition on $\mathbf{Y}$, in a sense $\mathbf{Y}$ ceases to be random, and it can be treated as a constant in the conditional expectation.

Theorem 3.7. *Let X be a discrete random variable and $\mathbf{Y} = (Y_1, \dots, Y_k)$ a discrete random vector. Let $g : \mathbb{R}^k \to \mathbb{R}$ be a function.*

(a) *If X and $\mathbf{Y}$ are independent, then $E[X|\mathbf{Y}] = E[X]$. In other words, the value of the random variable $E[X|\mathbf{Y}]$ is simply the constant $E[X]$.*

(b) *Assume that $g(\mathbf{Y})$ has a finite expectation. Then*

$$E[g(\mathbf{Y})|\mathbf{Y}] = g(\mathbf{Y}). \tag{3.8}$$

(c) *Assume that $Xg(\mathbf{Y})$ has a finite expectation. Then*

$$E[Xg(\mathbf{Y})|\mathbf{Y}] = g(\mathbf{Y})E[X|\mathbf{Y}]. \tag{3.9}$$

Proof. Part (a). For each $\mathbf{y}$ with $P(\mathbf{Y} = \mathbf{y}) > 0$ the conditional probability mass function of X is the same as the unconditional one by independence:

$$P(X = x|\mathbf{Y} = \mathbf{y}) = \frac{P(X = x, \mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} = \frac{P(X = x)\,P(\mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} = P(X = x).$$

Substituting this into equation (3.2) shows that $H(\mathbf{y}) = E[X|\mathbf{Y} = \mathbf{y}] = E[X]$ (constant function), and then $E[X|\mathbf{Y}] = H(\mathbf{Y}) = E[X]$.

Part (b) is a special case of part (c) where $X = 1$.

Part (c). First compute with formula (3.3). In the sum below a ranges over the values of X and $\mathbf{b}$ over the values of $\mathbf{Y}$.

$$E[Xg(\mathbf{Y})|\mathbf{Y}=\mathbf{y}] = \sum_{a,\mathbf{b}} ag(\mathbf{b})P(X=a, \mathbf{Y}=\mathbf{b}\,|\,\mathbf{Y}=\mathbf{y})$$

Evaluate the conditional probability:

$$\begin{aligned} P(X=a, \mathbf{Y}=\mathbf{b}\,|\,\mathbf{Y}=\mathbf{y}) &= \frac{P(X=a, \mathbf{Y}=\mathbf{b}, \mathbf{Y}=\mathbf{y})}{P(\mathbf{Y}=\mathbf{y})} \\ &= \begin{cases} 0, & \mathbf{b}\neq\mathbf{y} \\ \dfrac{P(X=a, \mathbf{Y}=\mathbf{y})}{P(\mathbf{Y}=\mathbf{y})} = P(X=a\,|\,\mathbf{Y}=\mathbf{y}), & \mathbf{b}=\mathbf{y}. \end{cases} \end{aligned}$$

Substitute this back above. The only nonzero term in the sum is the one for $\mathbf{b}=\mathbf{y}$.

$$\begin{aligned} E[Xg(\mathbf{Y})|\mathbf{Y}=\mathbf{y}] &= \sum_a ag(\mathbf{y})P(X=a\,|\,\mathbf{Y}=\mathbf{y}) \\ &= g(\mathbf{y})\sum_a a\,P(X=a\,|\,\mathbf{Y}=\mathbf{y}) = g(\mathbf{y})E[X\,|\,\mathbf{Y}=\mathbf{y}]. \end{aligned} \tag{3.10}$$

Equation (3.9) comes by substituting $\mathbf{Y}$ for $\mathbf{y}$. □

We illustrate these properties with random walk.

Example 3.8 (Symmetric simple random walk)**.** Let $\{X_n : n\geq 1\}$ be i.i.d. random variables with

$$P(X_n=1)=P(X_n=-1)=\tfrac{1}{2}$$

and $S_n=\sum_{k=1}^n X_k$ for $n\geq 1$. Find $E[S_n|S_m]$ and $E[S_n^2|S_m]$ for $1\leq m<n$.

If $1\leq m<n$ then $S_n = S_m+\sum_{k=m+1}^n X_k$, and S_m and $\sum_{k=m+1}^n X_k$ are independent. Then

$$E[S_n|S_m] = E\biggl[\,S_m+\sum_{k=m+1}^n X_k\,\bigg|\,S_m\biggr] = E[S_m|S_m]+E\biggl[\,\sum_{k=m+1}^n X_k\,\bigg|\,S_m\biggr].$$

We have $E[S_m|S_m]=S_m$ by (3.8). Since S_m and $\sum_{k=m+1}^n X_k$ are independent we have

$$E\biggl[\,\sum_{k=m+1}^n X_k\,\bigg|\,S_m\biggr] = E\biggl[\,\sum_{k=m+1}^n X_k\biggr] = 0.$$

This gives $E[S_n|S_m]=S_m$.

Similarly,

$$\begin{aligned} E[S_n^2|S_m] &= E\biggl[\biggl(S_m+\sum_{k=m+1}^n X_k\biggr)^2\,\bigg|\,S_m\biggr] \\ &= E[S_m^2|S_m]+E\biggl[2S_m\cdot\sum_{k=m+1}^n X_k\,\bigg|\,S_m\biggr]+E\biggl[\biggl(\sum_{k=m+1}^n X_k\biggr)^2\,\bigg|\,S_m\biggr]. \end{aligned}$$

We have $E[S_m^2|S_m] = S_m^2$ by (3.8). Using (3.9) and the independence of S_m and $\sum_{k=m+1}^n X_k$ we also have

$$\begin{aligned} E\Big[2S_m \cdot \sum_{k=m+1}^n X_k \,\Big|\, S_m\Big] &= 2S_m E\Big[\sum_{k=m+1}^n X_k \,\Big|\, S_m\Big] \\ &= 2S_m E\Big[\sum_{k=m+1}^n X_k\Big] = 0 \end{aligned}$$

and

$$E\bigg[\Big(\sum_{k=m+1}^n X_k\Big)^2 \,\bigg|\, S_m\bigg] = E\bigg[\Big(\sum_{k=m+1}^n X_k\Big)^2\bigg] = \mathrm{Var}\Big[\sum_{k=m+1}^n X_k\Big] = n - m.$$

Hence

$$E[S_n^2|S_m] = S_m^2 + n - m.$$

△

We record one more general property. The theorem below states that if $\mathbf{Z}$ is independent of $\mathbf{Y}$, then the conditional expectation of $g(\mathbf{Y}, \mathbf{Z})$ given $\mathbf{Y}$ can be computed by treating $\mathbf{Y}$ as a constant and taking expectation over $\mathbf{Z}$.

Theorem 3.9. *Let $\mathbf{Y} = (Y_1, \dots, Y_k)$ be a discrete random vector, $\mathbf{Z} = (Z_1, \dots, Z_m)$ a discrete random vector independent of $\mathbf{Y}$, and $g : \mathbb{R}^{k+m} \to \mathbb{R}$ a function for which $E[g(\mathbf{Y}, \mathbf{Z})]$ is finite. Then*

$$E[g(\mathbf{Y}, \mathbf{Z}) \,|\, \mathbf{Y}] = \sum_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z})\, g(\mathbf{Y}, \mathbf{z}). \tag{3.11}$$

Proof. As in the calculation (3.10) above, conditioning on $\mathbf{Y} = \mathbf{y}$ fixes the value $\mathbf{Y} = \mathbf{y}$ inside the expectation. Hence,

$$\begin{aligned} E[g(\mathbf{Y}, \mathbf{Z}) \,|\, \mathbf{Y} = \mathbf{y}] &= \sum_{\mathbf{z}} g(\mathbf{y}, \mathbf{z}) P(\mathbf{Z} = \mathbf{z} | \mathbf{Y} = \mathbf{y}) \\ &= \sum_{\mathbf{z}} g(\mathbf{y}, \mathbf{z}) P(\mathbf{Z} = \mathbf{z}). \end{aligned}$$

The second equality used the independence of $\mathbf{Y}$ and $\mathbf{Z}$. Equation (3.11) comes by substituting $\mathbf{Y}$ for $\mathbf{y}$. □

Example 3.10. Let Y and Z be independent random variables and $Z \sim \mathrm{Ber}(p)$. Compute $E[(Y + Z)^3 \,|\, Y]$.

The probability mass function of Z is $P(Z = 1) = p = 1 - P(Z = 0)$. According to Theorem 3.9,

$$\begin{aligned} E[(Y + Z)^3 \,|\, Y] &= \sum_z P(Z = z)\,(Y + z)^3 = (1 - p)Y^3 + p(Y + 1)^3 \\ &= Y^3 + p(3Y^2 + 3Y + 1). \end{aligned}$$

△

3.3. Martingales: definition and simple properties

Definition 3.11. Let $\{X_k : k \geq 0\}$ and $\{M_n : n \geq 0\}$ be two discrete stochastic processes defined on some probability space. The state space for X_k is arbitrary, but we assume $M_n \in \mathbb{R}$. Then $\{M_n : n \geq 0\}$ is a **martingale** with respect to $\{X_k : k \geq 0\}$ if the following three properties are satisfied for each $n \geq 0$:

(1) The random variable M_n is a function of $X_0, \dots, X_n$.

(2) M_n has finite expectation.

(3) The conditional expectation of the next value M_{n+1}, given the entire past of the X-process, equals the current value M_n:

$$E[M_{n+1}|(X_0, X_1, \dots, X_n)] = M_n. \tag{3.12}$$

We say that $\{M_n : n \geq 0\}$ is **adapted** to $\{X_n : n \geq 0\}$ if the first property (1) holds. The second property (2) is needed for the existence of all the conditional expectations. The third property (3) is called the **martingale property**. △

Remark 3.12. In many applications of martingales, the random variable X_n represents new information that we learn at the nth step. The vector of random variables $X_0, \dots, X_n$ represents all the information that we have after the nth time step. The assumption that M_n is a function of $X_0, \dots, X_n$ means that we can construct M_n from the information available after the nth step. The special case where $X_n = M_n$ for each n also satisfies the assumptions. The martingale property means that our best estimate of M_{n+1} in terms of the information available after the nth step is what we see M_n. △

Remark 3.13. When checking that a given process is a martingale, conditions (1) and (2) of Definition 3.11 are usually fairly self-evident, and the real issue is the conditional expectation condition (3).

Condition (3) can be replaced with the following equivalent version:

(4) For each $n \geq 0$ we have

$$E[M_{n+1} - M_n|(X_0, \dots, X_n)] = 0. \tag{3.13}$$

This equivalence follows because by Condition (1) of Definition 3.11, the random variable M_n is a function of $X_0, \dots, X_n$, and hence

$$E[M_n|(X_0, \dots, X_n)] = M_n.$$

Linearity of conditional expectation then gives

$$\begin{aligned} E[M_{n+1} - M_n|(X_0, \dots, X_n)] &= E[M_{n+1}|(X_0, \dots, X_n)] - E[M_n|(X_0, \dots, X_n)] \\ &= E[M_{n+1}|(X_0, \dots, X_n)] - M_n. \end{aligned}$$

This shows that we can replace Condition (3) of Definition 3.11 with condition (4) from above. △

An immediate consequence of the definition of a martingale is that its mean stays constant in time.

Lemma 3.14. *If $M_0, M_1, M_2, \dots$ is a martingale then $E[M_n] = E[M_0]$ for each n.*

Proof. By the martingale property, for all $n \geq 0$ we have

$$E[M_{n+1}|(X_0, \dots, X_n)] = M_n.$$

Taking expectations in both sides, and using the law of total expectation we get $E[M_{n+1}] = E[M_n]$. From this we get

$$E[M_0] = E[M_1] = \cdots = E[M_n]$$

for all $n \geq 1$. □

Example 3.15. Suppose that $X_1, X_2, X_3, \dots$ is a sequence of i.i.d. mean zero random variables. As before, define from these steps a random walk by $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$ for $n \geq 1$. Show that $S_0, S_1, \dots$ is a martingale with respect to the sequence $X_0, X_1, X_2, \dots$, where we set $X_0 = 0$.

The definition $S_n = X_1 + \cdots + X_n$ directly expresses S_n as a function of $X_0, X_1, \dots, X_n$. Hence Condition (1) of Definition 3.11 is satisfied. Since the X_ks have finite expectation, so does S_n, so Condition (2) is also satisfied. For the last condition we check that for $n \geq 0$ we have

$$E[S_{n+1} - S_n|(X_0, \dots, X_n)] = 0.$$

We have $S_{n+1} - S_n = X_{n+1}$, and X_{n+1} is independent of $X_0, X_1, \dots, X_n$. By Theorem 3.7(a),

$$E[S_{n+1} - S_n|X_0, \dots, X_n)] = E[X_{n+1}|X_0, \dots, X_n)] = E[X_{n+1}] = 0.$$

We have proved that $\{S_n : n \geq 0\}$ is a martingale with respect to $\{X_k : k \geq 0\}$. △

Example 3.16. Suppose that $Y_1, Y_2, \dots$ are i.i.d. mean 1 random variables. Let $M_0 = 1$ and $M_n = \prod_{k=1}^{n} Y_k$. Set $Y_0 = 1$. Show that $\{M_n : n \geq 0\}$ is a martingale with respect to $Y_0, Y_1, \dots$.

We check Condition (3) of Definition 3.11. From the definition $M_{n+1} = M_n Y_{n+1}$. Note that M_n is a function of $Y_0, \dots, Y_n$, and Y_{n+1} is independent of $Y_0, \dots, Y_n$. By (3.9),

$$E[M_{n+1}|(Y_0, \dots, Y_n)] = E[M_n Y_{n+1}|(Y_0, \dots, Y_n)] = M_n E[Y_n] = M_n$$

which proves the martingale property. △

Example 3.17 (Branching processes and martingales)**.** Consider a branching process X_n as defined in Section 2.7. Let μ denote the mean number of offspring. Then $M_n = X_n \mu^{-n}$ is a martingale with respect to $\{X_k : k \geq 0\}$.

According to the notation introduced in Section 2.7, $Z_{n,j}$ is the number of offspring of individual j from generation $n-1$. Then

$$X_{n+1} = \sum_{j=1}^{X_n} Z_{n+1,j}.$$

In particular,

$$E[X_{n+1}|X_n = m] = E\Big[\sum_{j=1}^{m} Z_{n+1,j}\Big| X_n = m\Big] = E\Big[\sum_{j=1}^{m} Z_{n+1,j}\Big] = m\mu,$$

since $\{Z_{n+1,j} : j \geq 1\}$ are independent of X_n and have mean μ. But this means that $E[X_{n+1}|X_n] = \mu X_n$, and hence

$$\begin{aligned} E[M_{n+1}|(X_0, \ldots, X_n)] &= E[\mu^{-(n+1)} X_{n+1}|(X_0, \ldots, X_n)] = \mu^{-(n+1)}\mu X_n \\ &= \mu^{-n} X_n = M_n. \end{aligned}$$

This verifies that $\{M_n : n \geq 0\}$ is a martingale. △

Example 3.18 (Pólya urn model)**.** We have an urn that initially contains one green ball and one yellow ball. We add balls to the urn by repeating the following step: draw a ball uniformly at random from the urn, note its color, put it back in the urn, and add one ball of the same color.

Let G_n denote the number of green balls and Y_n the number of yellow balls after the nth draw. The initial values are $G_0 = Y_0 = 1$. Suppose for example that the first four draws are green-green-yellow-green. Then $(G_1, Y_1) = (2, 1)$, $(G_2, Y_2) = (3, 1)$, $(G_3, Y_3) = (3, 2)$ and $(G_4, Y_4) = (4, 2)$.

Let $M_n = \frac{G_n}{G_n+Y_n}$ denote the ratio of green balls after the nth draw. We show that M_n is a martingale. The natural information at time n is $X_n = (G_n, Y_n)$, the contents of the urn after the nth draw. M_n is a function of X_n. Let $\mathbf{X}_n = (X_0, \ldots, X_n)$ denote the accumulated information after the nth draw, and $\mathbf{x} = ((g_0, y_0), \ldots, (g_n, y_n))$ a possible value of the random vector $\mathbf{X}_n$.

The conditional probability mass function of M_{n+1}, given $\mathbf{X}_n = \mathbf{x}$, follows from the rules of the process. With the current contents (g_n, y_n), a green ball is drawn with probability $\frac{g_n}{g_n+y_n}$ and then a green ball is added, and similarly for the opposite case:

$$P\biggl(M_{n+1} = \frac{g_n+1}{g_n+y_n+1} \,\bigg|\, \mathbf{X}_n = \mathbf{x}\biggr) = \frac{g_n}{g_n+y_n}$$

$$\text{and} \qquad P\biggl(M_{n+1} = \frac{g_n}{g_n+y_n+1} \,\bigg|\, \mathbf{X}_n = \mathbf{x}\biggr) = \frac{y_n}{g_n+y_n}.$$

From this we compute the conditional expectation:

$$\begin{aligned} E[M_{n+1} \,|\, \mathbf{X}_n = \mathbf{x}] &= \frac{g_n}{g_n+y_n} \cdot \frac{g_n+1}{g_n+y_n+1} + \frac{y_n}{g_n+y_n} \cdot \frac{g_n}{g_n+y_n+1} \\ &= \frac{g_n}{g_n+y_n}. \end{aligned}$$

Substituting in the random variables gives

$$E[M_{n+1} \,|\, \mathbf{X}_n] = \frac{G_n}{G_n+Y_n} = M_n$$

and we have shown the martingale property. △

Example 3.19 (Betting on a fair coin)**.** We are playing heads or tails in a casino. The coin is flipped repeatedly, and before the kth flip we are allowed to wager an amount of our choosing on the outcome. Encode heads as 1 and tails as -1 and let X_k denote the outcome of the kth flip. The amount that we wager on the kth flip is denoted by H_k, with the interpretation that $H_k > 0$ if we bet on heads, and $H_k < 0$ if we bet on tails. Then the amount we win (or lose) is exactly $H_k X_k$. (You should check that the signs work out correctly.)

After n games our total profit is $W_n = \sum_{k=1}^{n} H_k X_k$. (A negative amount represents a loss.) Let us call $H_1, H_2, H_3, \dots$ a *betting strategy* if H_1 is a constant, and for each $n \geq 2$, the random variable H_n is a function of $(X_1, \dots, X_{n-1})$ and has a finite expectation. In other words, the amount we bet on the nth coin flip can depend on n and the outcomes of the previous $n-1$ coin flips, but nothing else. We show that the sequence $W_1, W_2, W_3, \dots$ is a martingale with respect to $X_1, X_2, X_3, \dots$. We can add a zeroth term to both sequences as $X_0 = W_0 = 0$.

We check Condition (3) of Definition 3.11 and leave the other two conditions to the reader. Note that $W_{n+1} - W_n = X_{n+1}H_{n+1}$. By assumption X_{n+1} is independent of $(X_0, X_1, \dots, X_n)$, while H_{n+1} is a function of these random variables. Hence

$$\begin{aligned} E[W_{n+1} - W_n | (X_0, \dots, X_n)] &= E[X_{n+1}H_{n+1} | (X_0, \dots, X_n)] \\ &= H_{n+1}E[X_{n+1} | (X_0, \dots, X_n)] = H_{n+1}E[X_{n+1}] = H_{n+1} \cdot 0 = 0. \end{aligned}$$

Thus $\{W_n : n \geq 0\}$ satisfies the martingale property. △

A consequence of the constant mean of a martingale (Lemma 3.14) is that in a sequence of repeated independent fair games there is no betting strategy that would yield a profit on average, after any fixed number of games.

A small generalization of Example 3.19 leads to the following result. It is of more than cursory interest. This lemma will be used to prove further properties of martingales, such as Lemma 3.22 in the next section.

Lemma 3.20 (Betting on a martingale). *Suppose that M_n is a martingale with respect to the sequence $\{X_n : n \geq 0\}$. Let $\{H_n : n \geq 1\}$ be a sequence of random variables such that, for each $n \geq 1$, $E|H_n|$ and $E|H_n(M_n - M_{n-1})|$ are finite and H_n is a function of $(X_0, \dots, X_{n-1})$. Let*

$$W_0 = M_0 \quad \textit{and} \quad W_n = M_0 + \sum_{k=1}^{n} H_k(M_k - M_{k-1}) \quad \textit{for } n \geq 1.$$

Then $\{W_n : n \geq 0\}$ is a martingale with respect to $\{X_n : n \geq 0\}$.

Proof. The proof is almost the same as in the previous example.

$$\begin{aligned} E[W_{n+1} - W_n | (X_0, \dots, X_n)] &= E[H_{n+1}(M_{n+1} - M_n) | (X_0, \dots, X_n)] \\ &= H_{n+1}E[M_{n+1} - M_n | (X_0, \dots, X_n)] = H_{n+1} \cdot 0 = 0, \end{aligned}$$

where the second last equality used the martingale property of M_n.

□

Example 3.21 (Doubling the bet until a win). The reader may wonder about the following strategy for betting on coin flips that seems to offer a certain profit. Start by betting \$1 that the first coin flip is tails. If you win, stop and keep the one dollar profit. If the first flip is heads and you lose your dollar, then bet \$2 that the second coin flip is tails. Keep doubling the bet until the first tails and then stop. Suppose the first tails comes on the Nth flip. Then you have $N-1$ straight losses followed by a win at the Nth flip, so the net profit is

$$-1 - 2 - 4 - \cdots - 2^{N-2} + 2^{N-1} = 1. \tag{3.14}$$

(There are N terms in the sum above because the powers of 2 start from $1 = 2^0$.)

We know tails must come up eventually. Hence a one dollar profit appears to be certain. Why does this not contradict the mathematics we have developed above? Is this really a practical strategy?

There are two points to be made here. The first one is that the time N when the game stops is random. Consider what happens if we force the game to end by some fixed time n. By Lemma 3.14 the expected profit should be zero. We can check this directly. Let W_n be the profit earned when the game is forced to stop by time n. W_n can take two possible values, 1 (if there is a tails among the first n flips) and $-(1 + 2 + \cdots + 2^{n-1}) = -2^n + 1$ (if all n coin flips are heads). Thus

$$\begin{aligned} P(W_n = 1) &= P(\text{at least one tails among the } n \text{ flips}) = 1 - 2^{-n}, \\ P(W_n = -2^n + 1) &= P(\text{all } n \text{ coin flips are heads}) = 2^{-n} \end{aligned}$$

from which

$$E[W_n] = 1 \cdot (1 - 2^{-n}) + (-2^n + 1) \cdot 2^{-n} = 0. \tag{3.15}$$

In the next section we discuss the possibility to observing a martingale at a random time.

The second point is a practical one. There must be an amount beyond which the player cannot go into more debt. In other words, practical considerations force some upper bound on the duration of the game. Then we are back to (3.15) and zero expected profit. △

3.4. Martingales and stopping times

Recall the definition of a stopping time from Definition 2.23. T is a stopping time with respect to the sequence $X_0, X_1, X_2, \ldots$ if the event $\{T = n\}$ is determined by $X_0, X_1, \ldots, X_n$. An equivalent way to express this is that the values of the indicator $I\{T = n\}$ can be expressed as a function of $(X_0, X_1, \ldots, X_n)$. We can think of T as a *stopping rule* in terms of the sequence $\{X_n : n \geq 0\}$. The lemma below says that a martingale frozen at a stopping time remains a martingale, even as we let time march on past the stopping time. We use the abbreviation $a \wedge b = \min\{a, b\}$ for the minimum of two numbers.

Lemma 3.22. *Suppose that $\{M_n : n \geq 0\}$ is a martingale with respect to $\{X_n : n \geq 0\}$, and let T be a stopping time with respect to the same sequence $\{X_n : n \geq 0\}$. Define the stopped process $\{W_n : n \geq 0\}$ as $W_n = M_{T \wedge n}$ for $n \geq 0$. More explicitly,*

$$W_n = \begin{cases} M_n, & \textit{if} \quad n \leq T \\ M_k, & \textit{if} \quad n > T = k. \end{cases}$$

Then $\{W_n : n \geq 0\}$ is also a martingale with respect to $\{X_n : n \geq 0\}$.

Proof. The key to the proof is to find a betting strategy $\{H_k\}$ such that

$$M_{T \wedge n} = M_0 + \sum_{k=1}^{n} H_k (M_k - M_{k-1}) \tag{3.16}$$

and apply Lemma 3.20. Since the telescoping sum satisfies $M_0 + \sum_{k=1}^m (M_k - M_{k-1}) = M_m$, it is evident that we want to choose H_k so that the adding is stopped when T is reached. Thus we set

$$H_k = I(T \geq k),$$

the indicator that the event $T \geq k$ happens. Let us verify that this choice works.

If $T \geq n$ then $H_1 = \cdots = H_n = 1$ and the total profit at time n is

$$M_0 + \sum_{k=1}^{n} (M_k - M_{k-1}) = M_n.$$

If $T < n$ then $H_1 = \cdots = H_T = 1$ while $H_k = 0$ for $k > T$. Now the total profit at time n is

$$M_0 + \sum_{k=1}^{T} (M_k - M_{k-1}) = M_T.$$

The two cases together check that (3.16) holds.

In order to appeal to Lemma 3.20 to conclude that $M_{T \wedge n}$ is a martingale, we need to check that H_n is a function of $(X_0, \ldots, X_{n-1})$. Indeed,

$$H_n = I(T \geq n) = 1 - I(T < n) = 1 - \sum_{k=0}^{n-1} I(T = k).$$

By the definition of a stopping time, the value of the indicator $I\{T = k\}$ is a function of $X_0, \ldots, X_k$. The equality above then verifies that the value of H_n is determined by $(X_0, \ldots, X_{n-1})$. □

Applying now Lemma 3.14 we get that if T is a stopping time then

$$E[M_{T \wedge n}] = E[M_{T \wedge (n-1)}] = \cdots = E[M_{T \wedge 0}].$$

But $T \wedge 0 = 0$, which means that

$$E[M_{T \wedge n}] = E[M_0].$$

If T is an a.s. finite number then $\lim_{n\to\infty} T \wedge n = T$, since eventually n will be larger than T. Thus we have $\lim_{n\to\infty} M_{T \wedge n} = M_T$. If we could switch around the limit and the expectation then this would imply

$$E[M_T] = E[\lim_{n\to\infty} M_{T \wedge n}] \overset{?}{=} \lim_{n\to\infty} E[M_{T \wedge n}] = E[M_0]. \tag{3.17}$$

Unfortunately, the step with $\overset{?}{=}$ does not always hold, as the following counterexamples show.

Example 3.23. Here is a simple counterexample: consider the symmetric simple random walk $\{S_n : n \geq 0\}$ with $S_0 = 0$ built from the i.i.d. random variables $\{X_k : k \geq 1\}$ with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Then S_n is a martingale with respect to $\{X_k : k \geq 1\}$ by Example 3.15. Let T be the first hitting time of 1:

$$T = \min\{n \geq 1 : S_n = 1\}.$$

Since S_n is an irreducible recurrent Markov chain, we have $P(T < \infty) = 1$. By definition $S_T = 1$ and hence $E[S_T] = 1$. But $E[S_0] = 0$, which shows $E[S_T] \neq E[S_0]$. △

Example 3.21 gives another counterexample.

Example 3.24. In Example 3.21 we considered the betting strategy of always betting on tails until we first win, with doubling our bet after each loss. The total profit earned by game n is a martingale with respect to the sequence of coin flips. The stopped martingale has a value of 1 (see (3.14)). Hence in this case the expected value of the stopped martingale is not the same as the expected value at time 0 (which is zero). △

With some additional assumptions on the stopping time T and the martingale in question one can prove $E[M_T] = E[M_0]$. Any one of the additional assumptions in the theorem below allows us to justify the step $E[\lim_{n\to\infty} M_{T\wedge n}] = \lim_{n\to\infty} E[M_{T\wedge n}]$ in equation (3.17) above.

Theorem 3.25 (Optional stopping)**.** *Suppose that $\{M_n : n \geq 0\}$ is a martingale with respect to the discrete process $\{X_k : k \geq 0\}$. Let $T \geq 0$ be an a.s. finite stopping time with respect to the random variables $X_0, X_1, \ldots$. If any of the following conditions are satisfied then $E[M_T] = E[M_0]$.*

(i) *There is a constant $c > 0$ such that $P(T \leq c) = 1$.*

(ii) *There is a constant $c > 0$ such that $P(|M_{T\wedge n}| \leq c) = 1$ for all $n \geq 0$.*

(iii) *$E[T] < \infty$ and there is a constant $c > 0$ such that $P(|M_{T\wedge(n+1)} - M_{T\wedge n}| \leq c) = 1$ for all $n \geq 0$.*

Here are some applications of Theorem 3.25.

Example 3.26 (Back to gambler's ruin, symmetric case)**.** Consider a simple symmetric random walk $S_n, n \geq 0$, with $S_0 = 0$. For given integers $a, b > 0$ let

$$T = \inf_{n\geq 0}\{S_n \in \{-a, b\}\}$$

be the first hitting time of the set $\{-a, b\}$. Following Example 2.16 shows that $P(T < \infty) = 1$. The calculation there can be used to show that $P(T > (a+b)n) \leq (1 - (\frac{1}{2})^{a+b})^n$ which is strong enough for deducing that $E[T] < \infty$. (Exercise 2.15 gives an alternative derivation of $E[T] < \infty$, valid in general when the set of transient states is finite.)

The process $S_n, n \geq 0$ is a martingale, and T is a stopping time. Since $P(T < \infty) = 1$, the random variable S_T is either equal to $-a$ or b. In particular $|S_T| \leq \max(a, b)$, and we may use Theorem 3.25 to give

$$E[S_T] = E[S_0] = 0.$$

Using again the fact that S_T is either equal to $-a$ or b we get

$$\begin{aligned} 0 = E[S_T] &= -aP(S_T = -a) + bP(S_T = b) \\ &= -aP(S_T = -a) + b(1 - P(S_T = -a)) \\ &= b - (a+b)P(S_T = -a) \end{aligned}$$

which gives

$$P(S_T = -a) = \frac{b}{a+b}, \qquad P(S_T = b) = \frac{a}{a+b}. \tag{3.18}$$

We can also compute the expected value of T. By Exercise 3.8 the process $M_n = S_n^2 - n$ is a martingale. To apply optional stopping, we check that $|M_{T\wedge(n+1)} - M_{T\wedge n}|$ is bounded:

$$\begin{aligned}|M_{T\wedge(n+1)} - M_{T\wedge n}| &= \left|S^2_{T\wedge(n+1)} - S^2_{T\wedge n} - T\wedge(n+1) + T\wedge n\right| \\ &\le |S^2_{T\wedge(n+1)}| + |S^2_{T\wedge n}| + |T\wedge(n+1) - T\wedge n| \\ &\le 2\max(a^2, b^2) + 1.\end{aligned}$$

Now we can apply Theorem 3.25 to give

$$E[M_T] = E[M_0] = 0.$$

But $E[M_T] = E[S_T^2 - T] = E[S_T^2] - E[T]$, hence

$$E[T] = E[S_T^2].$$

Using (3.18) we can compute this mean:

$$E[T] = E[S_T^2] = a^2 P(S_T = -a) + b^2 P(S_T = b) = a^2\frac{b}{a+b} + b^2\frac{a}{a+b} = ab.$$

△

Example 3.27 (Gambler's ruin, asymmetric case)**.** Consider simple random walk $S_n = \sum_{k=1}^n X_k$, with $S_0 = 0$, and $X_,k \ge 1$ i.i.d. with $P(X_1 = 1) = 1 - P(X_1 = -1) = p$ with some $0 < p < 1$, $p \neq \frac{1}{2}$. We will use the notation $q = 1 - p$.

Again, for given integers $a, b > 0$ let

$$T = \inf_{n\ge 0}\{S_n \in \{-a, b\}\}$$

be the first hitting time of the set $\{-a, b\}$.

Consider the random variable $\left(\frac{q}{p}\right)^{X_k}$. These are i.i.d. random variables with expected value

$$E\left[\left(\frac{q}{p}\right)^{X_k}\right] = \frac{q}{p}\cdot p + \frac{p}{q}\cdot q = q + p = 1.$$

By Example 3.16 the process $M_n = \prod_{k=1}^n \left(\frac{q}{p}\right)^{X_k} = \left(\frac{q}{p}\right)^{S_n}$ is a martingale. The process $M_{T\wedge n}$ is bounded, hence by Theorem 3.25 we have

$$E[M_T] = E[M_0] = 1.$$

But S_T is either $-a$ or b, hence

$$\begin{aligned}1 = E[M_T] &= E\left[\left(\frac{q}{p}\right)^{S_T}\right] \\ &= \left(\frac{q}{p}\right)^{-a} P(S_T = -a) + \left(\frac{q}{p}\right)^{b} P(S_T = b).\end{aligned}$$

Since $P(T < \infty) = 1$ we have $P(S_T = b) = 1 - P(S_T = a)$. This leads to

$$P(S_T = -a) = \frac{\left(\frac{q}{p}\right)^{a+b} - \left(\frac{q}{p}\right)^{a}}{\left(\frac{q}{p}\right)^{a+b} - 1}, \qquad P(S_T = b) = \frac{\left(\frac{q}{p}\right)^{a} - 1}{\left(\frac{q}{p}\right)^{a+b} - 1}.$$

We can also compute $E[T]$ by considering the martingale $S_n - (p-q)n$. We get

$$E[S_T - (p-q)T] = 0,$$

which lead to

$$\begin{aligned} E[T] &= \frac{1}{p-q} E[S_T] = \frac{1}{p-q} \left(-aP(S_T = -a) + bP(S_T = b)\right) \\ &= \frac{1}{p-q} \left(-a \frac{\left(\frac{q}{p}\right)^{a+b} - \left(\frac{q}{p}\right)^{a}}{\left(\frac{q}{p}\right)^{a+b} - 1} + b \frac{\left(\frac{q}{p}\right)^{a} - 1}{\left(\frac{q}{p}\right)^{a+b} - 1} \right). \end{aligned}$$

△

The next example is the simplified version of the famous 'Monkey and the typewriter' problem. Imagine that we have a monkey who starts hitting the keys on a typewriter randomly (or hitting the keys of the keyboard of a laptop). It can be shown that with probability one, at some point the monkey will type exactly the text of your favorite novel. But how long do we have to wait for that? The next example describes the solution where we only have two keys on the keyboard (0 and 1), with the monkey choosing one of them with probability $1/2 - 1/2$. Exercise 3.18 asks you to extend the result to the general case.

Example 3.28 (Waiting for a pattern in a sequence of coin flips)**.** Suppose that $X_n, n \geq 1$ is an i.i.d. sequence of Bernoulli(1/2) random variables:

$$P(X_n = 1) = P(X_n = 0) = \tfrac{1}{2}.$$

Let $A = (a_1, \dots, a_k)$ be a finite sequence with $a_j \in \{0, 1\}$. We call A a 'word' or a 'pattern', and $X_n, n \geq 1$ random letters. How long do we have to wait on average to see the pattern A appear in the sequence of random letters? More precisely, let

$$T_A = \inf\{n \geq k : X_{n-k+1} = a_1, X_{n-k+2} = a_2, \dots, X_n = a_k\}$$

be the first time when we see A appear in the last k random letters. What is the value of $E[T_A]$?

If $A = (1, 1, \dots, 1)$ then T_A is the same as the hitting time of state k in the success run Markov chain with success probability $1/2$. The mean was calculated in Exercise 2.11. For a general pattern A one could form a Markov chain on the state space $\{0, 1\}^k$ that keeps track of the last k letters of the sequence $\{X_j, j \geq 1\}$. By finding the expected hitting time of the the state A one could access $E[T_A]$, although this gets complicated for large k. We outline a martingale method that gives a simple formula for $E[T_A]$.

We introduce the following process for $n \geq 0$, with $Z_0 = 0$:

$$\begin{aligned} Z_n &= -n + \sum_{j=1}^{n} 2^{k\wedge(n-j+1)} I(X_j = a_1, X_{j+1} = a_2, \ldots, X_{(j+k-1)\wedge n} = a_{k\wedge(n-j+1)}) \\ &= -n + \sum_{j=1}^{n-k+1} 2^k I(X_j = a_1, \ldots, X_{j+k-1} = a_k) \\ &\quad + \sum_{j=n-k+2}^{n} 2^{n+1-j} I(X_j = a_1, \ldots, X_n = a_{n+1-j}). \end{aligned} \tag{3.19}$$

The sum in the definition can be understood as follows: for each $1 \leq j \leq n-k+1$ we check whether the word $A = (a_1, \ldots, a_k)$ shows up from position j onwards, and if it does, we add 2^k to the sum. For $n-k+1 < j \leq n$ we check whether the beginning of the word A matches the letters that appear from position j to n, and if it does, we add 2^{n+1-j} to the sum. The length of the matching part is exactly $n+1-j$ in this last case.

We prove that $Z_n, n \geq 0$ is a martingale with respect to the sequence $X_j, j \geq 1$. We first give a proof by checking the definition, and then another one by showing that Z_n can be obtained from a betting strategy.

To show $E[Z_{n+1}|(X_1, \ldots, X_n)] = Z_n$ we evaluate the conditional expectations of the indicators appearing in the sum (3.19), with n replaced by $n+1$. There are two cases.

(i) If $1 \leq j+k-1 \leq n$ then

$$E[I(X_j = a_1, \ldots, X_{j+k-1} = a_k)|(X_1, \ldots, X_n)] = I(X_j = a_1, \ldots, X_{j+k-1} = a_k)$$

because in this case the indicator random variable is a function of $(X_1, \ldots, X_n)$.

(ii) If $n+1-k < j \leq n+1$ then the jth term in the sum (3.19) corresponding to $n+1$ has the indicator $I(X_j = a_1, \ldots, X_{n+1} = a_{n+2-j})$. Rewrite this as

$$\begin{aligned} &I(X_j = a_1, \ldots, X_{n+1} = a_{n+2-j}) \\ &\qquad = I(X_j = a_1, \ldots, X_n = a_{n+1-j}) I(X_{n+1} = a_{n+2-j}). \end{aligned}$$

By the independence of X_{n+1} and $(X_1, \ldots, X_n)$ we get

$$\begin{aligned} &E[I(X_j = a_1, \ldots, X_n = a_{n+1-j}) I(X_{n+1} = a_{n+2-j})|(X_1, \ldots, X_n)] \\ &\quad = I(X_j = a_1, \ldots, X_n = a_{n+1-j}) E[I(X_{n+1} = a_{n+2-j})|(X_1, \ldots, X_n)] \\ &\quad = \frac{1}{2} I(X_j = a_1, \ldots, X_n = a_{n+1-j}). \end{aligned}$$

If $j = n+1$ the last term reduces to just $\frac{1}{2}$.

Combine the cases and use linearity to get

$$
\begin{aligned}
E[Z_{n+1}|(X_1,\dots,X_n) &= -(n+1) \\
&+ \sum_{j=1}^{n-k+1} 2^k E[I(X_j = a_1,\dots,X_{j+k-1} = a_k)|(X_1,\dots,X_n)] \\
&+ \sum_{j=n-k+2}^{n+1} 2^{n+2-j} E[I(X_j = a_1,\dots,X_{n+1} = a_{n+2-j})|(X_1,\dots,X_n)] \\
&= -(n+1) + \sum_{j=1}^{n-k+1} 2^k I(X_j = a_1,\dots,X_{j+k-1} = a_k) \\
&\quad + \sum_{j=n-k+2}^{n} 2^{n+2-j}\frac{1}{2} I(X_j = a_1,\dots,X_{n+1} = a_{n+2-j}) + 2\cdot\frac{1}{2}.
\end{aligned}
$$

The last term is the contribution of the $j = n+1$ term. Comparison with (3.19) shows that the final form above is exactly Z_n. We have verified that $Z_n, n \geq 0$ is indeed a martingale.

Here is another proof of the martingale property of Z_n using a betting representation of Z_n. Suppose that we are betting on the letters $X_j, j \geq 1$: we can place a bet for the jth letter being 0 or 1, and we double or lose our bet depending on the outcome. (Note that this is a fair game!). Suppose that we bet 1$ on the first letter being a_1. If we lose, we stop betting, but if we win then we double our bet and put $2 on the second letter being a_2. We continue this until the kth letter, after which we do not bet anymore. This strategy is very similar to the one presented in Example 3.21 except we double our bet when we win (and we place the bets according to the letters of A). After the kth letter we either win $1 + 2 + \cdots + 2^{k-1} = 2^k - 1$, or if our first lost bet is the jth one (with $1 \leq j \leq k$) then our net winning is $1 + 2 + \cdots + 2^{j-2} - 2^{j-1} = -1$. Hence our net profit equals

$$2^k I(X_1 = a_1,\dots,X_k = a_k) - 1.$$

With a similar computation we can check that if $1 \leq j < k$ then after the jth letter our net winning is given by

$$2^j I(X_1 = a_1,\dots,X_j = a_j) - 1.$$

Now imagine that we are using this betting strategy to play *parallel* games where we start betting considering the letters $X_j, X_{j+1}, \dots$ for each $j \geq 1$. Another way to think about this is to imagine that we have an infinite group of gamblers, and the jth one begins betting starting with the jth letter (using the strategy described above). Then if we consider the cumulative net winning (from all the parallel games) after the nth coin flip then this is given exactly by Z_n. Since it is the result of a betting strategy involving a fair game, $Z_n, n \geq 0$ must be a martingale.

It can be shown (see Exercise 3.16) that $E[T_A] < \infty$ and $|Z_{T_A\wedge(n+1)} - Z_{T_A\wedge n}|$ is bounded, hence we can use Theorem 3.25 to get

$$E[Z_{T_A}] = E[Z_0] = 0.$$

By (3.19) we have

$$Z_{T_A} = -T_A + \sum_{j=1}^{T_A-k} 2^k I(X_j = a_1, X_{j+1} = a_2, \ldots, X_{j+k-1} = a_k)$$

$$+ \sum_{j=T_A-k+1}^{T_A} 2^{T_A-j+1} I(X_j = a_1, \ldots, X_{T_A} = a_{T_A-j+1}). \tag{3.20}$$

We claim that the sums in (3.20) are deterministic functions of A that we can evaluate. Indeed, since $X_{T_A-k+1}, X_{T_A-k+2} \ldots X_{T_A}$ gives the first appearance of the word A in the sequence $X_j, j \geq 1$, we know that $(X_j, X_{j+1}, \ldots, X_{j+k-1})$ cannot be equal to A for $j \leq T_A - k$. Hence

$$\sum_{j=1}^{T_A-k} 2^k I(X_j = a_1, X_{j+1} = a_2, \ldots, X_{j+k-1} = a_k) = 0.$$

On the other hand, since $(X_{T_A-k+1}, X_{T_A-k+2} \ldots X_{T_A}) = (a_1, \ldots, a_k)$, the terms in the second sum of (3.20) can be evaluated explicitly because for $1 \leq \ell \leq k$ we get

$$I(X_{T_A-\ell+1} = a_1, \ldots, X_{T_A} = a_\ell) = I(a_{k-\ell+1} = a_1, a_{k-\ell+2} = a_2, \ldots, a_k = a_\ell).$$

Therefore

$$Z_{T_A} = -T_A + \sum_{\ell=1}^{k} 2^\ell I(a_{k-\ell+1} = a_1, a_{k-\ell+2} = a_2, \ldots, a_k = a_\ell)$$

Taking expectations and using $E[Z_{T_A}] = E[Z_0] = 0$ we get

$$E[T_A] = \sum_{\ell=1}^{k} 2^\ell I(a_{k-\ell+1} = a_1, a_{k-\ell+2} = a_2, \ldots, a_k = a_\ell).$$

The terms in the sum correspond to 'overlaps' of A with its shifted version. The term corresponding to $\ell = k$ will always contribute 2^k, and the more overlaps the word A has with its shifted versions, the larger will this sum be, and the longer we have to wait on average. For example, for the four letter words 1111, 0101, 0111 we get $E[T_{1111}] = 2 + 2^2 + 2^3 + 2^4$, $E[T_{0101}] = 2^4 + 2^2$, and $E[T_{0111}] = 2^4$. △

3.5. Martingale limit theorem

Another valuable property of martingales is that they converge almost surely under some additional conditions.

Theorem 3.29 (Limit theorem for bounded martingales)**.** *Let $\{M_n : n \geq 0\}$ be a martingale with respect to $\{X_k : k \geq 0\}$. Suppose there exists a constant $c \in \mathbb{R}$ such that $P(M_n \geq c) = 1$ for all n. Then there exists a random variable M_∞ that is almost surely finite, and $\lim_{n\to\infty} M_n = M_\infty$ with probability one.*

The theorem states that if a martingale is bounded from below as a sequence by a deterministic number then it must converge to a limit. (See e.g. [Dur19] for a proof.) The same is true for martingales that are bounded from above. This can be seen by noting that if $\{M_n : n \geq 0\}$ is a martingale, then so is $\{-M_n : n \geq 0\}$, and if M_n is bounded from above then $-M_n$ is bounded from below.

Example 3.30 (Simple symmetric walk)**.** With SSRW S_n started from $S_0 = 0$ we can create both examples and non-examples of the theorem.

The first observation is that since $S_{n+1} = S_n \pm 1$ for all times n, S_n cannot converge as $n \to \infty$. Consistently with this, Theorem 3.29 does not apply because there is no lower bound $S_n \geq c$ that works for all n. This is because S_n can take the value $-n$.

By stopping the walk we can create an example of a convergent martingale. Let

$$T = \inf\{n \geq 0 : S_n = 1\}$$

be the first time when the walk visits state 1. T is a stopping time. By Lemma 3.22, the process $M_n = S_{T\wedge n}$ is a martingale. This martingale satisfies $M_n \leq 1$ because once M_n hits the point 1, it is frozen there. Theorem 3.29 applies and implies that there is a limit $M_n \to M_\infty$ as $n \to \infty$.

Can we describe the limit? Since S_n is an irreducible recurrent Markov chain, we have $P(T < \infty) = 1$. Thus with probability one, eventually n exceeds T, and after that $M_n = S_{T\wedge n} = S_T = 1$. In other words, $M_\infty = 1$, a constant. △

In the previous example the martingale $M_n = S_{T\wedge n}$ converged to a constant limit. In the next example the limit is genuinely random.

Example 3.31 (Pólya urn revisited)**.** In Example 3.18 we introduced the Pólya urn model. We showed that the ratio of green balls after the nth ball, $M_n = \frac{G_n}{G_n+Y_n}$, is a martingale with respect to the sequence $(G_k, Y_k), k \geq 0$. Since M_n is a ratio, it is always at least 0. By Theorem 3.29 there exists a random variable M_∞ such that $\lim_{n\to\infty} M_n = M_\infty$ with probability one. Concretely this means that, as we add more and more balls in the urn, over the long term the ratio of green balls stabilizes to a definite limiting value.

What can we say about the limit? We show that in fact the limit M_∞ is random, with uniform distribution on the unit interval $[0, 1]$. First we prove by induction that G_n is uniformly distributed on the set $\{1, 2, \dots, n+1\}$. (Note that after the nth draw the total number of balls in the urn is $n + 2$, and both colors are present in the urn.) Since $G_0 = 1$, the statement holds for $n = 0$. Assume that the statement holds for a particular value $n \geq 0$. In other words,

$$P(G_n = k) = \frac{1}{n+1} \qquad \text{for all } 1 \leq k \leq n+1.$$

For the induction step we need to show that $P(G_{n+1} = k) = \frac{1}{n+2}$ for $1 \leq k \leq n+2$. This separates naturally into three cases: $k = 1$, $k = n + 2$, and $1 < k < n + 2$.

(i) The definition of the urn process tells us that $G_{n+1} = 1$ can happen only if $G_n = 1$ (and $Y_n = n + 1$) and the $(n + 1)$st draw is yellow. Hence

$$P(G_{n+1} = 1) = P(G_n = 1,\ (n+1)\text{st draw is yellow}) = \frac{1}{n+1} \cdot \frac{n+1}{n+2} = \frac{1}{n+2}.$$

(ii) Similarly,

$$P(G_{n+1} = n+2) = P(G_n = n+1,\ (n+1)\text{st draw is green}) = \frac{1}{n+1} \cdot \frac{n+1}{n+2} = \frac{1}{n+2}.$$

(iii) For $1 < k < n+2$, if $G_{n+1} = k$ then G_n is either k or $k-1$. Hence we get

$$\begin{aligned} P(G_{n+1} = k) &= P(G_n = k,\ (n+1)\text{st draw is yellow}) \\ &\qquad + P(G_n = k-1,\ (n+1)\text{st draw is green}) \\ &= \frac{1}{n+1}\cdot\frac{n+2-k}{n+2} + \frac{1}{n+1}\cdot\frac{k-1}{n+2} \\ &= \frac{n+1}{(n+1)(n+2)} = \frac{1}{n+2}. \end{aligned}$$

The induction step has been verified. Thus for each n, G_n is uniformly distributed on the set $\{1, 2, \dots, n+1\}$. Consequently $M_n = \frac{G_n}{n+2}$ is uniform on the set $\{\frac{1}{n+2}, \frac{2}{n+2}, \dots, \frac{n+1}{n+2}\}$.

With a small calculation we deduce that the limiting random variable M_∞ must be uniformly distributed on $[0,1]$. Let $t \in [0,1]$ be real. An integer k satisfies $\frac{k}{n+2} \le t$ iff $k \le \lfloor t(n+2) \rfloor$, and so the cumulative distribution function of M_n is

$$P(M_n \le t) = \begin{cases} \frac{\lfloor t(n+2)\rfloor}{n+1} & \text{for } 0 \le t \le \frac{n+1}{n+2}, \\ 1 & \text{for } \frac{n+1}{n+2} \le t \le 1. \end{cases}$$

From the inequality

$$\frac{t(n+2)}{n+1} - \frac{1}{n+1} \le \frac{\lfloor t(n+2)\rfloor}{n+1} \le \frac{t(n+2)}{n+1}$$

we conclude that $P(M_n \le t) \to t$ as $n \to \infty$. (See Figure 1 for an illustration.) Consequently the cumulative distribution function of M_∞ is exactly that of the uniform distribution:

$$P(M_\infty \le t) = \lim_{n\to\infty} P(M_n \le t) = t \qquad \text{for } 0 \le t \le 1.$$

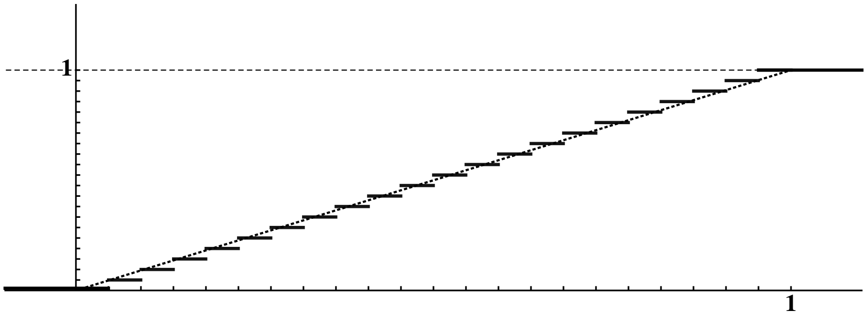

Figure 1. The cumulative distribution function of M_{20}. The slanted dotted line shows the cumulative distribution function of the uniform distribution.

The line above uses an unmentioned theorem: that almost sure convergence $M_n \to M_\infty$ implies the convergence of certain probabilities of M_n to those of M_∞. △

Example 3.32 (The scaled limit of the branching process population)**.** We have discussed in Example 3.17 that if X_n is branching process with mean number of offspring μ, then $M_n = X_n\mu^{-n}$ is a martingale. Because X_n denotes the size of the population at time n, we have $X_n \geq 0$ and hence M_n is bounded from below. By Theorem 3.29 there is an almost surely finite random variable M_∞ so that $\lim_{n\to\infty} M_n = M_\infty$ with probability one.

In Section 2.7 we proved that if $\mu \leq 1$ then the branching process dies out with probability one, meaning that eventually we will have $X_n = 0$, and hence $M_n = 0$. This means that for $\mu \leq 1$ we have $M_\infty = 0$ (and we would not even need Theorem 3.29 for this limit).

If $\mu > 1$ then we know that the process X_n survives forever with a positive probability. Moreover, we have seen that if the process survives forever then its value must go to infinity. Now we can actually say something about the speed with which it grows. By the martingale limit theorem $X_n\mu^{-n}$ converges to a random value M_∞. This means that X_n cannot grow faster than the function μ^n, and on the event $\{M_\infty > 0\}$ the process will go to ∞ exponentially fast, at the same rate as μ^n.

The exact growth behavior was described by Kesten and Stigum in article [**KS66**], who showed that $P(M_\infty > 0) > 0$ if and only if $\sum_{k\geq 1} \beta_k k \log k < \infty$ where $\{\beta_k\}_{k\geq 0}$ is the offspring distribution. △

Exercises

Exercise 3.1. Suppose $E[\,|X|\,] < \infty$ and $P(\mathbf{Y} = \mathbf{y}) > 0$. Show that then

$$E[\,|X|\,|\,\mathbf{Y} = \mathbf{y}] \leq \frac{1}{P(\mathbf{Y} = \mathbf{y})} E[\,|X|\,].$$

This verifies that the conditional expectation is a well-defined finite number.

Hint. Simplify the defining formula of the conditional expectation.

Exercise 3.2. Suppose that $X \sim \text{Binom}(n, p)$ and $Y \sim \text{Binom}(m, p)$ are independent, and denote $Z = X + Y$.

(a) Find the conditional probability mass function of X given Z.

(b) Find the conditional expectation $E[X|Z]$.

Exercise 3.3. Suppose that $\{X_k : k \geq 1\}$ are i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Let $S_n = \sum_{k=1}^n X_k$. Find $E[S_n^3|S_m]$ for $1 \leq m \leq n$.

Exercise 3.4. Suppose that X is a random variable with a finite expectation, Y is a random vector, and $Z = E[X|Y]$. Express $E[Z|Y]$ in terms of X, Y and Z.

Exercise 3.5. Prove directly from the definitions that $E[g(\mathbf{Y})|\mathbf{Y}] = g(\mathbf{Y})$, without appealing to Theorem 3.7.

Exercise 3.6. Suppose that X is a random variable with a finite expectation and $\{Y_k : k \geq 1\}$ are random variables.

(a) Show that for any $1 \leq m < n$ we have

$$E[E[X|(Y_1, \dots, Y_m)]|(Y_1, \dots, Y_n)] = E[X|(Y_1, \dots, Y_m)]$$

(b) Show that for any $1 \leq m < n$ we have

$$E[E[X|(Y_1, \ldots, Y_n)]|(Y_1, \ldots, Y_m)] = E[X|(Y_1, \ldots, Y_m)]$$

Exercise 3.7. Suppose that $\{X_k : k \geq 1\}$ is a sequence of i.i.d. random variables with finite expectation $E[X_1] = \mu$. Let $S_n = \sum_{k=1}^n X_k$. Show that $S_n - \mu n, n \geq 1$ is a martingale with respect to $\{X_k : k \geq 1\}$.

Exercise 3.8. Suppose that $\{X_k : k \geq 1\}$ is a sequence of i.i.d. random variables with $E[X_1] = 0$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Let $S_n = \sum_{k=1}^n X_k$. Show that $S_n^2 - n\sigma^2, n \geq 1$ is a martingale with respect to $\{X_k : k \geq 1\}$.

Exercise 3.9. Suppose that $\{X_k : k \geq 1\}$ is a sequence of i.i.d. random variables with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Let $S_n = \sum_{k=1}^n X_k$. Find a degree three polynomial of S_n (with coefficients possibly depending on n) that is a martingale with respect to $\{X_k : k \geq 1\}$.

Exercise 3.10. Let $0 < p < 1$ and $\{X_k : k \geq 1\}$ be a sequence of i.i.d. random variables with $P(X_1 = 1) = p$ and $P(X_1 = -1) = 1-p$. Let $S_0 = 0$, $S_n = \sum_{k=1}^n X_k$ and let $U_n = \max_{0 \leq k \leq n} S_k$ be the running maximum of the random walk S_n. Let $M_0 = 0$ and

$$M_n = U_n - p \sum_{k=0}^{n-1} I_{\{S_k = U_k\}} \qquad \text{for } n \geq 1.$$

Show that M_n is a martingale with respect to $\{X_k : k \geq 1\}$.

Exercise 3.11. Suppose that X is a discrete random variable with finite expectation. Let $Y_n, n \geq 0$ be a sequence of discrete random variables and set $M_n = E[X|(Y_0, \ldots, Y_n)]$. Show that $M_n, n \geq 0$ is a martingale with respect to $\{Y_k : k \geq 0\}$.

Exercise 3.12. Suppose that $\{X_n : n \geq 0\}$ is a martingale with respect to the sequence $\{Y_k : k \geq 0\}$. Assume that $E[X_n^2] < \infty$ for all $n \geq 0$, and let $D_n = X_n - X_{n-1}$ for $n \geq 1$.

(a) Show that $E[D_{n+1}D_n] = 0$ for all $n \geq 1$.
Hint. Take conditional expectation with respect to $(Y_0, \ldots, Y_n)$ inside the expectation.

(b) Show that $E[D_m D_n] = 0$ for all $1 \leq m < n$.

Exercise 3.13. Let $\{X_n : n \geq 0\}$ is a Markov chain on a finite state space $\mathcal{S}$ with transition probability p, and $f : \mathcal{S} \to \mathbb{R}$ is a function. Suppose that there is a $\lambda \neq 0$ so that

$$\sum_{y \in \mathcal{S}} p(x, y) f(y) = \lambda f(x), \qquad \text{for all } x \in \mathcal{S}.$$

Show that $M_n = f(X_n)\lambda^{-n}$ defines a martingale $M_n, n \geq 0$ with respect to $\{X_k : k \geq 0\}$.

Exercise 3.14. Suppose that $\{X_k : k \geq 1\}$ is a sequence of i.i.d. random variables with $P(X_1 = 1) = P(X_1 = -1)$. Let $S_n = \sum_{k=1}^n X_k$ with $S_0 = 0$. Let $T = \min\{n \geq 1 : S_n = 1\}$ be the first hitting time of 1. Show that $P(T < \infty) = 1$.

Exercise 3.15. We perform a random walk on $\mathbb{Z}^2$ starting from $(0,0)$. For this let $X_n, n \geq 1$ be i.i.d. random variables so that X_n is one of $(0,1),(1,0),(0,-1),(-1,0)$ with probability equal probabilities. Then $S_n = \sum_{k=1}^n X_k$ is the position of the random walker after the nth step. Let R_n denote the distance from the origin after n steps. Show that $R_n^2 - n$ is a martingale

Exercise 3.16. Consider the martingale $Z_n, n \geq 0$ and the stopping time T_A defined in Example 3.28. Show that $E[T_A] < \infty$ (without computing its value) and that $|Z_{T_A \wedge (n+1)} - Z_{T_A \wedge n}|, n \geq 0$ is bounded by a deterministic value uniformly in n.
This step was needed in Example 3.28 in order to use Theorem 3.25 to compute the exact value of $E[T_A]$.
Hint. Compare T_A to the random time of seeing A in a block of the form $X_{jk+1}, \dots, X_{(j+1)k}$ for some $j \geq 0$.

Exercise 3.17. Consider the setup of Example 3.28. Let $A = (a_1, \dots, a_k)$ and $B = (b_1, \dots, b_\ell)$ be two distinct finite sequences with $a_i, b_j \in \{0,1\}$. We assume that we cannot get one of the words from the other by adding 0s or 1s at the end of the the other word. (E.g. $A = (1,0,1)$ and $B = (1,0,1,1,1)$ are not allowed.) Let T_A and T_B be waiting times to see A and B first, respectively, and let $T = T_A \wedge T_B$ be the smaller of these two waiting times.

(1) Find the probability $P(T = T_A)$.
(This is the probability of the word A 'beating' the word B.)

(2) Find the expected value $E[T]$.

Exercise 3.18. Suppose that we have a finite 'alphabet' where the set of letters is given by $\mathcal{A}$. Suppose that we fix a probability measure $p_a, a \in \mathcal{A}$ on $\mathcal{A}$, and generate i.i.d. random letters $X_n, n \geq 1$ with this distribution. Find the expected waiting time of the appearance of a finite word (or pattern) $(a_1, \dots, a_k)$ by generalizing the approach of Example 3.28.

Exercise 3.19. Consider a subcritical branching process, that is , one whose mean number of offspring satisfies $\mu < 1$. Start with $X_0 = 1$ (a single individual) and let $T = \inf\{n \geq 1 : X_n = 0\}$ denote the extinction time. Show that $ET < \infty$.
Hint. Do not attempt an exact computation but instead find a bound for $P(T > n)$ that is good enough for deducing $ET < \infty$. Recall Lemma B.4 for this purpose. To get an estimate, you might start with this: $E[X_n] = E[X_n \, I_{\{X_n \geq 1\}}] \geq P(X_n \geq 1)$.

$$P(T > n) = P(X_n \geq 1) \leq E[X_n] = \mu^n$$

from which

$$ET = \sum_{n=0}^{\infty} P(T > n) \leq \sum_{n=0}^{\infty} \mu^n = \frac{1}{1-\mu} < \infty.$$

The series converges precisely when $0 < \mu < 1$.

Chapter 4

Poisson processes

Recall that a renewal process $\{N_t : t \geq 0\}$ keeps track of the cumulative number of arrivals up to time t. The ingredients are a sequence of i.i.d. inter-arrival times $\{X_k\}_{k\geq 1}$, arrival times $T_0 = 0$, $T_n = \sum_{k=1}^{n} X_k$, and

$$N_t = \max\{n \geq 0 : T_n \leq t\}$$

that gives the number of arrivals during the interval $[0, t]$. ($T_0 = 0$ does not count as an arrival.) The special case with i.i.d. exponential waiting times is called the Poisson process.

Definition 4.1. Let λ be a positive real parameter. The rate λ **Poisson process** on $\mathbb{R}_{\geq 0}$ is the renewal process $\{N_t : t \geq 0\}$ defined with i.i.d. waiting times $\{X_k : k \geq 1\}$ where each X_k has exponential distribution with parameter λ. △

As we will see, this process has other characterizations and many useful properties. These properties follow from special properties of the exponential distribution, reviewed briefly below.

4.1. Properties of the exponential distribution

Definition 4.2. The random variable X has the **exponential distribution with parameter (rate)** λ, abbreviated $X \sim \text{Exp}(\lambda)$, if X has the probability density function $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and zero otherwise. △

The cumulative distribution function of such a random variable is $F(x) = (1 - e^{-\lambda x})$ for $x > 0$ and zero otherwise. Equivalenty, the tail probability satisfies

$$P(X > x) = 1 - F(x) = e^{-\lambda x} \quad \text{for } x \geq 0.$$

The expectation and variance of the $\mathrm{Exp}(\lambda)$ distribution can be computed using integration by parts:

$$E[X] = \int_0^\infty \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}, \qquad E[X^2] = \int_0^\infty \lambda x^2 e^{-\lambda x} dx = \frac{2}{\lambda^2},$$
$$\mathrm{Var}(X) = E[X^2] - (E[X]\,)^2 = \frac{1}{\lambda^2}.$$

Multiplying an random variable by a constant produces a new exponential random variable with a different parameter: by comparing the cumulative distribution functions or the tail probabilities, one sees that if $Y \sim \mathrm{Exp}(1)$ then $\lambda^{-1}Y \sim \mathrm{Exp}(\lambda)$.

Example 4.3 (Memoryless property of the exponential distribution)**.** If $X \sim \mathrm{Exp}(\lambda)$ and $s, t > 0$ then

$$P(X > t + s \,|\, X > s) = P(X > t).$$

This follows from the definition:

$$P(X > t + s \,|\, X > s) = \frac{P(X > t+s, X > t)}{P(X > s)} = \frac{P(X > t+s)}{P(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}.$$

In words this means that given that X is larger than s, the random variable $X - s$ has the same distribution as X. △

The exponential distribution is a special member of the two-parameter family of gamma distributions. We first introduce the *gamma function*

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx, \quad \text{for real } r > 0. \tag{4.1}$$

The gamma function generalizes the factorial function: if n is a positive integer then $\Gamma(n) = (n-1)!$.

Definition 4.4. Let $r, \lambda > 0$. A random variable X has the **gamma distribution with parameters** (r, λ) if X is positive and has probability density function

$$f(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} \qquad \text{for } x > 0, \tag{4.2}$$

with $f(x) = 0$ for $x \le 0$. We abbreviate this $X \sim \mathrm{Gamma}(r, \lambda)$.

Since $\Gamma(1) = 1$, the $\mathrm{Gamma}(1, \lambda)$ distribution is the same as the $\mathrm{Exp}(\lambda)$ distribution. The following lemma shows that we can obtain the $\mathrm{Gamma}(r, \lambda)$ distribution for integer r as the sum of r i.i.d. $\mathrm{Exp}(\lambda)$ random variables. The proof is an application of the convolution formula for absolutely continuous random variables.

Lemma 4.5. *Suppose that $X_1, \dots, X_n$ are i.i.d.* $\mathrm{Exp}(\lambda)$ *distributed random variables and $S_n = \sum_{i=1}^n X_i$. Then $S_n \sim Gamma(n, \lambda)$. In particular S_n has the probability density function*

$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} \qquad \text{for } x > 0, \tag{4.3}$$

with $f(x) = 0$ for $x \le 0$.

A consequence of this lemma is that the nth arrival time in a rate λ Poisson process has $\text{Gamma}(n, \lambda)$ distribution.

The following lemma on independent exponentials will be of importance both for Poisson processes and continuous-time Markov chains in Chapter 6.

Lemma 4.6 (Exponential races)**.** *Let $\lambda_1, \dots, \lambda_n$ be positive real numbers. Suppose that $X_1, \dots, X_n$ are independent random variables with $X_k \sim \text{Exp}(\lambda_k)$. Let $Y = \min(X_1, \dots, X_n)$ and let K be the index of the minimal X_k, that is, $Y = X_K$. Then*

Then the following statements hold.

(a)

$$Y \sim \text{Exp}\Big(\sum_{j=1}^{n} \lambda_j\Big) \tag{4.4}$$

(b) *For $1 \le k \le n$ we have*

$$P(K = k) = P(Y = X_k) = \frac{\lambda_k}{\sum_{j=1}^{n} \lambda_j}. \tag{4.5}$$

(c) *K and Y are independent.*

Imagine that X_k is the time in which racer k finishes the track. Then if the X_ks are independent exponentials, the lemma states that the winning time Y is also an exponential. Furthermore, the identity K of the winner is independent of the winning time Y.

Since the X_ks are independent and have continuous distributions, $P(X_i = X_j) = 0$ for each pair $i \neq j$. That is, with probability one, there are no ties. Consequently there is a unique minimal X_k and the random index K is well-defined.

Proof. Note that the event $\{\min(X_1, \dots, X_n) > t\}$ is the same as $\{X_1 > t, \dots, X_n > t\}$ which is the intersection of independent events. Using this we get, for $t \ge 0$,

$$\begin{aligned} P(Y > t) &= P(\min(X_1, \dots, X_n) > t) = P(X_1 > t, \dots, X_n > t) \\ &= \prod_{j=1}^{n} P(X_j > t) = \prod_{j=1}^{n} e^{-\lambda_j t} = e^{-(\sum_{j=1}^{n} \lambda_j)t}. \end{aligned}$$

Thus the tail probabilitities of Y are the same as those of an $\text{Exp}(\sum_{j=1}^{n} \lambda_j)$ distributed random variable, which identifies the distribution of Y.

We deduce an expression for the joint distribution of (K, Y) by integrating the joint density function $f(x_1, \dots, x_n) = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i x_i}$ of the random variables

$X_1, \dots, X_n$. Let $k \in \{1, \dots, n\}$ and $t \geq 0$.

$$\begin{aligned}
&P(K = k,\, Y > t) = P(X_k > t,\, X_j > X_k \text{ for } j \neq k) \\
&= \underset{\substack{x_k > t \\ x_j > x_k \text{ for } j \neq k}}{\int \cdots \int} \prod_{j=1}^{n} \lambda_j e^{-\lambda_j x_j}\, dx_1 \cdots dx_n \\
&= \int_t^\infty \lambda_k e^{-\lambda_k x_k} \Bigl(\prod_{j: j \neq k} \int_{x_k}^\infty \lambda_j e^{-\lambda_j x_j}\, dx_j \Bigr)\, dx_k \\
&= \int_t^\infty \lambda_k e^{-\lambda_k x_k} \Bigl(\prod_{j: j \neq k} e^{-\lambda_j x_k} \Bigr)\, dx_k \\
&= \int_t^\infty \lambda_k e^{-(\sum_{j=1}^n \lambda_j) x_k}\, dx_k = \frac{\lambda_k}{\sum_{j=1}^n \lambda_j} \int_t^\infty \Bigl(\sum_{j=1}^n \lambda_j \Bigr) e^{-(\sum_{j=1}^n \lambda_j) x_k}\, dx_k \\
&= \frac{\lambda_k}{\sum_{j=1}^n \lambda_j}\, e^{-(\sum_{j=1}^n \lambda_j) t}.
\end{aligned}$$

By taking $t = 0$ we get the distribution of K:

$$P(K = k) = P(K = k,\, Y > 0) = \frac{\lambda_k}{\sum_{j=1}^n \lambda_j}.$$

(By summing over k we also obtain a second proof of $P(Y > t) = e^{-(\sum_{j=1}^n \lambda_j)t}$.) The outcome of the calculation can now be stated as

$$P(K = k,\, Y > t) = P(K = k)P(Y > t).$$

Since this is true for all $k \in \{1, \dots, n\}$ and all $t \geq 0$, this identity is rich enough to imply the independence of K and Y. □

4.2. Counting measure and the Poisson point process

It is often beneficial to think about the Poisson process not in terms of the counting function N_t, but rather in terms of the collection of arrival times T_k. The next definition introduces an important concept for this setup.

Definition 4.7. The **counting measure** of a renewal process is the function $N(A)$ of subsets A of the real line that gives the number of arrival times that lie in the set A:

$$N(A) = \#\{n \geq 1 : T_n \in A\}.$$

Again, no arrival at time 0 is counted. △

For intervals $A = (a, b]$ with $0 \leq a < b$ we drop the extra parentheses from $N(A) = N((a, b])$ and simplify the notation to $N(a, b]$. In particular, we have $0 \leq a < b$ we have

$$N(a, b] = N_b - N_a.$$

To illustrate with an example, if $T_1 = 0.5$, $T_2 = 0.9$ and $T_3 = 1.2$, then $N(0.6, 0.8] = 0$, $N(0.6, 0.95] = 1$, and $N(0.4, 1.3] \geq 3$.

The next theorem describes the properties of the counting measure of the Poisson process. Item (i) accounts for the name of the process.

Theorem 4.8. *Let $\{N_t : t \geq 0\}$ be a rate λ Poisson process. Then the random variables $\{N(a,b] : 0 \leq a < b\}$ have the following properties.*

(i) *$N(a,b]$ has Poisson distribution with parameter $\lambda(b-a)$.*

(ii) *For disjoint intervals $(a_1,b_1], (a_2,b_2], \dots, (a_n,b_n]$ the random variables $N(a_1,b_1]$, $N(a_2,b_2], \dots, N(a_n,b_n]$ are independent.*

Before we discuss the proof, recall that X is Poisson(λ) distributed with $\lambda > 0$ if X takes on nonnegative integer values and

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad \text{for } k \geq 0.$$

If $X \sim \text{Poisson}(\lambda)$ then $E[X] = \text{Var}(X) = \lambda$. In particular, part (i) of the theorem above implies that in a Poisson point process of rate λ the expected number of points in an interval equals λ times the length of the interval.

If X and Y are independent, $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, then $X + Y \sim \text{Poisson}(\lambda+\mu)$. This can be shown with the convolution formula for discrete distributions, or by computing the probability generating function of $X+Y$.

The last fact about sums of independent Poissons is implicitly contained in Theorem 4.8. Namely, if $a_0 < a_1 < \cdots < a_n$ then counting arrivals implies that

$$N(a_0,a_n] = \sum_{k=0}^{n-1} N(a_k,a_{k+1}].$$

According to Theorem 4.8, we have a Poisson($\lambda(a_n - a_0)$) random variable on the left, and the sum of independent Poisson random variables with parameters $\lambda(a_{k+1}-a_k)$ on the right. Since we can choose the points a_i arbitrarily, it must be the case that sums of independent Poissons are always Poisson. Since expectations are additive in general, it must then be that the parameter of the sum equals the sum of the parameters.

We used above left-open right-closed intervals $(s,t]$ because of the convenient connection between the counting measure and the counting function: $N(s,t] = N_t - N_s$. The reader should note though that, with probability one, including or excluding the endpoints of the intervals does not change the random variable $N(s,t]$. This is because each arrival time T_n is a continuous random variable and hence takes any particular value with probability zero. So we can reason as follows to show that $N(s,t]$ and $N[s,t]$ differ with probability zero (in case $s=0$, remembering also that $T_0 = 0$ does not count as an arrival):

$$\begin{aligned} P\{N(s,t] \neq N[s,t]\,\} &= P(T_n = s \text{ for some } n \geq 1) \\ &\leq \sum_{n=1}^{\infty} P(T_n = s) = \sum_{n=1}^{\infty} 0 = 0. \end{aligned}$$

The inequality above comes from the subadditivity (1.4) of probability measures.

The Poisson distribution is valid for *all* subsets $A \subset \mathbb{R}_{\geq 0}$, as long as A has a well-defined *length* $\ell(A)$: if N_t is a rate λ Poisson process, then $N(A) \sim \text{Poisson}(\lambda\ell(A))$, in other words,

$$P\{N(A)=k\} = e^{-\lambda\ell(A)}\frac{(\lambda\ell(A))^k}{k!} \qquad \text{for } k \in \mathbb{Z}_{\geq 0}. \tag{4.6}$$

A natural example of a subset A would be a union of pairwise disjoint intervals, in which case $\ell(A)$ would be the sum of the lengths of the intervals.

'Proof' of Theorem 4.8. Denote the arrivals of the Poisson process by $\{T_k : k \geq 0\}$. By Lemma 4.5 we have $T_k \sim \text{Gamma}(\lambda, k)$. To see that $N(0,t] = N_t$ has Poisson(λt) distribution, note that $N_t = k$ means that the kth arrival happened by time t but the $(k+1)$st arrival did not:

$$P(N_t = k) = P(T_k \leq t < T_{k+1}) = P(T_{k+1} > t) - P(T_k > t).$$

This probability can be computed using the probability density function of the gamma distribution:

$$\begin{aligned} P(T_{k+1} > t) - P(T_k > t) &= \int_t^\infty \frac{\lambda^{k+1} s^k}{k!} e^{-\lambda s} ds - \int_t^\infty \frac{\lambda^k s^{k-1}}{(k-1)!} e^{-\lambda s} ds \\ &= -\int_t^\infty \lambda^k \frac{d}{ds}\left(\frac{s^k}{k!} e^{-\lambda s}\right) ds = \frac{\lambda^k t^k}{k!} e^{-\lambda t}. \end{aligned}$$

The intermediate step came from the observation that

$$\frac{d}{ds}\left(\frac{s^k}{k!} e^{-\lambda s}\right) = \frac{s^{k-1} e^{-\lambda s}}{(k-1)!} - \frac{\lambda s^k e^{-\lambda s}}{k!}.$$

This shows that $N_t \sim \text{Poisson}(\lambda t)$. Using the memoryless property of the exponential distribution one can produce an argument to show (i) and (ii) of the theorem.

For a somewhat more convincing (but still not fully rigorous) argument using the joint probability density function of the independent interarrival times of the Poisson process, see Section 4.6 of [**ASV18**]. A rigorous proof can be found in Section 4.8 of [**Res92**]. □

Remark 4.9 (Point processes)**.** Through the counting measure $\{N(A) : A \subset \mathbb{R}_{\geq 0}\}$ we make contact with a large class of stochastic processes called *point processes*. Point processes are models of collections of random points, and they are represented by counting measures. When we work with the counting measure $N(A)$ of a Poisson process rather than with its counting function N_t, we often change the terminology also and call $\{N(A) : A \subset \mathbb{R}_{\geq 0}\}$ a *Poisson point process*.

This is much more than a terminological switch. While the Poisson process N_t is a strictly one-dimensional time-indexed stochastic process, the Poisson point process $N(A)$ can be defined on higher dimensional spaces, and indeed even on arbitrary abstract spaces. △

The next theorem states that the properties of the Poisson process given in Theorem 4.8 in fact characterize or define the Poisson process. It is this characterization that can be generalized to many other spaces. For the statement of the theorem, let $\{T_k : k \geq 1\}$ be a given strictly increasing sequence of positive random variables: $0 < T_1 < T_2 < T_3 < \cdots$ Think of them as random arrival times on $(0, \infty)$. For $0 \leq a < b$ let $N(a, b]$ denote the number of random points $\{T_k\}_{k\geq 1}$ that lie in the interval $(a, b]$.

Theorem 4.10. *Suppose the collection of random variables $\{N(a,b] : 0 \leq a < b\}$ satisfy* (i) *and* (ii) *of Theorem 4.8. Define $N_0 = 0$ and $N_t = N(0,t]$ for $t > 0$. Then $\{N_t : t \geq 0\}$ is a rate λ Poisson process. In particular, the interarrival times $\{T_1, T_k - T_{k-1} : k \geq 1\}$ are i.i.d.* $\mathrm{Exp}(\lambda)$ *random variables.*

From this characterization we can show that if we start recording arrivals in a Poisson process from some time $t_0 > 0$ then the resulting process is also a Poisson process with the same rate.

Lemma 4.11. *Suppose that $\{N_t : t \geq 0\}$ is a rate λ Poisson process. Then for any fixed $t_0 > 0$ the process $M_t = N_{t+t_0} - N_{t_0}, t \geq 0$ is also a rate λ Poisson process.*

Proof. For a given $0 \leq a < b$ the number of points in $(a,b]$ for $M_t, t \geq 0$ is just the number of points in $(a+t_0, b+t_0]$ for the $\{N_t : t \geq 0\}$ process. Using this we see that properties (i) and (ii) in Theorem 4.8 are satisfied for the M_t process since they are true for the N_t process. Theorem 4.10 now implies that $M_t, t \geq 0$ is a rate λ Poisson point process. □

Example 4.12. Suppose we believe that the arrival times of frogs to a pond can be reasonably modeled by a Poisson point process. We suppose that frogs are arriving at a rate of 3 per hour.

(a) What is the probability that no frogs will arrive in the next hour?

(b) What is the probability that 12 or fewer frogs arrive in the next five hours?

(c) What is the probability that the waiting time between the fourth and the fifth frog arriving is more than half an hour?

(d) What is the probability that exactly 2 frogs arrive during the first hour, and exactly 3 frogs during the 4th hour?

(e) What is the probability that exactly 2 frogs arrive during time interval $(1,3]$, and exactly 3 frogs arrive during time interval $(2,5]$?

Each question can be answered either by the renewal process characterization or by the counting measure characterization of the process, and sometimes by both. Let the time unit be an hour. Let the arrival times be $T_n = X_1 + \cdots + X_n$ with i.i.d. $\mathrm{Exp}(3)$ interarrival times $\{X_k\}_{k\geq 1}$. Let $\{N_t : t \geq 0\}$ denote the counting process of the frogs, by assumption a Poisson process of rate 3.

(a) We need to find $P(N_1 = 0)$. Since $N_1 = N(0,1] \sim \mathrm{Poisson}(3)$, we get

$$P(N_1 = 0) = e^{-3}.$$

Equivalently, this is the probability that the first arrival comes after one hour: since $X_1 \sim \mathrm{Exp}(3)$,

$$P(T_1 > 1) = P(X_1 > 1) = e^{-3}.$$

(b) We need $P(N_5 \leq 12)$. Since $N_5 \sim \mathrm{Poisson}(3 \cdot 5)$, we get

$$P(N_5 \leq 12) = \sum_{k=0}^{12} \frac{15^k}{k!} e^{-15}.$$

(c) Directly from the interarrival times:

$$P(X_5 > 1/2) = e^{-3\cdot\frac{1}{2}} = e^{-3/2}.$$

Alternatively, if we declare the fourth arrival T_4 to be the new time origin and let $M_t = N_{T_4+t} - N_{T_4}$ denote the process of arrivals after this new time origin, this probability equals

$$P(M_{1/2} = 0) = e^{-3\cdot\frac{1}{2}} = e^{-3/2}.$$

Note that even though we put the new time origin at the *random* time T_4, the renewal process $\{M_t : t \geq 0\}$ is still a rate 3 Poisson process because it is determined by the interarrival times $\{X_k\}_{k\geq 5}$, which are i.i.d. Exp(3) random variables.

(d) The probability desired is $P(N(0,1] = 2, N(3,4] = 3)$. Since the intervals $(0,1]$ and $(3,4]$ are disjoint, the events $\{N(0,1] = 2\}$ and $\{N(3,4] = 3\}$ are independent. Hence the probability of their intersection equals the product of their individual probabilities:

$$\begin{aligned} P(N(0,1] = 2, N(3,4] = 3) &= P(N(0,1] = 2)P(N(3,4] = 3) \\ &= \frac{3^2}{2!}e^{-3} \cdot \frac{3^3}{3!}e^{-3} = \frac{81}{4}e^{-6}. \end{aligned}$$

(e) The intervals $(1,3]$ and $(2,5]$ are not disjoint. We decompose these intervals into smaller components, until we have disjoint intervals and can use independence.

$$\begin{aligned} P\big(N(1,3] = 2, N(2,5] = 3\big) &= \sum_{j=0}^{2} P\big(N(1,2] = j, N(2,3] = 2-j, N(2,5] = 3\big) \\ &= \sum_{j=0}^{2} P\big(N(1,2] = j, N(2,3] = 2-j, N(3,5] = 1+j\big) \\ &= \sum_{j=0}^{2} P(N(1,2] = j)\, P(N(2,3] = 2-j)\, P(N(3,5] = 1+j) \\ &= \sum_{j=0}^{2} e^{-3}\frac{3^j}{j!} \cdot e^{-3}\frac{3^{2-j}}{(2-j)!} \cdot e^{-6}\frac{6^{1+j}}{(1+j)!} = 54e^{-12}\sum_{j=0}^{2}\frac{6^j}{j!(2-j)!(1+j)!}. \end{aligned}$$

△

Poisson process as a limit of independent trials. The Poisson distribution appears when we are modeling counts of rare events. A representation of this fact is the following lemma that is usually covered in an introductory probability course.

Lemma 4.13. *Set $\lambda > 0$, let $X \sim$ Poisson(λ), and $S_n \sim$ Binom$(n, \frac{\lambda}{n})$ for $n \geq \lambda$. Then*

$$\lim_{n\to\infty}\sum_{k=0}^{\infty} |P(X = k) - P(S_n = k)| = 0.$$

Another way to state the lemma is that if $\varepsilon > 0$ is small and n is large then a Binom(n, ε) distributed random variable has a distribution that is close to that of a Poisson$(n\varepsilon)$ random variable. Note that the Binom(n, ε) distribution can be represented as the distribution of the sum of n i.i.d. Bernoulli(ε) distributed random variable.

The following theorem provides a quantitative estimate for this approximation, in a more general setting.

Theorem 4.14. *Let* $X \sim \text{Binom}(n,p)$ *and* $Y \sim \text{Poisson}(np)$*. Then for any subset* $A \subseteq \{0,1,2,\dots\}$*, we have*

$$\big| P(X \in A) - P(Y \in A) \big| \le np^2. \tag{4.7}$$

Proof of this theorem can be found, for example, in [Dur19].

One can extend this statement to modeling modeling the appearance of rare events in time. Here is a non-rigorous version of this theorem. If a point process (collection of random points) satisfies the following conditions, then it is a Poisson point process with rate λ:

(a) it has independent increments (this is property (ii) in Theorem 4.8),

(b) it is very unlikely that there are more than one points in a given short interval,

(c) the probability of seeing exactly one point in a given short interval is close to the length of the interval times λ.

4.3. Thinning and superposition

Imagine that a Poisson process represents the arrival of cars at a gas station. Suppose further that each driver buys a car wash with probability p, independently of the other drivers. We can split the original arrival process process into two processes: the process of cars that get a car wash, and the process of cars that do not get a car wash. How are these two processes related to each other? It turns out that they are both Poisson processes and, perhaps surprisingly, independent of each other.

To put this in mathematical terms, let $\{T_n : n \ge 1\}$ be the arrival times of a rate λ Poisson process $\{N_t : t \ge 0\}$. Each arrival is assigned a random type from the set $\{1,\dots,\ell\}$, independently of the other arrivals. The probability of type j is p_j. Assume that each $p_j > 0$ and $\sum_{j=1}^{\ell} p_j = 1$. For each j, let $N_t^{(j)}$ be the process of type j arrivals. Technically speaking, we can let $\{Y_k : k \ge 1\}$ be i.i.d. random variables with distribution $P(Y_k = j) = p_j$, independent of the Poisson process N_t. Then for each type $j \in \{1,\dots,\ell\}$, the process $\{N_t^{(j)} : t \ge 1\}$ consists of the arrival times $\{T_k : k \ge 1, Y_k = j\}$.

Theorem 4.15 (Thinning a Poisson process)**.** *For each* $j \in \{1,\dots,\ell\}$*, the process* $\{N_t^{(j)} : t \ge 0\}$ *of type* j *arrivals is a Poisson process with rate* $p_j\lambda$*. Moreover, the processes* $\{N_t^{(j)} : t \ge 0\}$*,* $1 \le j \le \ell$*, are* ℓ *mutually independent Poisson processes.*

The proof comes from a property of the Poisson distribution captured in the lemma below.

Lemma 4.16. *Let* Y *be a Poisson(*μ*) random variable. Imagine that* Y *represents the number of items in a box. The items are labeled independently with integers from the set* $\{1,\dots,\ell\}$*, so that each item receives label* j *with probability* p_j*, independently* *of the labels of the other items. For $j \in \{1, \ldots, \ell\}$, let Y_j be the number of items that receive label j. These random counting variables satisfy $Y_1 + \cdots + Y_\ell = Y$. Then $Y_1, \ldots, Y_\ell$ are independent random variables with marginal distributions $Y_j \sim \text{Poisson}(p_j\lambda)$.*

Proof. Let $n_1, \ldots, n_\ell$ be nonnegative integers and set $n = n_1 + \cdots + n_\ell$. The key is to recognize that once we condition on $Y = n$, the distribution of $Y_1, \ldots, Y_\ell$ is multinomial.

$$\begin{aligned}
&P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_\ell = n_\ell)\\
&\quad = P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_\ell = n_\ell, Y = n)\\
&\quad = P(Y = n)\, P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_\ell = n_\ell \,|\, Y = n)\\
&\quad = \frac{e^{-\mu}\mu^n}{n!} \cdot \frac{n!}{n_1!\, n_2! \cdots n_\ell!}\, p_1^{n_1} p_2^{n_2} \cdots p_\ell^{n_\ell}\\
&\quad = \frac{e^{-p_1\mu}\,(p_1\mu)^{n_1}}{n_1!} \cdot \frac{e^{-p_2\mu}\,(p_2\mu)^{n_2}}{n_2!} \cdots \frac{e^{-p_\ell\mu}\,(p_\ell\mu)^{n_\ell}}{n_\ell!}.
\end{aligned}$$

In the last step we split the exponential as $e^{-\mu} = e^{-p_1\mu} \cdots e^{-p_\ell\mu}$, split the power of μ as $\mu^n = \mu^{n_1} \cdots \mu^{n_\ell}$ and recombined factors. The last line above is a product of $\text{Poisson}(p_j\mu)$ probabilities. It identifies the joint distribution of $\{Y_j : 1 \le j \le \ell\}$ and shows that these random variables are independent. □

Proof of Theorem 4.15. For this proof the counting measure characterization given in Theorem 4.8 is convenient. Since arrivals in distinct time intervals are independent, it is enough to check that in a given interval the random labeling produces independent Poisson random variables with the correct parameters. This is a consequence of Lemma 4.16. □

Example 4.17. Suppose cars arrive at a gas station as a rate λ Poisson process, with time unit of hours. Assume each driver buys a car wash with probability p, independently of the other drivers.

(a) Given that four cars arrived at the gas station between 1 PM and 2 PM, what is the probability that exactly three of them bought a car wash?

This question is answered directly from the description of the model. The answer does not involve the time spent at all or the rate λ, but only the independent labels of *wash* and *no wash*. Given that there are four arrivals, the number of *wash* labels has the binomial distribution $\text{Bin}(4, p)$. Let N_t be the process of all arrivals, $N_t^{(w)}$ the process of arrivals that get a wash, and $N_t^{(nw)}$ the process of arrivals that decline the car wash.

$$P\big(N^{(w)}(13, 14] = 3 \,\big|\, N(13, 14] = 4\big) = 4p^3(1-p).$$

(b) What is the probability that between 9 AM and 11 AM, exactly 5 cars get a car wash, and between 10 AM and 2 PM, exactly 2 cars arrive at the gas station without buying a car wash?

According to Theorem 4.15, $N_t^{(w)}$ is a rate $p\lambda$ Poisson process, $N_t^{(nw)}$ is a rate $(1-p)\lambda$ Poisson process, and the two processes are independent. We find the answer as follows:

$$\begin{aligned}&P\{N^{(w)}(9,11)=5,\ N^{(nw)}(10,14)=2\}\\&\qquad=P\{N^{(w)}(9,11)=5\}\,P\{N^{(nw)}(10,14)=2\}\\&\qquad=e^{-2p\lambda}\frac{(2p\lambda)^5}{5!}\cdot e^{-4(1-p)\lambda}\frac{(4(1-p)\lambda)^2}{2!}=\frac{32}{15}e^{-(4-2p)\lambda}p^5(1-p)^2\lambda^7.\end{aligned}$$

(c) What is the probability that between 8 AM and 11 AM, at least 5 cars bought a car wash?

This question involves only the process $N_t^{(w)}$.

$$P\{N^{(w)}(8,11)\ge 5\}=\sum_{k=5}^{\infty}P\{N^{(w)}(8,11)=k\}=\sum_{k=5}^{\infty}e^{-3p\lambda}\frac{(3p\lambda)^k}{k!}.$$

△

The opposite of splitting a Poisson process into separate streams would be to combine several processes into a single process. This is called *superposition*. Now we start with ℓ independent Poisson processes $\{N_t^{(j)} : t \ge 0\}$ where the index j ranges over $\{1,\dots,\ell\}$. Let λ_j be the rate of $N_t^{(j)}$. Set $N_t = \sum_{j=1}^{\ell} N_t^{(j)}$. N_t is the process formed by combining all the arrivals together. Equivalently, the set of arrivals of N_t is the union of the arrivals of the processes $N_t^{(j)}$, $1 \le j \le \ell$. Let $\lambda = \sum_{j=1}^{\ell}\lambda_j$ be the sum of the rates.

Theorem 4.18 (Superposition of independent Poisson processes)**.** *The process $\{N_t : t \ge 0\}$ is a Poisson process with rate λ. Moreover, for each specific arrival of N_t, the probability that this particular arrival came from the process $N_t^{(j)}$ is λ_j/λ, independently of the other arrivals.*

The proof follows again by checking the conditions of Theorem 4.8.

Remark 4.19. The exponential race of Lemma 4.6 is closely associated with Theorem 4.18. If $T_1^{(j)}$ denotes the first arrival time of the jth process $N_t^{(j)}$, then $T_1^{(j)} \sim \text{Exp}(\lambda_j)$ and these are independent. The first arrival T_1 of N_t is the minimum of the $T_1^{(j)}$ variables. Since N_t is a rate λ Poisson process, $T_1 = \min(T_1^{(1)},\dots,T_1^{(\ell)}) \sim \text{Exp}(\lambda)$. The event that the first arrival is in the jth process is exactly $T_1^{(j)} = \min(T_1^{(1)},\dots,T_1^{(\ell)})$, and by the theorem this has probability λ_j/λ. △

Example 4.20. Suppose cars go by a house as a rate λ Poisson process $N_t^{(c)}$ and vans go by as a rate μ Poisson process $N_t^{(v)}$. Let the time unit be an hour. Assume that the two Poisson processes $N^{(c)}$ and $N^{(v)}$ are independent of each other.

(a) What is the probability that exactly five vehicles went by between 9 AM and 11 AM?

According to Theorem 4.18, the combined process $N_t = N_t^{(c)} + N_t^{(v)}$ of vehicles is a rate $\lambda+\mu$ Poisson process. Hence

$$P\{N(9,11)=5\}=e^{-2(\lambda+\mu)}\frac{2^5(\lambda+\mu)^5}{5!}=\frac{4}{15}e^{-2(\lambda+\mu)}(\lambda+\mu)^5.$$

(b) What is the probability that exactly two cars and three vans went by between 9 AM and 11 AM?

If we specify the types of the vehicles, then we have to use the original processes and their independence.

$$\begin{aligned}
&P\{N^{(c)}(9,11)=2,\, N^{(v)}(9,11)=3\}\\
&\quad = P\{N^{(c)}(9,11)=2\}\, P\{N^{(v)}(9,11)=3\}\\
&\quad = e^{-2\lambda}\frac{2^2\lambda^2}{2!}\cdot e^{-2\mu}\frac{2^3\mu^3}{3!} = \frac{8}{3}e^{-2(\lambda+\mu)}\lambda^2\mu^3.
\end{aligned}$$

(c) Given that five vehicles went by during a two-hour period, what is the probability that exactly two of these were cars?

According to Theorem 4.18, each vehicle is a car with probability $\lambda/(\lambda+\mu)$ independently of the other vehicles. Thus the time period does not enter the answer at all. Instead we simply ask, what is the probability that out of five independent vehicle types, exactly two are cars. This is a question about a binomial distribution:

$$P(\text{exactly 2 cars}\,|\,\text{5 vehicles}) = \binom{5}{2}\left(\frac{\lambda}{\lambda+\mu}\right)^2\left(\frac{\mu}{\lambda+\mu}\right)^3.$$

△

4.4. Conditioning the Poisson process on the number of arrivals

In this section we derive the distribution of the arrival times under a conditioning on the total number of arrivals in an interval. We start with a calculation.

Example 4.21. Let $\{N_t : t \ge 0\}$ be a rate λ Poisson process. Fix time values $0 < r < s < t$. Find the distribution of $N(r,s]$, given that $N_t = n$.

Under this conditioning the values of $N(r,s]$ are restricted to the range $0,1,\dots,n$ and so $N(r,s]$ cannot be a Poisson variable any more. Let $k \in \{0,1,\dots,n\}$.

$$\begin{aligned}
P\big(N(r,s]=k \,\big|\, N_t=n\big) &= \frac{P\{N(r,s]=k,\, N_t=n\}}{P(N_t=n)}\\
&= \frac{P\{N(r,s]=k,\, N((0,r]\cup(s,t])=n-k\}}{P(N_t=n)}\\
&= \frac{e^{-\lambda(s-r)}\frac{(\lambda(s-r))^k}{k!}\cdot e^{-\lambda(r+t-s)}\frac{(\lambda(r+t-s))^{n-k}}{(n-k)!}}{e^{-\lambda t}\frac{(\lambda t)^n}{n!}}\\
&= \binom{n}{k}\left(\frac{s-r}{t}\right)^k\left(\frac{r+t-s}{t}\right)^{n-k}.
\end{aligned}$$

Above we applied (4.6) to the set $A=(0,r]\cup(s,t]$ whose length is $r+t-s$. The independence in the numerator is a consequence of the disjointness of $(r,s]$ and A. In the last step the exponentials and the powers of λ were cancelled.

The calculation shows that, conditionally on $N_t=n$, $N(r,s]\sim\text{Bin}(n,\frac{s-r}{t})$. To interpret the result, note that $\frac{s-r}{t}$ is the probability of $X\in(r,s]$ for a random variable X that is uniformly distributed on $[0,t]$. Thus we can paraphrase the outcome as follows: given $N_t = n$, the random number of arrivals in $(r,s]$ is determined by putting down n i.i.d. uniform points in $[0,t]$ and counting how many land in $(r,s]$.

An important qualitative conclusion is also that the conditional distribution of $N(r,s]$ does not involve λ. This points to an important feature of the Poisson process: the role of the parameter λ is only to modulate the number of arrivals in a given interval, and once this number has been chosen from a Poisson distribution, λ does not influence how the arrivals are spread out in the interval. △

The next theorem captures this result. Note however an important subtlety. The *ordered* arrival times $0 < T_1 < \cdots < T_n$ cannot be put down independently because that would not guarantee the ordering. This task is addressed in part (b) of the theorem.

Theorem 4.22. *Let $\{N_t : t \geq 0\}$ be a rate λ Poisson process, and denote the arrival times by $\{T_n : n \geq 1\}$. Fix a real $t > 0$ and a positive integer n. Then, conditionally on the event $N_t = n$, the following is true.*

(a) *As an unordered set, the arrival times in $[0,t]$ are distributed as n independent Unif$[0,t]$ random variables.*

(b) *The joint probability density function of the vector of ordered arrival times $(T_1,\dots,T_n)$ equals the constant $n!/t^n$ on the set $W_{n,t} = \{(t_1,\dots,t_n) : 0 < t_1 < \cdots < t_n < t\}$, and zero elsewhere.*

Some clarification of the theorem.

Echoing Example 4.21 above, part (a) applies for example to questions about the numbers of arrivals in subsets of $[0,t]$. To illustrate, let $0 < r < s < t$. Since a Unif$[0,t]$ variable lies in $(r,s]$ with probability $\frac{s-r}{t}$,

$$P(N(r,s] = 3 \,|\, N_t = 5) = \binom{5}{3}\left(\frac{s-r}{t}\right)^3\left(1 - \frac{s-r}{t}\right)^2.$$

Thus under the conditioning, the number of arrivals in a given interval inside $(0,t]$ is not Poisson but binomial. Furthermore, under the conditioning the numbers of arrivals in disjoint intervals inside $(0,t]$ cannot be independent since they constrain each other. For example, conditional on $N_t = 5$, $N(s,t] = 5 - N(0,s]$ and consequently $N(s,t]$ and $N(0,s]$ are not conditionally independent.

In part (b) the set $W_{n,t}$ of ordered n-vectors is the set where the random vector $(T_1,\dots,T_n)$ takes its values, and hence the joint density function must vanish outside $W_{n,t}$.

Parts (a) and (b) of the theorem are connected by the following general fact about uniform random variables. Let $(U_1,\dots,U_n)$ be i.i.d. Unif$[0,t]$ random variables, and let $V_1 < \cdots < V_n$ be the increasing rearrangement of this sequence. In other words, V_1 is the smallest value among the $U_1,\dots,U_n$, V_2 the second smallest, and so on. $(V_1,\dots,V_n)$ are called the *order statistics* of $(U_1,\dots,U_n)$. Then it follows that the joint density function f of $(V_1,\dots,V_n)$ is given by

$$f(t_1,\dots,t_n) = \begin{cases} n!/t^n, & (t_1,\dots,t_n) \in W_{n,t} \\ 0, & \text{otherwise.} \end{cases}$$

The appearance of the density function $n!/t^n$ is explained by the observation that $n!/t^n$ is the reciprocal of the *volume* $t^n/n!$ of $W_{n,t}$. Since the density function of the random vector $(V_1,\dots,V_n)$ is the reciprocal of the volume of its range $W_{n,t}$, this vector is *uniformly distributed on the set* $W_{n,t}$.

That the volume of $W_{n,t}$ equals $t^n/n!$ is checked by integration. You can also reason it in a probabilistic way. The volume of the n-dimensional cube $[0,t]^n$ is t^n. This is the range of the vector $(U_1,\dots,U_n)$ of Unif$[0,t]$ random variables. Every ordering of $(U_1,\dots,U_n)$ is equally likely. Thus $(U_1,\dots,U_n)$ lies in $W_{n,t}$ with probability $1/n!$. Hence the volume of $W_{n,t}$ must be $(1/n!)\cdot t^n$.

Example 4.23. We illustrate the use of parts (a) and (b) of Theorem 4.22. Let $0<r<t$ and compute the probability $P(T_2>r\,|\,N_t=2)$, namely that conditional on two arrivals in $[0,t]$, the second arrival comes after time r.

We calculate this probability first by using part (a) and counting arrivals:

$$\begin{aligned}P(T_2>r\,|\,N_t=2)&=P(N_r\le 1\,|\,N_t=2)\\&=P(N_r=0\,|\,N_t=2)+P(N_r=1\,|\,N_t=2)\\&=\left(\frac{t-r}{t}\right)^2+2\,\frac{r}{t}\cdot\frac{t-r}{t}=\frac{t^2-r^2}{t^2}.\end{aligned}$$

The second calculation integrates the joint density function $2!/t^2$ of (T_1,T_2) over the set $W=\{(t_1,t_2):0<t_1<t_2<t\}$, subject to the further restriction that $t_2>r$.

$$\begin{aligned}P(T_2>r\,|\,N_t=2)&=\frac{2!}{t^2}\iint\limits_{\substack{0<t_1<t_2\le t\\ t_2>r}}dt_1\,dt_2\\&=\frac{2!}{t^2}\int_r^t\left(\int_0^{t_2}dt_1\right)dt_2=\frac{2!}{t^2}\int_r^t t_2\,dt_2=\frac{t^2-r^2}{t^2}.\end{aligned}$$

From this calculation we can derive the conditional cumulative distribution function of T_2, given that $N_t=2$:

$$F_{T_2}(s\,|\,N_t=2)=P(T_2\le s\,|\,N_t=2)=\frac{s^2}{t^2},\quad 0\le s\le t,$$

and then its conditional density function:

$$f_{T_2}(s\,|\,N_t=2)=\frac{d}{ds}F_{T_2}(s\,|\,N_t=2)=\frac{2s}{t^2},\quad 0<s<t.$$

△

Since a Poisson process can be restarted at any time point, Theorem 4.22 applies equally well to any interval $(a,b]$. Namely, conditionally on $N(a,b]=n$, the unordered arrival times behave like uniform random variables on $(a,b]$.

Proof of Theorem 4.22. Here is a somewhat formal proof which can be turned rigorous. Let $0<x_1<\cdots<x_n<t$ and suppose that $\varepsilon>0$ is smaller than $x_1, t-x_n$ and $x_{i+1}-x_i$ for all i. We will compute the conditional probability

$$P(N(x_1,x_1+\varepsilon]=1,N(x_2,x_2+\varepsilon]=1,\dots,N(x_n,x_n+\varepsilon]=1\,|\,N_t=n).\tag{4.8}$$

If we can show that this is close to $n!\varepsilon^n t^{-n}$ then this would indicate that the conditional distribution of $T_1, \dots, T_n$ is the same as the ordered version of n independent Uniform$[0,t]$ random variables. We can compute the conditional probability (4.8) from the definition. Let $P(A|B)$ denote the probability in (4.8), with A and B denoting the two events that appear there. So $P(B) = P(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$.

The event $A \cap B$ is the same as having no arrivals in the intervals $[0, x_1]$, $(x_1 + \varepsilon, x_2]$, $\dots$, $(x_n + \varepsilon, t]$, and exactly one arrival in each of the intervals $(x_j, x_j + \varepsilon]$. Since these are non-overlapping intervals, the corresponding events are independent, and we can compute the probability of their intersection by multiplication:

$$\begin{aligned} P(A \cap B) &= e^{-\lambda x_1} e^{-\lambda(x_2 - x_1 - \varepsilon) \cdots e^{-\lambda(t - x_n - \varepsilon)}} \prod_{j=1}^{n} \lambda \varepsilon e^{-\lambda \varepsilon} \\ &= \lambda^n \varepsilon^n e^{-\lambda t}. \end{aligned}$$

But this gives

$$P(A|B) = \frac{\lambda^n \varepsilon^n e^{-\lambda t}}{\frac{(\lambda t)^n}{n!} e^{-\lambda t}} = \varepsilon^n n! \frac{1}{t^n},$$

which is exactly what we wanted. □

Example 4.24. Let $\{N_t : t \geq 0\}$ be a rate λ Poisson point process. For $0 < s < t$ find the conditional distribution of N_s given $N_t = n$.

The arrivals in $[0, t]$ are distributed as n independent Uniform$[0, t]$ random variables. For each one of these the probability of being in $[0, s]$ is $\frac{s}{t}$. Hence the conditional distribution of N_s given $N_t = n$ is binomial with parameters $(n, \frac{s}{t})$. △

4.5. Generalizations: inhomogeneous and spatial Poisson point processes

Recall the counting measure characterization of Theorem 4.8 of the rate λ Poisson point process. Note that condition (i) can be rewritten to say that the number of points in an interval $(a, b]$ is given by a Poisson distribution where the parameter is the length of the interval times λ.

There are several ways we can generalize this definition. One possibility is to measure the length of intervals using the integral of a non-negative function. Thus leads to inhomogeneous Poisson processes.

Definition 4.25 (Inhomogeneous Poisson process on $\mathbb{R}_{\geq 0}$)**.** Let $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a function with the property that $\int_a^b \lambda(s) ds < \infty$ for all $0 \leq a < b$. We say that a point process on $\mathbb{R}_{\geq 0}$ is a Poisson process with intensity $\lambda(t)$ if

- With probability one there are no overlapping points.
- For any $0 \leq a < b$ the number of points in $(a, b]$ is Poisson distributed with parameter $\int_a^b \lambda(s) ds$.
- If $A_1, A_2, \dots A_n$ are non-overlapping finite intervals then the number of points in the respective intervals are independent.

Another way to extend the definition of a Poisson process is to consider point processes in higher dimensions. This can be used to model rare events which happen in a two or three dimensional region (or in space-time), e.g. the positions (or positions and times) of earthquakes in a region.

Definition 4.26 (Spatial Poisson process)**.** A homogeneous spatial Poisson process in $\mathbb{R}^n$ with parameter λ is a collection of random points in $\mathbb{R}^n$ with the following properties:

- With probability one there are no overlapping points.
- In any bounded region A the number of points is Poisson distributed with parameter $\lambda \cdot \text{area}(A)$.
- If $A_1, A_2, \dots A_n$ are non-overlapping regions with finite areas then the number of points in the respective regions are independent.

Example 4.27. You are standing in a forest where the trees' positions form a spatial Poisson point process with parameter λ. Order the distances of the trees from your position in increasing order and denote them by $T_1, T_2, \dots$. What is the distribution of T_n?

Note that $\{T_n \le x\}$ is the same event as having at least n points in the disk of radius x with your position as its center. The number of points in that disk is Poisson distributed with parameter $\lambda \cdot (\text{area of the disk}) = \lambda\pi x^2$, hence

$$P(T_n \le x) = \sum_{k=n}^{\infty} \frac{(\lambda\pi x^2)^k}{k!} e^{-\lambda\pi x^2} = 1 - \sum_{k=0}^{n-1} \frac{(\lambda\pi x^2)^k}{k!} e^{-\lambda\pi x^2}$$

This gives the cumulative distribution function of T_n. We can also get the density function by differentiating this:

$$\begin{aligned}\frac{d}{dt} P(T_n \le x) &= -\sum_{k=0}^{n-1} \left(-2x\pi\lambda \frac{(\lambda\pi x^2)^k}{k!} e^{-\lambda\pi x^2} + 2x\pi\lambda k \frac{(\lambda\pi x^2)^{k-1}}{k!} e^{-\lambda\pi x^2} \right) \\ &= \sum_{k=0}^{n-1} \left(2x\pi\lambda \frac{(\lambda\pi x^2)^k}{k!} e^{-\lambda\pi x^2} - 2x\pi\lambda \frac{(\lambda\pi x^2)^{k-1}}{(k-1)!} e^{-\lambda\pi x^2} \right) \\ &= 2x\pi\lambda \frac{(\lambda\pi x^2)^{n-1}}{(n-1)!} e^{-\lambda\pi x^2}\end{aligned}$$

(The square of T_n would be a Gamma distributed random variable.) △

Of course, we can also combine spatial and inhomogeneous Poisson point processes. This would require a function $\lambda : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$.

Example 4.28. In a forest there are trees with an average density of 5 trees per 100 yd^2. Assume that the tree trunks are perfect disks with a radius of 1 foot. (1yd = 3ft.) We are standing 100 yd from the edge inside the forest and we shoot a bullet towards the edge. What is the probability that we will hit one of the trees? (We will disregard the problem that trees might 'overlap' in that model. The bullet can be considered as a point, and its path is a straight line.)

The centers of the trees form a (spatial) Poisson process with parameter $5/100yd^2$. Consider the rectangle which is formed by a 1foot region on both side of the 100 yd path of the bullet. The bullet will not hit any trees if none of the centers of the trees are in this rectangle. The area of this rectangle is $2/3 \cdot 100 = \frac{200}{3} yd^2$. The number of tree centers X in this rectangle is a Poisson random variable with parameter $5/100 \frac{200}{3} = \frac{10}{3}$. Thus the probability in question is $P(X > 0) = 1 - e^{-10/3}$. $\triangle$

Chapter 5

Return to renewal processes

Suppose that buses arriving to a bus stop can be modeled by a renewal process: $T_n, n \geq 1$ are the arrival times and $N_t, t \geq 0$ is the corresponding renewal process counting the arrivals up to time t. Suppose that you arrive at the bus stop at a given time $s > 0$. How long do you have to wait until the next bus?

There have been N_s buses that arrived up to time s, so the next bus arrives at T_{N_s+1}. Our wait time is

$$Z_s = T_{N_s+1} - s. \tag{5.1}$$

This quantity is also called the residual life at time s

Another related quantity is the time from the previous bus (or zero if there were no buses):

$$A_s = s - T_{N_s}, \tag{5.2}$$

here T_0 is defined as 0. This is called the age at time s. (Note that A_s cannot be larger than s.)

What can we say about these random variables for a fixed s or as $s \to \infty$?

5.1. Age and residual life in the Poisson process

In the case of the rate λ Poisson process we can compute the distribution of Z_s and A_s directly, and even identify the joint distribution.

For given $0 \leq x < s, 0 \geq y$ the event $\{A_s > x, Z_s > y\}$ is the same as the event of having no arrivals in the interval $[s - x, s + y]$. The probability of this can be computed using the counting measure representation of the Poisson process:

$$P(A_s > x, Z_s > y) = P(N(s - x, s + y] = 0) = e^{-\lambda(x+y)}.$$

We can get the distribution of Z_s by setting $x = 0$:

$$P(Z_s > y) = e^{-\lambda y}.$$

This shows that $Z_s \sim \text{Exp}(\lambda)$. The distribution of A_s can be obtained similarly, the only difference is that A_s cannot be larger than s:

$$P(A_s > x) = e^{-\lambda x} \qquad \text{if } 0 \le x < s, \qquad P(A_s = s) = P(N_s = 0) = e^{-\lambda s}.$$

Hence A_s has the same distribution as $\min(T, s)$ where $T \sim \text{Exp}(\lambda)$. Since $P(A_s > x, Z_s > y) = P(A_s > x)P(Z_s > y)$, we also get that A_s and Z_s are independent in this case.

As $s \to \infty$ the distribution of Z_s remains the same, and the distribution of A_s gets closer and closer to that of an $\text{Exp}(\lambda)$ random variable. This proves the following statement.

Theorem 5.1. *Suppose that $N_t, t \ge 0$ is a rate λ Poisson process. Then for a fixed $s > 0$ the random variables A_s, Z_s are independent, $Z_s \sim$ Exp(λ) and $A_s \sim \min(T, s)$ where $T \sim$ Exp(λ). As $s \to \infty$ the joint distribution of (A_s, Z_s) converges to two independent Exp(λ) random variables.*

A consequence of this theorem is that if the buses arrive according to a rate λ Poisson point process then the expected wait time for the next bus is always $\frac{1}{\lambda}$. This is somewhat strange, as this is also the expected wait time between any two consecutive buses! We will revisit this issue in the next section in a more general setting.

5.2. Age and residual life in the general case

Consider now a general renewal process. Again, we would like to understand the joint distribution of A_s, Z_s, so it would be useful to compute the probability $P(A_s > x, Z_s > y)$ for given $0 \le x < s, 0 \ge y$. By considering the possible values for the number of arrivals up to s, we can express this probability using the distributions of the arrival times $T_n, n \ge 1$:

$$\begin{aligned} P(A_s > x, Z_s > y) &= P(\text{no arrivals in } [s-x, s+y]) \\ &= \sum_{k=0}^{\infty} P(\text{no arrivals in } [s-x, s+y],\, k \text{ arrivals up to } s) \\ &= \sum_{k=0}^{\infty} P(T_k \le s-x, s+y < T_{k+1}) \end{aligned}$$

In general, the last sum is not easy to evaluate. Because of this, we will try to solve a similar, but easier question: for given $x, y > 0$ on average in what proportion of time will it be true that $A_s > x$, and $Z_s > y$? More precisely, we are interested in the long term behavior of the quantity

$$\frac{1}{t}\int_0^t I(A_s > x, Z_s > y)ds$$

It turns out that we can evaluate the limit of this quantity using a renewal reward process. Denote the i.i.d. interarrival times by $X_n, n \ge 1$, and the mean interarrival time by μ. Let R_n be the time during $(T_{n-1}, T_n]$ when the age is larger than x and the residual life is less than y:

$$R_n = \int_{T_{n-1}}^{T_n} I(A_s > x, Z_s > y)ds$$

If we consider this as the 'reward' of the nth cycle, and disregard the contribution of the last incomplete cycle, the renewal reward limit theorem tells us that

$$\frac{1}{t}\int_0^t I(A_s > x, Z_s > y)ds \to \frac{E[R_1]}{\mu},$$

with probability one.

To evaluate $E[R_1]$, first note that if $T_1 \leq x + y$ then there is no time when the the age is larger than x and the residual life is less than y. Otherwise the length of time when this happens is exactly $X_1 - (x + y)$. Thus $R_1 = (X_1 - (x + y))^+$ and

$$E[R_1] = E[(X_1 - (x + y))^+].$$

This leads to the following result.

Theorem 5.2. *Let $N_t, t \geq 0$ be a renewal process with expected mean interarrival times $X_n, n \geq 1$. Then*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t I(A_s > x, Z_s > y)ds = \frac{E[(X_1 - (x + y))^+]}{E[X_1]}$$

with probability one.

In the case when X_1 is absolutely continuous with probability density function f then we can write

$$\begin{aligned} E[(X_1 - (x + y))^+] &= \int_{x+y}^{\infty} (z - (x + y))f(z)dz = \int_{x+y}^{\infty}\int_{x+y}^{z} f(z)du\, dz \\ &= \int_{x+y}^{\infty}\int_{z}^{\infty} f(z)du\, dz = \int_{x+y}^{\infty} P(X_1 > z)dz. \end{aligned}$$

It turns out that the last formula holds without assuming absolute continuity:

$$E[(X_1 - (x + y))^+] = \int_{x+y}^{\infty} P(X_1 > z)dz.$$

By setting $x = 0$ or $y = 0$ we get the following two limits (both hold with probability one):

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{t}\int_0^t I(A_s > x)ds &= \frac{\int_x^\infty P(X_1 > z)dz}{E[X_1]}, \\ \lim_{t\to\infty} \frac{1}{t}\int_0^t I(Z_s > y)ds &= \frac{\int_y^\infty P(X_1 > z)dz}{E[X_1]}. \end{aligned}$$

It turns out that these limits hold in a stronger sense as well, if we assume that X_1 does not take its values from an arithmetic progression.

Theorem 5.3. *Assume that $N_t, t \geq 0$ is a renewal process with i.i.d. interarrival times $X_n, n \geq 1$. Assume that there for each $a > 0$ we have*

$$P(X_1 \in \{a, 2a, 3a, \dots\}) < 1.$$

Then the following limits (in distribution) hold:

$$\lim_{t\to\infty} P(A_t > x, Z_t > y) = \frac{\int_{x+y}^{\infty} P(X_1 > z)dz}{E[X_1]}$$
$$\lim_{t\to\infty} P(A_t > x) = \frac{\int_{x}^{\infty} P(X_1 > z)dz}{E[X_1]}$$
$$\lim_{t\to\infty} P(Z_t > y) = \frac{\int_{y}^{\infty} P(X_1 > z)dz}{E[X_1]}.$$

The theorem states that the age and residual life both converge in distribution. The limiting distribution of A_t and Z_t both have tail probabilities given by the formula $\frac{\int_x^\infty P(X_1>z)dz}{E[X_1]}$. Note that

$$\frac{\int_x^\infty P(X_1 > z)dz}{E[X_1]} = \int_x^\infty \frac{P(X_1 > z)}{E[X_1]}dz = \int_x^\infty g(z)dz$$

with

$$g(z) = \frac{P(X_1 > z)}{E[X_1]}. \tag{5.3}$$

We claim that $g(z)$ is a probability density function on $\mathbb{R}_+$, we just need to check $\int_0^\infty g(z)dz = 1$. Assume for simplicity that X_1 is absolutely continuous with probability density function f. Then

$$\int_0^\infty g(z)dz = \int_0^\infty \frac{\int_z^\infty f(u)du}{E[X_1]}dz = \frac{1}{E[X_1]}\int_0^\infty uf(u)du = \frac{E[X_1]}{E[X_1]} = 1.$$

(We switched the order of integration in the second step.) It can be shown that this identity holds without the absolute continuity assumption as well.

To restate the results of Theorem 5.3: under the stated conditions both the age and the residual life converges in distribution to an absolutely continuous random variable with probability density function given by g from (5.3).

Example 5.4. We have already seen that in the case of a rate λ Poisson process the residual life Z_s (and hence its limit as well) has the same $\text{Exp}(\lambda)$ distribution as the interarrival time.

We will show that if the interarrival times are absolutely continuous and the limiting residual life distribution has the same distribution as the original interarrival times then those interarrival times must be exponentially distributed. Hence the Poisson process is the only one with this special property.

now check that if the interarrival times are absolutely continuous then this is the only case when this happens. Denote the probability density function of X_1 by f. Then (5.3) gives

$$g(z) = \frac{\int_z^\infty f(y)dy}{E[X_1]}.$$

Differentiating both sides gives

$$g'(z) = -\frac{f(z)}{E[X_1]}.$$

If $g(z) = f(z)$ then we get the differential equation

$$f'(z) = -\frac{f(z)}{E[X_1]}$$

which has the general solution $f(z) = ce^{-\frac{z}{E[X_1]}}$, where c can be any parameter. Since f is the probability density function of a nonnegative random variable we must have $\int_0^\infty f(z)dz = 1$ which gives

$$1 = \int_0^\infty ce^{-\frac{z}{E[X_1]}}dz = E[X_1]c,$$

and $c = \frac{1}{E[X_1]}$. This means that X_1 must be exponentially distributed with parameter $\frac{1}{E[X_1]}$. △

Example 5.5 (The inspection paradox)**.** The sum of the residual life and the age is always equal to the corresponding interarrival time. Let us compute the expected value of the limiting distribution of the residual life (and age).

This is given by

$$\int_0^\infty zg(z)dz = \frac{1}{E[X_1]}\int_0^\infty zP(X_1 > z)dz.$$

Assuming for simplicity that X_1 has a probability density function f we can continue with

$$\begin{aligned}\frac{1}{E[X_1]}\int_0^\infty zP(X_1 > z)dz &= \frac{1}{E[X_1]}\int_0^\infty z\int_z^\infty f(u)du\,dz\\ &= \frac{1}{E[X_1]}\int_0^\infty \int_0^z udu f(z)\,dz\\ &= \frac{1}{E[X_1]}\int_0^\infty \frac{u^2}{2}f(u)du = \frac{E[X_1^2]}{2E[X_1]}.\end{aligned}$$

(The last formula can be proved without the absolute continuity assumption.) This is the expected value of the limiting distribution of the residual life and also the limiting distribution for the age. Hence the expected value of the limiting distribution of the current interarrival time is $2\frac{E[X_1^2]}{2E[X_1]} = \frac{E[X_1^2]}{E[X_1]}$. But if X_1 is not a constant then $E[X_1]^2 < E[X_1^2]$, which implies $\frac{E[X_1^2]}{E[X_1]} > E[X_1]$.

Hence the expected value of the limiting distribution of the interarrival time (or cycle length) corresponding to t is larger than the expected value of the the i.i.d. interarrival times! Using the bus stop model, this implies that the waiting time between the bus that just left and the one that will arrive next in expectation is larger than the the expected wait time between buses!

To resolve the paradox consider the following argument. The interarrival times are random: some are larger than the mean, some are shorter. On a long run, the cycles that are larger than average take up more time on the timeline than the ones that are shorter. Hence if we go to the bus stop, it is more likely that we arrive during one of the unusually long interarrival times, which explains why in expectation we observe a longer cycle. △

In some cases even the expected value of the limiting residual age can be larger than the expected interarrival time.

Example 5.6. Suppose that buses are arriving to a bus stop according to a renewal process, with i.i.d. interarrival times that have a probability density function $f(x) = \frac{1}{(\log a)x}$ on $1 \le x \le a$, and 0 otherwise. What is the expected wait time if we go down to the bus stop?

The first and second moments of the interarrival times are given by

$$E[X_1] = \int_1^a \frac{1}{\log a} dx = \frac{a-1}{\log a}, \qquad E[X_1^2] = \int_1^a \frac{x}{\log a} dx = \frac{a^2-1}{2\log a}.$$

Hence the expected value of limiting distribution of the residual age is

$$\frac{E[X_1^2]}{2E[X_1]} = \frac{\frac{a^2-1}{2\log a}}{2\frac{a-1}{\log a}} = \frac{a+1}{4}.$$

If $a > 1$ is large enough then $\frac{a+1}{4} > \frac{a-1}{\log a}$, which means that the expected wait time for the next bus will be larger than the expected interarrival time! △

The setup of the renewal process can be generalized by allowing the first arrival to have a different distribution then the others. In this case $X_n, n \ge 0$ are independent, $X_n, n \ge 1$ are i.i.d., and the nth arrival is $T_n = \sum_{k=0}^{n-1} X_k$. This process is sometimes called delayed renewal process.

The changing of the distribution of the first waiting time does not change the asymptotic behavior of the process, all of the results that we discussed will still hold. What is interesting is that if we choose the distribution of X_0 so it has probability density function given by g from (5.3) then the delayed process becomes stationary: the process $M_t = N_{t+t_0} - N_t, t \ge 0$ will have the same distribution as $N_t, t \ge 0$. Note that this is just a generalization of the stationarity property of the rate λ Poisson process.

Chapter 6

Continuous-time Markov chains

The previous chapters studied Markov chains and martingales in discrete time and renewal processes and Poisson processes in continuous time. This chapter develops Markov chains in continuous time. The resulting processes combine features of discrete-time Markov chains and Poisson processes.

The price of developing Markov chains in continuous time is that technical complications arise. We are forced to accept some fundamental facts without proof, even more than in previous chapters. Why then study continuous-time Markov chains, especially since any computer simulation has to approximate continuous time with a discrete grid of time points? There are at least two reasons that make this important.

- Many much studied and much applied Markovian models of stochastic phenomena acquire their best and most useful formulation as continuous-time Markov chains. This includes examples such as models of queueing networks, population genetics, and chemical reaction networks, that possess enormous theoretical and practical significance.
- The theory of stochastic processes that follows this introductory course, including topics such as Brownian motion, stochastic differential equations, diffusion processes, general Markov processes, and interacting particle systems, takes place mainly in a continuous-time setting. The study of continuous-time Markov chains is a point of entry into this subject.

6.1. Markov property in continuous time

The key structural property of Markov chains in continuous time is the same as in discrete time: the future of the process, given the past, depends only on the present state. A precise formulation for a continuous time process $\{X_t : t \geq 0\}$ goes as follows: for any $t_0 \geq 0$, the conditional distribution of the process $\{X_t : t \geq t_0\}$

from time t_0 onwards, *given* its past $\{X_t : 0 \le t \le t_0\}$ up to time t_0, depends only on the state X_{t_0} at time t_0.

In contrast with discrete time, transition probabilities in continuous time have to depend explicitly not just on the state variables but also on the time increment. This is a consequence of the simple fact that there is no single time value s whose multiples $0, s, 2s, 3s, \ldots$ account for all possible time values.

Let $\mathcal{S}$ be a discrete state space, that is, $\mathcal{S}$ is either finite or countably infinite.

Definition 6.1. A function $\mathbf{P}_t = \{p_t(x,y)\}_{x,y\in\mathcal{S}}$ of a real time variable $t \ge 0$ and states $x, y \in \mathcal{S}$ is a **transition probability function** if $p_t(x,y) \ge 0$,

$$\sum_{y\in\mathcal{S}} p_t(x,y) = 1 \qquad \text{for all } t \ge 0 \text{ and all } x \in \mathcal{S}$$

and

$$p_0(x,y) = \begin{cases} 1, & x = y \\ 0, & x \ne y. \end{cases}$$

△

Next we state the definition of a time-homogeneous Markov chain in continuous time.

Definition 6.2. Let $\mathbf{P}_t$ be a transition probability function and μ a probability distribution on the discrete state space $\mathcal{S}$. Then an $\mathcal{S}$-valued continuous-time process X_t is a **continuous-time Markov chain with transition probability function $\mathbf{P}_t$ and initial distribution** μ if

$$P(X_0 = x) = \mu(x) \qquad \text{for all states } x \in \mathcal{S}$$

and

$$P(X_{s+t} = y \,|\, X_{s_0} = x_0, \ldots, X_{s_n} = x_n, X_s = x) = p_t(x,y) \tag{6.1}$$

for all $n \in \mathbb{Z}_{\ge 0}$, all states $x_0, \ldots, x_n, x, y \in \mathcal{S}$ and all time points $0 \le s_0 < s_1 < \cdots < s_n < s < s + t$, whenever the conditioning event on the left has positive probability. △

The case $n = 0$ of (6.1) gives the familiar meaning of the transition probability as the one-step conditional probability:

$$p_t(x,y) = P(X_{s+t} = y \,|\, X_s = x). \tag{6.2}$$

As in Chapter 2 we write P_μ for the probability when the initial distribution of the process is μ, and specialize to P_x when the initial state is x: $P_\mu(X_0 = y) = \mu(y)$ and $P_x(X_0 = x) = 1$ for states x, y. The formula for the finite-dimensional distributions of a continuous-time Markov chain in terms of the initial distribution and the transition probability function is entirely analogous to formula (2.23) of discrete-time chains: for $0 = s_0 < s_1 < \cdots < s_n$ and states $x_0, \ldots, x_n$,

$$\begin{aligned} &P_\mu(X_0 = x_0, X_{s_1} = x_1, \ldots, X_{s_n} = x_n) \\ &\qquad = \mu(x_0) \prod_{j=1}^{n} p_{s_j - s_{j-1}}(x_{j-1}, x_j). \end{aligned} \tag{6.3}$$

The proof of (6.3) is analogous to that of (2.23).

Furthermore, the Markov property extends to the infinite future and to arbitrary conditioning events, exactly as it did for discrete-time chains in Theorem 2.21. Let U be any set of $\mathcal{S}$-valued functions of time, and let B be any event determined by the process $\{X_s : 0 \le s \le t\}$ restricted to the time interval $[0,t]$. Then

$$P_\mu\big(\{X_{t+s} : s \ge 0\} \in U \,\big|\, X_t = x, B\big) = P_x\big(\{X_s : s \ge 0\} \in U\,\big) \tag{6.4}$$

provided the conditioning event has positive probability.

The strong Markov property also holds. In continuous time, T is a stopping time if for each finite t, the event $\{T \le t\}$ is determined by the process $\{X_s : 0 \le s \le t\}$ up to time t. Then for any event B determined by the process $\{X_s : 0 \le s \le T\}$ up to the stopping time T, we have the identity

$$P_\mu\big(\{X_{T+s} : s \ge 0\} \in U \,\big|\, T < \infty, X_T = x, B\big] = P_x\big(\{X_s : s \ge 0\} \in U\,\big) \tag{6.5}$$

provided the conditioning event has positive probability.

Before further general discussion we go over some examples that turn a discrete-time Markov chain into a continuous one by attaching a Poisson clock to it.

Example 6.3 (Homogeneous Poisson process)**.** Let $\{N_t : t \ge 0\}$ be a rate α Poisson process. Take the state space to be $\mathcal{S} = \mathbb{Z}_{\ge 0}$. We show that N_t satisfies Definition 6.2 and identify the transition probability.

Let $0 \le k_0 \le k_1 \le \dots \le k_n \le k \le \ell$ be integers and $0 \le s_0 < s_1 < \dots < s_n < s < s+t$ real time points. Use below the independent increments and the Poisson distribution in the last step. To simplify the notation of intermediate steps, write also $s_{n+1} = s$, $k_{n+1} = k$, $s_{n+2} = s+t$, and $k_{n+2} = \ell$.

$$\begin{aligned}
&P(N_{s+t} = \ell \,|\, N_{s_0} = k_0, \dots, N_{s_n} = k_n, N_s = k)\\
&\quad= \frac{P(N_{s_0} = k_0, \dots, N_{s_n} = k_n, N_s = k, N_{t+s} = \ell)}{P(N_{s_0} = k_0, \dots, N_{s_n} = k_n, N_s = k)}\\
&\quad= \frac{P(N(0,s_0] = k_0, N(s_0,s_1] = k_1 - k_0, \dots, N(s_{n+1},s_{n+2}] = k_{n+2} - k_{n+1})}{P(N(0,s_0] = k_0, N(s_0,s_1] = k_1 - k_0, \dots, N(s_n,s_{n+1}] = k_{n+1} - k_n)}\\
&\quad= \frac{P(N(0,s_0] = k_0)\prod_{i=1}^{n+2} P(N(s_{i-1},s_i] = k_i - k_{i-1})}{P(N(0,s_0] = k_0)\prod_{i=1}^{n+1} P(N(s_{i-1},s_i] = k_i - k_{i-1})}\\
&\quad= P(N(s_{n+1},s_{n+2}] = k_{n+2} - k_{n+1})\\
&\quad= P(N(s,s+t] = \ell - k) = e^{-\alpha t}\frac{(\alpha t)^{\ell-k}}{(\ell-k)!}.
\end{aligned}$$

Thus a homogeneous Poisson process is a continuous-time Markov chain. Its transition probability is

$$p_t(k,\ell) = e^{-\alpha t}\frac{(\alpha t)^{\ell-k}}{(\ell-k)!} \qquad \text{for } k \le \ell,$$

and $p_t(k,\ell) = 0$ for $k > \ell$ (because the Poisson process never jumps down).

Up to now we have always assumed the initial value to be $N_0 = 0$. But nothing prevents us from allowing other initial conditions. Let X_0 be a $\mathbb{Z}_{\ge 0}$-valued random variable, independent of $\{N_t : t \ge 0\}$. Then a rate α Poisson process with initial value X_0 can be defined by $X_t = X_0 + N_t$. The calculation above works again to give the Markov property since X_0 is independent of all the future N-increments.

A Poisson process with a nonzero initial condition can have natural applications. For example, if X_t represents the number of customer arrivals to a service station by time t, then X_0 can be the number of customers who are already waiting when the station opens. △

The second example generalizes Example 6.3 to a large class of examples.

Example 6.4 (Discrete-time Markov chain run by a Poisson process)**.** Let $\{Y_n : n \geq 0\}$ be a discrete-time Markov chain on a discrete state space $\mathcal{S}$ with transition probability matrix $U = \{u(x, y)\}_{x,y \in \mathcal{S}}$. Let $\{N_t : t \geq 0\}$ be a rate α Poisson process that is independent of $\{Y_n : n \geq 0\}$. Define a continuous-time process by setting $X_t = Y_{N_t}$ for $t \geq 0$. In other words, if $0 = S_0 < S_1 < S_2 < \cdots$ are the jump times of N_t, then for all $t \geq 0$ and $n \geq 0$,

$$X_t = Y_n \quad \text{ for } \; t \in [S_n, S_{n+1}).$$

The two processes have the same initial state $X_0 = Y_0$ and X_t visits the same sequence of states in the same order as Y_n does. But while Y_n moves exactly at integer times, X_t waits $\text{Exp}(\alpha)$-distributed random times between its jumps.

We show that X_t is a continuous-time Markov chain and derive its transition probability function $p_t(x, y)$.

We compute the joint distribution of $X_0, X_{s_1}, \dots, X_{s_n}$ for arbitrary time points $0 = s_0 < s_1 < s_2 < \cdots < s_n$. Let $x_0, \dots, x_n$ be states in $\mathcal{S}$. The first step below decomposes the event $\{X_0 = x_0, X_{s_1} = x_1, \dots, X_{s_n} = x_n\}$ according to the number of jumps that the Poisson process N_t takes in the intervals $(0, s_1], (s_1, s_2], \dots, (s_{n-1}, s_n]$. Once the numbers of jumps have been chosen, we know which Y_n-state gives each X_{s_j}.

$$\begin{aligned}
&P(X_0 = x_0, X_{s_1} = x_1, X_{s_2} = x_2, \dots, X_{s_n} = x_n) \\
&= \sum_{k_1,\dots,k_n \geq 0} P\bigl(N(0, s_1] = k_1, N(s_1, s_2] = k_2, \dots, N(s_{n-1}, s_n] = k_n, \\
&\qquad\qquad\qquad X_0 = x_0, X_{s_1} = x_1, X_{s_2} = x_2, \dots, X_{s_n} = x_n\bigr) \\
&= \sum_{k_1,\dots,k_n \geq 0} P\bigl(N(0, s_1] = k_1, N(s_1, s_2] = k_2, \dots, N(s_{n-1}, s_n] = k_n, \\
&\qquad\qquad\qquad Y_0 = x_0, Y_{k_1} = x_1, Y_{k_1+k_2} = x_2, \dots, Y_{k_1+\cdots+k_n} = x_n\bigr) \\
&= \sum_{k_1,\dots,k_n \geq 0} \biggl\{ \prod_{j=1}^{n} e^{-\alpha(s_j - s_{j-1})} \frac{(\alpha(s_j - s_{j-1}))^{k_j}}{k_j!} \biggr\} \cdot \biggl\{ P(Y_0 = x_0) \prod_{j=1}^{n} u^{(k_j)}(x_{j-1}, x_j) \biggr\}.
\end{aligned}$$

The last step above used first the independence of $\{Y_n : n \geq 0\}$ and $\{N_t : t \geq 0\}$ and then the joint distributions of these two processes.

Take the initial probability $P(Y_0 = x_0) = P(X_0 = x_0)$ outside the sum and rearrange the sum of products as a product of sums:

$$= P(Y_0 = x_0) \sum_{k_1,\dots,k_n \ge 0} \prod_{j=1}^{n} \left\{ e^{-\alpha(s_j - s_{j-1})} \frac{(\alpha(s_j - s_{j-1}))^{k_j}}{k_j!} \, u^{(k_j)}(x_{j-1}, x_j) \right\}$$

$$= P(X_0 = x_0) \prod_{j=1}^{n} \left\{ \sum_{k_j \ge 0} e^{-\alpha(s_j - s_{j-1})} \frac{(\alpha(s_j - s_{j-1}))^{k_j}}{k_j!} \, u^{(k_j)}(x_{j-1}, x_j) \right\}. \tag{6.6}$$

The product form above suggests a Markovian structure. Rename $s_{n-1} = s$, $x_{n-1} = x$, $s_n = s + t$, and $x_n = y$ and substitute the formula above into the conditional probability in equation (6.2). The conditional probability becomes a ratio of two products of the kind on line (6.6). All but the last factor in the numerator cancels and the calculation yields the following:

$$\begin{aligned} &P(X_{s+t} = y \,|\, X_{s_0} = x_0, \dots, X_{s_{n-2}} = x_{n-2}, X_s = x) \\ &= \frac{P(X_{s_0} = x_0, \dots, X_{s_{n-2}} = x_{n-2}, X_s = x, X_{s+t} = y)}{P(X_{s_0} = x_0, \dots, X_{s_{n-2}} = x_{n-2}, X_s = x)} \\ &= \sum_{k_n \ge 0} e^{-\alpha(s_n - s_{n-1})} \frac{(\alpha(s_n - s_{n-1}))^{k_n}}{k_n!} \, u^{(k_n)}(x_{n-1}, x_n) \\ &= \sum_{k \ge 0} e^{-\alpha t} \frac{(\alpha t)^k}{k!} u^{(k)}(x, y). \end{aligned}$$

The last step above just cleaned up the notation. We have verified (6.1) and thereby identified the transition probability function

$$p_t(x, y) = \sum_{k=0}^{\infty} e^{-\alpha t} \frac{(\alpha t)^k}{k!} \, u^{(k)}(x, y). \tag{6.7}$$

We obtain Example 6.3 as a special case by taking $u(k, k+1) = 1$ for nonnegative integers k. Then in particular $Y_0 = 0$ implies $Y_n = n$ for all $n \ge 0$, and the CTMC is $X_t = Y_{N_t} = N_t$, namely the Poisson process itself. △

Example 6.5 (Continuous-time symmetric simple random walk)**.** As another particular case of Example 6.4 we can define a rate α continuous-time SSRW on $\mathbb{Z}$ by taking Y_n to be discrete-time SSRW with transition matrix $u(k, k \pm 1) = \frac{1}{2}$ for $k \in \mathbb{Z}$.

To illustrate the formula (6.7), we calculate the probability $p_t(0, 0)$ that after t time units the walk is at the origin, having started there at time zero.

$$\begin{aligned} p_t(0,0) &= \sum_{k=0}^{\infty} e^{-\alpha t} \frac{(\alpha t)^k}{k!} \, u^{(k)}(0,0) = \sum_{n=0}^{\infty} e^{-\alpha t} \frac{(\alpha t)^{2n}}{(2n)!} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} \\ &= \sum_{n=0}^{\infty} \frac{e^{-\alpha t}}{(n!)^2} \left(\frac{\alpha t}{2}\right)^{2n}. \end{aligned} \tag{6.8}$$

The series above includes the possibility that the process never moved from the origin during the time interval $[0, t]$ (this is the term $e^{-\alpha t}$ for $n = 0$) and for each positive n the possibility of making exactly $2n$ jumps during $[0, t]$ that bring the process back to the origin. For odd k, $u^{(k)}(0,0) = 0$ because the discrete-time walk cannot return to its starting point in an odd number of steps. △

Definition 6.2 and identity (6.3) suggest that transition probabilities should again be central in the study of continuous-time Markov chains, as they were in the discrete-time theory. However, this turns out to be impractical in most situations because the transition probabilities are complicated. Even when the transition probabilities are accessible, they do not necessarily give us a useful representation of the evolution.

The complexity and limitations of transition probabilities are already illustrated by (6.7). Suppose we were interested in something seemingly simple such as the probability distribution of the first jump time and the location of the process after that jump. There does not seem to be a way to get this information from the formula for $p_t(x, y)$. (See Example 6.8 below.)

Thus the starting point of the study must be something else, in terms of which the transition probabilities $\mathbf{P}_t$ are then later defined. Two possibilities appear in the literature:

(i) a hands-on construction of the process X_t, or

(ii) an analytic definition of the transition probabilities in terms of given jump rates.

We opt for the former. In Section 6.2 we construct the random evolution X_t from basic ingredients and then define the *jump rates*. Section 6.3 defines the *generator matrix* in terms of the jump rates and then derives the Kolmogorov differential equations satisfied by the transition probabilities. Along the way we see that the generator matrix is the most convenient representation of the rules of evolution of a continuous-time Markov chain .

6.2. Construction of continuous-time Markov chains

We give three constructions. The first construction is the main one. The second one turns the previous Example 6.4 into a construction, but this one cannot cover all the important examples. The third one gives a valuable alternative perspective on the evolution, in terms of the exponential races of Lemma 4.6. Example 6.9 compares the three constructions of the continuous-time Markov chain on two states. After that we present a series of examples. Remark 6.13 clarifies how simultaneous occurrences of events never happen in continuous-time Markov chains, unless of course built in as features of the model. The last part of the section discusses the possibility of explosion, which is the scenario where the process essentially leaves the state space in finite time.

First construction: jump chain and holding times. On the most basic level, a description of a temporal evolution $\{X_t : 0 \le t < \infty\}$ in the state space $\mathcal{S}$ requires two ingredients:

- The sequence of states visited, denoted by $Z_0, Z_1, Z_2, \dots$ where $Z_n \neq Z_{n+1}$ so that only jumps from a state to a different one are recorded.
- The times when jumps are made, denoted by $0 = J_0 < J_1 < J_2 < \cdots$.

From these ingredients the full evolution is defined, for all $t \geq 0$ and $n \in \mathbb{Z}_{\geq 0}$, by

$$X_t = Z_n \quad \text{for} \quad t \in [J_n, J_{n+1}). \tag{6.9}$$

This formulation includes two tacit assumptions.

We assumed $J_n < J_{n+1}$ for each n, which means that whenever X_t moves to a state, it stays there for some nonzero amount of time before moving again. This is eminently reasonable: if $J_{n+1} = J_n$ then X_t spends no time in state Z_n and this step might as well be eliminated from the description.

Equation (6.9) makes the path $t \mapsto X_t$ *right-continuous*, or equivalently, at the moment of a jump the process already resides at the target state. This convention is followed by all modern literature on stochastic processes. Our earlier definitions of Poisson and renewal processes made them right-continuous.

Next we investigate what properties the ingredients $\{Z_n\}$ and $\{J_n\}$ should have. For X_t to be Markovian requires that, conditional on $X_s = x$ and everything that happened in the process up to time s, the process starts anew from the state x.

- The next state can depend only on the present state x and not on the previous states. In other words, the sequence $\{Z_n\}$ of states visited is a Markov chain. We denote its transition probability matrix by $R = \{r(x, y)\}_{x,y \in \mathcal{S}}$.
- The time till the next jump can depend on the present state x but it must have the memoryless property because it cannot depend on the amount of time already spent at x. Hence the random time till the next jump must be an exponential random variable whose parameter $\lambda(x)$ can depend on x.

Thus far the description has ignored the possibility of an absorbing state. We take care of it by stipulating that, if Z_n is the first visit to an absorbing state, then $J_{n+1} = \infty$ and the remaining states $\{Z_k\}_{k \geq n+1}$ are ignored. The convention is that *an exponential random variable with parameter zero is infinite with probability one.* This is consistent with $T \sim \text{Exp}(0)$ satisfying

$$P(T > t) = e^{-0 \cdot t} = e^0 = 1 \qquad \text{for all } t \in [0, \infty).$$

$\{Z_n\}$ is called the *jump chain* or the *embedded Markov chain*. Its transition matrix R is called the *jump matrix* or the *routing matrix*. The parameter $\lambda(x)$ is the *holding parameter* at x or the *rate to jump away from* x. The next two bullets identify the properties of these parameters.

- If x is an absorbing state for the process, then $r(x, x) = 1$ (no jump out of x) and $\lambda(x) = 0$ (zero rate to leave x).
- For all other states x, $r(x, x) = 0$ because the jump chain records only jumps between distinct states, and $0 < \lambda(x) < \infty$ because the process spends some positive time at x (implied by $\lambda(x) < \infty$) but eventually leaves x (implied by $\lambda(x) > 0$).

The data required for constructing a continuous-time Markov chain is summarized as follows.

- an initial distribution μ;
- a routing matrix $R = \{r(x, y)\}_{x,y \in \mathcal{S}}$;
- holding time parameters $\lambda(x) \in [0, \infty)$ for $x \in \mathcal{S}$.

To construct the random evolution X_t, generate a discrete-time Markov chain $\{Z_n : n \geq 0\}$ with initial distribution μ and transition probability matrix R, and then, an independent sequence of i.i.d. random variables $\{\tau_k : k \geq 0\}$ that all have the same Exp(1) distribution. Let $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}_\mu)$ denote the probability space of these random variables. If μ is concentrated on a single state x we write $\widehat{P}_x$ instead of $\widehat{P}_\mu$.

Set $J_0 = 0$ and for $n \geq 1$ define the jump times by

$$J_n = \sum_{k=0}^{n-1} \frac{\tau_k}{\lambda(Z_k)}. \tag{6.10}$$

The convention is that if $\lambda(Z_k) = 0$ then $J_n = \infty$ for $n \geq k+1$. Then, as already explained above, define the process X_t by

$$X_t = Z_n \qquad \text{for } J_n \leq t < J_{n+1}. \tag{6.11}$$

Thus X_t remains in state Z_n for the *holding time* $J_{n+1} - J_n = \tau_n / \lambda(Z_n)$.

To understand the construction, let us derive the conditional distribution of the holding times $J_{k+1} - J_k = \tau_k/\lambda(Z_k)$, given the jump chain $\{Z_n\}$. Let $x_0, \dots, x_n \in \mathcal{S}$ and $s_0, \dots, s_n > 0$.

$$\begin{aligned}
&\widehat{P}_\mu\big(Z_0 = x_0, \dots, Z_n = x_n,\ J_1 - J_0 > s_1, \dots, J_{n+1} - J_n > s_n\big)\\
&= \widehat{P}_\mu\Big(Z_0 = x_0, \dots, Z_n = x_n,\ \frac{\tau_0}{\lambda(Z_0)} > s_1, \dots, \frac{\tau_n}{\lambda(Z_n)} > s_n\Big)\\
&= \widehat{P}_\mu\big(Z_0 = x_0, \dots, Z_n = x_n, \tau_0 > \lambda(x_0)s_1, \dots, \tau_n > \lambda(x_n)s_n\big)\\
&= \mu(x_0)\prod_{i=1}^n r(x_{i-1}, r_i) \cdot \prod_{i=0}^n e^{-\lambda(x_i)s_i}\\
&= \widehat{P}_\mu\big(Z_0 = x_0, \dots, Z_n = x_n\big) \cdot \prod_{i=0}^n e^{-\lambda(x_i)s_i}.
\end{aligned}$$

The second last equality above used the independence of $\{Z_k\}$ and $\{\tau_k\}$ and the joint distributions of these two processes. Dividing by the probability of the jump chain gives the conditional probability:

$$\begin{aligned}
&\widehat{P}_\mu\big(\, J_1 - J_0 > s_1, \dots, J_{n+1} - J_n > s_n \,\big|\, Z_0 = x_0, \dots, Z_n = x_n\big)\\
&\qquad = \prod_{i=0}^n e^{-\lambda(x_i)s_i}.
\end{aligned} \tag{6.12}$$

Thus, given the evolution of the jump chain $\{Z_k\}$, the holding times $\{J_{k+1} - J_k\}_{k \geq 0}$ are conditionally independent with distributions $J_{k+1} - J_k \sim \text{Exp}(\lambda(Z_k))$.

The calculation above tells us that the process X_t operates according to this description: (i) X_t visits the states $Z_0, Z_1, Z_2, \dots$ as prescribed by the jump chain, and (ii) given the sequence of states visited, the holding time in state Z_n is exponentially distributed with parameter $\lambda(Z_n)$ and independent of the other holding times.

From this construction we define the transition probabilities $\mathbf{P}_t = \{p_t(x,y)\}_{x,y\in\mathcal{S}}$ as

$$p_t(x,y) = \widehat{P}_x(X_t = y) \qquad \text{for } t \geq 0 \text{ and } x,y \in \mathcal{S}. \tag{6.13}$$

Theorem 6.6. *The process X_t thus constructed satisfies Definition 6.2 of a Markov chain with initial distribution μ and transition probability function $\mathbf{P}_t$.*

To summarize, given the routing matrix R and the holding parameters $\{\lambda(x) : x \in \mathcal{S}\}$, we have now constructed a continuous-time Markov chain for each initial distribution μ. For probabilities that concern the process X_t we write simply P_μ and P_x and drop the complicated notation $\widehat{P}_\mu$ and $\widehat{P}_x$ associated with the previous construction.

Define the (*infinitesimal*) *jump rates* of the process by

$$q(x,y) = \lambda(x)r(x,y) \qquad \text{for states } \ x \neq y. \tag{6.14}$$

The collection of jump rates $\{q(x,y) : x \neq y\}$ gives an equivalent way of summarizing the data of the process because the routing matrix and the holding parameters can be derived from the jump rates. To see this, the first step is that

$$\lambda(x) = \sum_{y:y\neq x} q(x,y) \tag{6.15}$$

for all states x: if x is absorbing, then $\lambda(x) = 0$ and the right-hand side of (6.14) equals zero for all $y \neq x$; while if x is non-absorbing, then $\sum_{y:y\neq x} r(x,y) = 1$ by the properties of the routing matrix. In the second step derive the routing matrix: if $\sum_{y:y\neq x} q(x,y) = 0$ then x is absorbing and set $r(x,x) = 1$ and $r(x,y) = 0$ for $y \neq x$; in the complementary case $\sum_{y:y\neq x} q(x,y) > 0$ set $r(x,x) = 0$ and

$$r(x,y) = \frac{q(x,y)}{\sum_{y:y\neq x} q(x,y)} \qquad \text{for } \ x \neq y. \tag{6.16}$$

It is helpful to keep both equivalent parametrizations in mind: on the one hand the jump rates $\{q(x,y) : x \neq y\}$, on the other hand the routing matrix $R = \{r(x,y) : x,y \in \mathcal{S}\}$ and the holding parameters $\{\lambda(x) : x \in \mathcal{S}\}$. As we gain familiarity with these quantities through examples and further theory, their meaning will become clearer. We illustrate them with the examples from Section 6.1.

Example 6.7 (Continuation of Example 6.3)**.** The process is $X_t = X_0 + N_t$, a rate α Poisson process N_t with an independent initial condition X_0, on the state space of nonnegative integers. Since only jumps from k to $k+1$ are allowed and the waiting times between jumps are i.i.d. Exp(α), for all $k \geq 0$ we have $r(k,k+1) = 1$ and $q(k,k+1) = \lambda(k) = \alpha$ and $r(k,\ell) = q(k,\ell) = 0$ for $\ell \neq k+1$. △

Example 6.8 (Continuation of Example 6.4)**.** The process is $X_t = Y_{N_t}$, a discrete-time Markov chain Y_n run by an independent rate α Poisson process N_t. The transition matrix of Y_n is $U = \{u(x,y)\}_{x,y\in\mathcal{S}}$.

The routing matrix R is not necessarily the same as the transition matrix U because the latter may satisfy $0 < u(x,x) < 1$ for some state x. This corresponds to the possibility that $Y_{n+1} = Y_n$ even though Y_n is not in an absorbing state. By contrast, the jump chain $Z_0, Z_1, Z_2, \ldots$ must satisfy $Z_{n+1} \neq Z_n$ as long as Z_n is not in an absorbing state.

The correct routing matrix is obtained by

$$r(x,y) = \begin{cases} \frac{u(x,y)}{1-u(x,x)}, & y \neq x \text{ and } u(x,x) < 1 \\ 0, & y = x \text{ and } u(x,x) < 1 \\ 0, & y \neq x \text{ and } u(x,x) = 1 \\ 1, & y = x \text{ and } u(x,x) = 1. \end{cases} \tag{6.17}$$

The first two lines are the case where x is not an absorbing state ($u(x,x) < 1$), and come from

$$r(x,y) = P_x(Y_1 = y \,|\, Y_1 \neq x) = \frac{P_x(Y_1 = y, Y_1 \neq x)}{P_x(Y_1 \neq x)}. \tag{6.18}$$

The last two lines of (6.17) are the case of an absorbing state x. (This problem was solved without absorbing states in Exercise 2.13 of Chapter 2.)

The next item is the holding parameter of a nonabsorbing state x. We can deduce this without explicit calculation by thinning the rate α Poisson process N_t that runs the process. Namely, when the Markov chain Y_n starts at x, it makes repeated attempts to jump away from x, each time with probability $1 - u(x,x)$, until it succeeds. The time J_1 of the first jump of X_t is the first jump time of N_t at which Y_n moves away from x. Thus J_1 is the first jump time of the process obtained by thinning N_t with probability $1 - u(x,x)$. The thinned process is a Poisson process with rate $\alpha(1 - u(x,x))$. Consequently $J_1 \sim \text{Exp}\big(\alpha(1 - u(x,x))\big)$.

We conclude that the holding parameters of nonabsorbing states are given by

$$\lambda(x) = \alpha(1 - u(x,x)). \tag{6.19}$$

This same equation works also for absorbing states x because then $u(x,x) = 1$ and we get $\lambda(x) = 0$, as required by the definition of the holding parameters.

As the last item, the jump rates come from (6.14): for $x \neq y$,

$$q(x,y) = \lambda(x)r(x,y) = \alpha u(x,y). \tag{6.20}$$

△

Second construction: discrete-time Markov chain run by a Poisson process. We take what we learned in Example 6.8 and turn it into an alternative construction of a continuous-time Markov chain. The only restriction to the applicability of this method comes from the identity $\lambda(x) = \alpha(1 - u(x,x))$. This cannot hold unless $\lambda(x) \leq \alpha$ for all states x. We make this an assumption.

Given jump rates $\{q(x,y) : x \neq y\}$, or equivalently the routing matrix R and the holding parameters $\{\lambda(x) : x \in \mathcal{S}\}$, assume that

$$\text{there exists a constant } \alpha \text{ such that } \lambda(x) \leq \alpha \text{ for all states } x. \tag{6.21}$$

Under this assumption, a possible construction of the continuous-time Markov chain X_t with the given parameters goes as follows. Let N_t be a rate α Poisson process. Let Y_n be an independent discrete-time Markov chain with transition matrix U defined by

$$u(x,y) = \begin{cases} \frac{q(x,y)}{\alpha}, & y \neq x \\ 1 - \frac{\lambda(x)}{\alpha}, & y = x. \end{cases} \tag{6.22}$$

Observe that if x is absorbing for the continuous-time Markov chain, then $\lambda(x) = q(x,y) = 0$ for all $y \neq x$, and the definition above gives $u(x,x) = 1$ and $u(x,y) = 0$ for all $y \neq x$. Thus x is also absorbing for U. If x is not absorbing, $\sum_{y:y\neq x} q(x,y) = \lambda(x)\sum_{y:y\neq x} r(x,y) = \lambda(x)$, and we see that $\sum_y u(x,y) = 1$, as should be the case for a transition matrix.

Now define $X_t = Y_{N_t}$. The calculations in Example 6.8 show that X_t has exactly the desired parameters $q(x,y)$, $\lambda(x)$ and $r(x,y)$. Equation (6.7) gives the transition probabilities of X_t.

To explain in plain English why this works, note that the transition matrix (6.22) forces the discrete-time chain Y_n to execute false jumps from a state back to itself. Thus, even though the clock N_t runs at rate α which may be too fast for the Markov chain X_t, the false jumps of Y_n slow down the motion by just the right amount so that X_t jumps with the given rates $q(x,y)$.

Third construction: Poisson clocks. This third construction gives the jump rates $q(x,y)$ a concrete meaning. Imagine an arrow diagram (technically speaking, a *directed graph*) of the Markov chain: the states are the vertices, and a directed arrow is drawn from state x to state $y \neq x$ whenever $q(x,y) > 0$. If there is an arrow from x to y, associate a rate $q(x,y)$ Poisson process $\{N_t^{x,y} : 0 \leq t < \infty\}$ to this arrow. These Poisson processes are independent of each other.

It is instructive to view the process X_t as a "particle" that moves on the arrow diagram by hopping from state to state along the arrows, and the Poisson processes as random "clocks" that control the particle's motion. The Poisson clock $N^{x,y}$ "rings" whenever it experiences a jump.

The rule of evolution of X_t is simply as follows. Whenever some clock $N^{x,y}$ rings, check whether the particle is at state x. If it is, move it instantaneously to y. If it is not, do nothing.

From the perspective of the particle the motion goes as follows. Suppose the particle moved to some state x_1 at time t_1. Then the particle remains at x_1 until some Poisson clock $\{N_{t_1+s}^{x_1,y} : s \geq 0\}$ on an arrow out of x_1 rings after time t_1. Suppose the first ring happens at Poisson process N^{x_1,x_2} at time $t_2 > t_1$. Then at time t_2 the particle moves instantaneously from x_1 to x_2.

This step is then repeated. The process remains at x_2 until some clock among the Poisson processes $N^{x_2,y}$ rings after time t_2. At the moment of the ring the process moves along the arrow whose Poisson clock was the first to ring. And so on.

We can again let $Z_0, Z_1, Z_2, \dots$ be the sequence of states visited and $0 = J_0 < J_1 < J_2 < \cdots$ be the jump times so that J_n is the moment when the process X_t moves from state Z_{n-1} to state Z_n. To justify that this construction yields the same process X_t as the previous constructions, we argue that this process obeys the routing matrix R and the holding times $\lambda(x)$.

If the process ever arrives at an absorbing state x, it remains there forever because $q(x,y) = 0$ for all $y \neq x$ and hence there are no outgoing arrows from x.

The fairly advanced key technical point is that the time J_n when the process arrives in state Z_n is a *stopping time* simultaneously for the entire collection of Poisson processes $N^{x,y}$. Precisely, this means that the event $\{J_n \le t\}$ depends only on what happened in the Poisson processes up to time t. The collection of Poisson processes satisfies the strong Markov property, so from time J_n onwards they are all again independent Poisson processes. This allows us to reason as if we were starting everything from time zero.

In particular, if the new state is $Z_n = x$ that is nonabsorbing, then the time till the next jump is the minimum of the waiting times of the Poisson processes $N^{x,y}$ on arrows out of x. These waiting times are independent exponential random variables with rates $q(x,y)$. By Lemma 4.6 their minimum is an exponential random variable with rate

$$\sum_{y:\, y\neq x} q(x,y) = \lambda(x) \sum_{y:\, y\neq x} r(x,y) = \lambda(x). \tag{6.23}$$

This shows that the holding time in a nonabsorbing state x has $\text{Exp}(\lambda(x))$ distribution.

The target of the jump from x is the state y whose Poisson clock $N^{x,y}$ rang first. Again by Lemma 4.6 the probability that this is a particular state y' equals

$$\frac{q(x,y')}{\sum_{y:\, y\neq x} q(x,y)} = \frac{\lambda(x) r(x,y')}{\lambda(x)} = r(x,y'). \tag{6.24}$$

Thus the choice of where to jump is governed by the routing matrix. This completes the justification that this construction gives the same process (in the sense of probability distributions) as the first construction.

The next example juxtaposes the three constructions of the two-state continuous-time Markov chain.

Example 6.9 (Two-state Markov chain)**.** Consider the Markov chain with state space $\mathcal{S} = \{1,2\}$ and jump rates $q(1,2) = \lambda$ and $q(2,1) = \mu$.

(a) *Holding times and jump chain.* The routing matrix is

$$R = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and the holding parameters are $\lambda(1) = \lambda$ and $\lambda(2) = \mu$. The jump chain has only two possible evolutions:

$$\text{if the initial state is } \begin{cases} 1, & \text{then } (Z_0, Z_1, Z_2, Z_3, \dots) = (1,2,1,2,\dots) \\ 2, & \text{then } (Z_0, Z_1, Z_2, Z_3, \dots) = (2,1,2,1,\dots). \end{cases}$$

In terms of the i.i.d. Exp(1) variables $\{\tau_k\}_{k\ge 0}$ the jump times J_n of (6.10) satisfy equations such as the following: if the initial state is $X_0 = Z_0 = 1$, then $J_1 = \tau_0/\lambda$,

$J_2 - J_1 = \tau_1/\mu$, and more generally, for example if n is even, then

$$J_n = \sum_{k=0}^{n-1} \frac{\tau_k}{\lambda(Z_k)} = \frac{\tau_0 + \tau_2 + \cdots + \tau_{n-2}}{\lambda} + \frac{\tau_1 + \tau_3 + \cdots + \tau_{n-1}}{\mu}.$$

As stipulated by the construction, J_n is the time when the process X_t moves from state Z_{n-1} to Z_n.

(b) *Discrete-time Markov chain run by a Poisson process.* Fix a number $\alpha \geq \lambda \vee \mu$ and let N_t be a rate α Poisson process. As stipulated in (6.22), define the transition matrix U as

$$U = \begin{bmatrix} 1 - \frac{\lambda}{\alpha} & \frac{\lambda}{\alpha} \\ \frac{\mu}{\alpha} & 1 - \frac{\mu}{\alpha} \end{bmatrix}$$

and let Y_n be a discrete-time Markov chain with transition matrix U and initial distribution as desired. Then $X_t = Y_{N_t}$ gives a construction of the Markov chain that has the same probability distribution as the construction above.

(c) *Poisson clocks.* To construct the process X_t with Poisson clocks, we need three independent ingredients: the initial state X_0, a rate λ Poisson process $N_t^{1,2}$, and a rate μ Poisson process $N_t^{2,1}$.

At time zero place the particle in the initial position X_0. The evolution is defined by repeatedly obeying these rules: (i) Whenever the particle is in state 1, it remains in state 1 until the next ring of $N^{1,2}$ at which time the particle moves instantaneously to state 2. (ii) Whenever the particle is in state 2, it remains in state 2 until the next ring of $N^{2,1}$ at which time the particle moves instantaneously to state 1. △

Further examples. We present here several well-known examples. Each example is described by giving its jump rates. This is the most useful way of describing continuous-time Markov chains. *Rates must not be confused with probabilities.* In Section 6.3 we take a closer look at how rates and transition probabilities connect with each other.

Example 6.10 (M/M/1 queue)**.** Imagine a service station with a single server. These are the rules. Only one customer is served at a time. Customers who arrive when the server is busy get in line to wait. Customers depart after their service is done. Customers obey FIFO (first in first out) queueing discipline: customers are served in the order of their arrival.

For a concrete example, think of a gas station with a single pump. Only one car can pump gas at a time. If I drive up and the pump is in use, I get in line to wait for my turn. There is no cutting in line (FIFO discipline). When my tank is full, I drive away.

Assume that customers arrive according to a rate λ Poisson process. The server spends an $\text{Exp}(\mu)$ distributed random time with each customer, independently of everything else. The parameter λ is the *arrival rate* and the parameter μ the *service rate.*

The continuous-time Markov chain is defined by letting X_t denote the number of customers in the system at time t (that is, both in service and waiting in the queue). The state space is $\mathcal{S} = \mathbb{Z}_{\geq 0}$, the set of nonnegative integers. The jump rates are $q(k, k+1) = \lambda$ for $k \geq 0$, corresponding to the arrival of a new customer, and $q(k, k-1) = \mu$ for $k \geq 1$, corresponding to the departure of a customer whose service is complete. △

Remark 6.11 (Queueing nomenclature)**.** The name M/M/1 queue of the example above is explained as follows. The first M is for memoryless arrivals, the second M for memoryless services, and 1 for a single server. This acronym extends in various ways. If there are s servers that serve customers simultaneously, then the model is an M/M/s queue (Example 6.12 below).

The model of Example 6.10 has unbounded waiting room for customers. An M/M/1/ℓ queue is as in Example 6.10 but with the restriction that there can be at most ℓ customers in the system. The state space is reduced to $\mathcal{S} = \{0, 1, \dots, \ell\}$. The rates are as in Example 6.10 except that $q(k, k+1)$ is defined only for $0 \leq k \leq \ell - 1$ and $q(k, k-1)$ only for $1 \leq k \leq \ell$. The assumption behind this model is that customers who arrive when the system is full turn away and never enter the system.

If the arrivals come from a Poisson process but the service time of each customer is some other ("general") distribution, the model is an M/G/1 queue. A crucial point is that only M/M/ type queues are Markov chains because without the memoryless property of the exponential distribution we cannot have the Markov property. A consequence is that there is a well-developed rich theory of M/M/ queues, but much less for non-Markovian queues. △

Example 6.12 (M/M/s queue)**.** Suppose now that there are s servers taking care of customers in parallel. (There are s pumps at the gas station which can be in use simultaneously.) Customers still arrive according to a rate λ Poisson process. If one of the servers is free, an arriving customer goes immediately into service. If all servers are busy an arriving customer waits in line until there is a free server. For simplicity, assume that each server spends an $\text{Exp}(\mu)$ distributed random time with every customer, independently of everything else. That is, each server serves at the same rate μ.

Again X_t denotes the number of customers in the system at time t. The arrival rates are still $q(k, k+1) = \lambda$ for $k \geq 0$. The departure rates now depend on the number of customers in the system:

$$q(k, k-1) = \begin{cases} k\mu & 1 \leq k \leq s \\ s\mu & k \geq s+1. \end{cases} \tag{6.25}$$

The explanation for the rates above comes from Lemma 4.6. When there are $k \leq s$ customers in the system they are all in service. The time till the next departure is the *minimum* of the k independent service times of these customers. The minimum of k independent $\text{Exp}(\mu)$ random variables is an $\text{Exp}(k\mu)$ variable. When all servers are full ($k \geq s$), customers depart at the maximal rate $s\mu$. △

Remark 6.13 (Simultaneous events)**.** The previous examples raise the following questions. Can there not be a jump from k to $k - 2$ when two customers finish service at the same time? What if there is an arrival and a departure at the exact same moment? Then the state X_t does not change, so does the model not recognize that something happened in the system?

The resolution of these questions highlights a powerful advantage of continuous time: *two distinct events never happen simultaneously.* This harks back to the fact that if two random variables (X, Y) have a joint density function $f(x, y)$, then $X \neq Y$ happens with probability one. To see this from the basic rules governing random variables with density functions, let $D = \{(x, x) : x \in \mathbb{R}\}$ denote the diagonal of the plane $\mathbb{R}^2$. Then we can compute

$$\begin{aligned} P(X = Y) &= P\big((X, Y) \in D\big) = \iint_D f(x, y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} \bigg(\int_x^x f(x, y)\,dy \bigg)\,dx = \int_{-\infty}^{\infty} 0\,dx = 0. \end{aligned}$$

Geometrically speaking, the reason behind this is that D has zero area.

This result applies to the models above because the waiting times for different events are independent exponential random variables. Thus any two of them have a joint density function. Consequently no two customers ever finish service at the same moment, or arrive and depart at the same moment.

If we desire more than one event to happen together, we have to build this possibility into the model. The next example illustrates. △

Example 6.14 (M/M/1 queue with batch arrivals)**.** If we want an M/M/1 queue that allows a random number of customers to arrive at the same time, we model these as batch arrivals. Individual batches arrive at rate λ, and each batch contains j customers with probability p_j, assumed to satisfy $\sum_{j\geq 1} p_j = 1$. Now the arrival rates change to $q(k, k + j) = p_j\lambda$ for $k \geq 0$ and $j \geq 1$.

The philosophy is again that of thinning: the original rate λ Poisson arrival process is decomposed into independent arrival processes of rates $p_j\lambda$. The rate $p_j\lambda$ process marks a jump from k to $k + j$. The non-existence of simultaneous occurrences still holds: two distinct batches never arrive at the same moment. △

Example 6.15 (Pure birth process, or Yule process)**.** Consider the following continuous time branching process. Individuals never die. Each individual gives birth to offspring (one at a time) according to a rate λ Poisson process, independently of other individuals. The offspring generate further offspring according to the same rule. Let X_t denote the number of individuals alive at time t. Then X_t is a continuous-time Markov chain with rates $q(k, k + 1) = k\lambda$ for $k \geq 1$. If zero is included in the state space then it is an absorbing state but irrelevant because it cannot be reached from any other state. △

Example 6.16 (Birth and death process)**.** In Example 6.15 assume also that each individual has an Exp(μ)-distributed random life time. Now the rates are

$$q(k, k + 1) = k\lambda \quad \text{and} \quad q(k, k - 1) = k\mu \text{ for } k \geq 1.$$

The state 0 is absorbing (extinction). △

Explosion. Before we turn to develop the theory of continuous-time Markov chains, a technical issue with the construction must be addressed. Namely, definition (6.11) determines X_t only as far in time as the intervals $[J_n, J_{n+1})$ go. Since the sequence

$\{J_n\}$ is increasing, there is a (possibly random) limit

$$J_\infty = \lim_{n\to\infty} J_n = \lim_{n\to\infty} \sum_{k=0}^{n-1} \frac{\tau_k}{\lambda(Y_k)} = \sum_{k=0}^{\infty} \frac{\tau_k}{\lambda(Y_k)}.$$

If $J_\infty = \infty$ the construction is complete: X_t is defined for all time $0 \le t < \infty$. But if J_∞ is finite, we have managed to define X_t only for $0 \le t < J_\infty$. Furthermore, since every time point J_n marks a jump of the process X_t, the process has made *infinitely many jumps* in the finite time interval $[0, J_\infty)$. Thus it is appropriate to use the term *explosion* for the event $J_\infty < \infty$. In a sense, the process accelerates so fast that it "runs off to infinity" in finite time.

We can get a precise understanding of when this happens through our deduction (6.12) of the distribution of the holding times. It turns out that an infinite series of independent exponential random variables either converges with probability one or diverges with probability one, and which case happens can be checked from the parameters. This fact is established by the next lemma.

Lemma 6.17. *Let $\{\alpha_k\}_{k\ge 0}$ be a sequence of positive real numbers and $\{\xi_k\}_{k\ge 0}$ independent exponential random variables with marginal distributions $\xi_k \sim \operatorname{Exp}(\alpha_k)$. Then this dichotomy holds:*

$$P\Big(\sum_{k=0}^{\infty} \xi_k < \infty\Big) = 1 \quad \textit{iff} \quad \sum_{k=0}^{\infty} \frac{1}{\alpha_k} < \infty$$
$$\textit{and} \quad P\Big(\sum_{k=0}^{\infty} \xi_k = \infty\Big) = 1 \quad \textit{iff} \quad \sum_{k=0}^{\infty} \frac{1}{\alpha_k} = \infty.$$

Proof. We prove only the *if* part of the first statement. (The *only if* part of the second statement is immediate. The rest requires advanced mathematical techniques.) By the monotone convergence theorem (Theorem B.5),

$$E\Big[\sum_{k=0}^{\infty} \xi_k\Big] = \sum_{k=0}^{\infty} E[\xi_k] = \sum_{k=0}^{\infty} \frac{1}{\alpha_k} < \infty.$$

By Lemma B.3, finiteness of the expectation implies $P\big(\sum_{k=0}^{\infty} \xi_k < \infty\big) = 1$. □

Theorem 6.18. *Conditional on the jump chain evolution $\{Z_n\}_{n\ge 0}$, explosion happens with probability zero if $\sum_{n=0}^{\infty} \frac{1}{\lambda(Z_n)} = \infty$, and with probability one if $\sum_{n=0}^{\infty} \frac{1}{\lambda(Z_n)} < \infty$.*

Proof. By (6.12), conditional on the jump chain evolution $\{Z_n\}_{n\ge 0}$, the holding times are independent exponential random variables with rates $\lambda(Z_n)$. Thus the theorem follows by application of the previous lemma. □

The next example illustrates how an explosion is produced by making the jump rates grow fast enough.

Example 6.19. Consider the continuous-time Markov chain with state space $\mathcal{S} = \mathbb{Z}_{\ge 0}$, routing matrix $r(k, k+1) = 1$ and holding parameters $\lambda(k) = (k+1)^\gamma$ where

γ is a constant. The jump chain satisfies $Z_n = Z_0 + n$. Thus

$$\sum_{n=0}^{\infty} \frac{1}{\lambda(Z_n)} = \sum_{n=0}^{\infty} (Z_0 + n)^{-\gamma}.$$

For any $Z_0 \geq 0$, this series diverges to ∞ iff $\gamma \leq 1$. Thus if $\gamma > 1$, explosion happens with probability one for every initial condition Z_0. △

The criterion of Theorem 6.18 can be difficult to apply since it requires tracking the random evolution of the holding rates $\lambda(Z_k)$ along the jump chain. However, some general facts can be established, listed in the next theorem.

Theorem 6.20. *The following are three sufficient conditions for preventing explosion of the continuous-time Markov chain X_t with routing matrix R and holding parameters $\{\lambda(x)\}$.*

(i) *If there is a constant α such that $\lambda(x) \leq \alpha$ for all states x, then no explosion can happen no matter how the process is started.*

(ii) *Suppose the state x is a recurrent state for the jump chain with transition matrix R. Then no explosion happens for the continuous-time Markov chain started from initial state x.*

(iii) *Let $\mathcal{A}_n(x)$ be the set of states that can be reached from the state x with n or fewer steps along the arrow diagram. Assume there is a constant $C(x)$ such that*

$$\max_{z \in \mathcal{A}_n(x)} \lambda(z) \leq C(x)n \qquad \text{for all } n \geq 1. \tag{6.26}$$

Then no explosion happens for the continuous-time Markov chain started from initial state x.

We leave the proof as an exercise. Condition (iii) says that as long as the holding parameter grows at most linearly, explosion is prevented. This was precisely the case $\gamma \leq 1$ in Example 6.19. Condition (i) is a special case of condition (iii). A finite state space $\mathcal{S}$ is a special case of condition (i).

An explosive Markov chain can be extended to all time in different ways. One way is to add an absorbing "cemetery state" Δ to the state space and send the process to Δ at time J_∞. Another way would be to restart the process from some particular state or distribution after each explosion. Explosive Markov chains will not be of interest to us in this book. Unless we specifically want to illustrate explosion, the examples we take up satisfy one of the conditions of Theorem 6.20.

6.3. Evolution of the transition probability function

The constructions in the previous section left the connection between rates and transition probabilities somewhat elusive. In this section we discuss differential equations that connect the two.

Rates from transition probabilities. Recall that by definition $p_0(x,x) = 1$ and $p_0(x,y) = 0$ for $x \neq y$.

Theorem 6.21. *For states $x \neq y$,*

$$q(x,y) = \lim_{h\to 0+} \frac{p_h(x,y)}{h} = \frac{d}{dt}p_t(x,y)\Big|_{t=0+} \tag{6.27}$$

and for each state x,

$$-\lambda(x) = \lim_{h\to 0+} \frac{p_h(x,x)-1}{h} = \frac{d}{dt}p_t(x,x)\Big|_{t=0+} \tag{6.28}$$

The theorem says that the rates are right derivatives of the transition probabilities at time zero. The limits can be taken only from the right because the transition probabilities are not defined for negative times. An alternative way to express (6.27) that illuminates the meaning of infinitesimal rates is to write

$$p_h(x,y) = hq(x,y) + o(h) \tag{6.29}$$

where $o(h)$ (pronounced "little-oh of h") is a quantity that satisfies $o(h)/h \to 0$ as $h \to 0$. Equation (6.29) says that for very small times h, the probability of moving from x to y is approximately $hq(x,y)$, with an error that is much smaller than h.

Proof of Theorem 6.21 for bounded rates. We prove Theorem 6.21 under the assumption that $\lambda(x) \leq \alpha$ for all states x, and use the construction $X_t = Y_{N_t}$ in terms of a discrete-time chain Y_n run by a rate α Poisson process N_t. Equation (6.7) gives the transition probabilities.

When $x \neq y$, $u^{(0)}(x,y) = 0$ and the term $k = 0$ vanishes in the series below. Separate the term $k = 1$ and treat the rest as an error term:

$$\begin{aligned}\frac{p_h(x,y)}{h} &= \frac{1}{h}\sum_{k=0}^{\infty} e^{-\alpha h}\frac{(\alpha h)^k}{k!}u^{(k)}(x,y)\\ &= e^{-\alpha h}\alpha\, u(x,y) + \frac{1}{h}\sum_{k=2}^{\infty} e^{-\alpha h}\frac{(\alpha h)^k}{k!}u^{(k)}(x,y).\end{aligned}$$

The last term is nonnegative and can be bounded as follows. Since we are letting $h \to 0+$ we can assume that $h \leq 1$. Use also $e^{-\alpha h} \leq 1$ and $u^{(k)}(x,y) \leq 1$. First take h^2 outside the sum.

$$\begin{aligned}\frac{1}{h}\sum_{k=2}^{\infty} e^{-\alpha h}\frac{(\alpha h)^k}{k!}u^{(k)}(x,y) &= h\sum_{k=2}^{\infty} e^{-\alpha h}\frac{\alpha^k h^{k-2}}{k!}u^{(k)}(x,y)\\ &\leq h\sum_{k=2}^{\infty}\frac{\alpha^k}{k!} \leq he^{\alpha}.\end{aligned} \tag{6.30}$$

This estimate shows that the error term tends to 0 has $h \to 0$. Thus from above we get

$$\lim_{h\to 0+}\frac{p_h(x,y)}{h} = \lim_{h\to 0+} e^{-\alpha h}\alpha u(x,y) = \alpha u(x,y) = q(x,y).$$

The last equality came from expression (6.20) for the jump rates. Equation (6.27) has been proved.

To deduce (6.28), from

$$p_h(x,x) = e^{-\alpha h} + e^{-\alpha h}\alpha h\, u(x,x) + \sum_{k=2}^{\infty} e^{-\alpha h}\frac{(\alpha h)^k}{k!}u^{(k)}(x,x)$$

we get

$$\frac{p_h(x,x)-1}{h} = \frac{e^{-\alpha h}-1}{h} + e^{-\alpha h}\alpha u(x,x) + \frac{1}{h}\sum_{k=2}^{\infty} e^{-\alpha h}\frac{(\alpha h)^k}{k!}u^{(k)}(x,x).$$

The last error term satisfies (6.30) with $x = y$ and the limit gives

$$\lim_{h\to 0+}\frac{p_h(x,x)-1}{h} = -\alpha + \alpha u(x,x) = -\lambda(x).$$

The last equality is from (6.19). □

Definition 6.22. Let $\{q(x,y) : x \neq y\}$ be the jump rates of a continuous-time Markov chain . Define the *generator matrix* as $Q = \{q(x,y)\}_{x,y\in\mathcal{S}}$ where the off-diagonal entries $q(x,y)$ for $x \neq y$ are the jump rates, and the diagonal entries are the negatives of the holding parameters:

$$q(x,x) = -\lambda(x). \tag{6.31}$$

△

The generator matrix is also called the *Q-matrix* of the process. It satisfies these three conditions: (i) the off-diagonal entries are nonnegative, (ii) the diagonal entries are nonpositive, and (iii) the row sums are zero. Conversely, any matrix with bounded entries, finite or infinite in dimension, is the generator matrix of continuous-time Markov chain if it satisfies the three conditions above.

Introduce also the shorthand notation $p'_t(x,y) = \frac{d}{dt}p_t(x,y)$ for the time derivatives of the transition probabilities and $\mathbf{P}'_t = \{p'_t(x,y)\}_{x,y\in\mathcal{S}}$ for the matrix of derivatives. Then the outcome of Theorem 6.21 can be expressed as

$$\mathbf{P}'_0 = Q. \tag{6.32}$$

The equation is to be understood entry-by-entry, as a shorthand for the system of equations

$$(d/dt)p_t(x,y)|_{t=0} = q(x,y) \qquad \text{for } x,y\in\mathcal{S}.$$

If $\mathcal{S}$ is finite, $\mathbf{P}'_t$ does also make sense as the derivative of the matrix-valued function $t \mapsto \mathbf{P}_t$.

Example 6.23 (Continuation of Example 6.8)**.** Let $X_t = Y_{N_t}$, a discrete-time Markov chain Y_n with transition matrix $U = \{u(x,y)\}_{x,y\in\mathcal{S}}$ run by an independent rate α Poisson process N_t. From the definition of Q and equations (6.19) and (6.20) we get

$$q(x,y) = \begin{cases} \alpha u(x,y), & y \neq x \\ -\lambda(x) = \alpha u(x,x) - \alpha, & y = x \end{cases} \tag{6.33}$$

which can be summarized by the matrix equation

$$Q = \alpha(U - I). \tag{6.34}$$

△

Chapman-Kolmogorov equations. Exactly as in discrete time, multiplication of transition matrices has probabilistic significance. Let $x, y \in \mathcal{S}$ and $s, t > 0$.

$$\begin{aligned} p_{s+t}(x,y) = P_x(X_{s+t} = y) &= \sum_{z \in \mathcal{S}} P_x(X_{s+t} = y, X_s = z) \\ &= \sum_{z \in \mathcal{S}} P_x(X_s = z) P_x(X_{s+t} = y \,|\, X_s = z) \\ &= \sum_{z \in \mathcal{S}} P_x(X_s = z) P_z(X_t = y) = \sum_{z \in \mathcal{S}} p_s(x,z)\, p_t(z,y). \end{aligned}$$

The penultimate equality used the Markov property to restart the process from state z at time s. The equation above is trivially true in the cases where s or $t = 0$ because $p_0(x,y) = I_{x=y}$.

To summarize, these are the Chapman-Kolmogorov equations for continuous-time Markov chains:

$$p_{s+t}(x,y) = \sum_{z \in \mathcal{S}} p_s(x,z)\, p_t(z,y) \qquad \text{for all } x, y \in \mathcal{S} \text{ and } s, t \geq 0. \tag{6.35}$$

Their matrix form is

$$\mathbf{P}_{s+t} = \mathbf{P}_s \mathbf{P}_t. \tag{6.36}$$

Transition probabilities in terms of rates.

Theorem 6.24. *The transition probabilities satisfy the following systems of differential equations.*

(a) Backward Kolmogorov equations:

$$\frac{d}{dt} p_t(x,y) = \sum_{z \neq x} q(x,z) p_t(z,y) - \lambda(x) p_t(x,y) \qquad \textit{for all } x, y \in \mathcal{S} \tag{6.37}$$

and in matrix form

$$\mathbf{P}_t' = Q\mathbf{P}_t \quad \textit{with initial condition } \mathbf{P}_0 = I. \tag{6.38}$$

(b) Forward Kolmogorov equations:

$$\frac{d}{dt} p_t(x,y) = \sum_{z \neq y} p_t(x,z) q(z,y) - \lambda(y) p_t(x,y) \qquad \textit{for all } x, y \in \mathcal{S} \tag{6.39}$$

and in matrix form

$$\mathbf{P}_t' = \mathbf{P}_t Q \quad \textit{with initial condition } \mathbf{P}_0 = I. \tag{6.40}$$

Proof. We give a proof that is rigorous for the right derivative of the finite state space case but otherwise only heuristic.

By the Chapman-Kolmogorov equations, for $t \geq 0$ and $h > 0$,

$$\mathbf{P}_{t+h} - \mathbf{P}_t = \mathbf{P}_h \mathbf{P}_t - \mathbf{P}_t = (\mathbf{P}_h - I)\mathbf{P}_t$$

and

$$\mathbf{P}_{t+h} - \mathbf{P}_t = \mathbf{P}_t \mathbf{P}_h - \mathbf{P}_t = \mathbf{P}_t(\mathbf{P}_h - I).$$

Divide by $h > 0$ and use (6.32) in the $h \to 0$ limit. This gives us the backward equation

$$\frac{d}{dt}\mathbf{P}_t = \lim_{h\to 0}\frac{\mathbf{P}_{t+h}-\mathbf{P}_t}{h} = \Bigl(\lim_{h\to 0}\frac{\mathbf{P}_h - I}{h}\Bigr)\mathbf{P}_t = \mathbf{P}_0'\mathbf{P}_t = Q\mathbf{P}_t.$$

and the forward equation

$$\frac{d}{dt}\mathbf{P}_t = \lim_{h\to 0}\frac{\mathbf{P}_{t+h}-\mathbf{P}_t}{h} = \mathbf{P}_t\Bigl(\lim_{h\to 0}\frac{\mathbf{P}_h - I}{h}\Bigr) = \mathbf{P}_t\mathbf{P}_0' = \mathbf{P}_t Q.$$

□

Kolmogorov's equations for finite state space. When the state space is finite we can augment Theorem 6.24 with a uniqueness statement and a representation of the solution $\mathbf{P}_t$.

To point us in the right direction, consider first the scalar case of this ordinary differential equation for an unknown differentiable function $x(t)$:

$$x'(t) = qx(t), \quad x(0) = 1. \tag{6.41}$$

To solve this, multiply by an integrating factor e^{-qt} and notice that the derivative of a product arises:

$$\begin{aligned} x'(t) = qx(t) \iff x'(t) - qx(t) = 0 \iff e^{-qt}x'(t) - qe^{-qt}x(t) = 0 \\ \iff \frac{d}{dt}\bigl(e^{-qt}x(t)\bigr) = 0. \end{aligned}$$

Since the function $e^{-qt}x(t)$ must be constant in time, we get $x(t) = x(0)e^{qt} = e^{qt}$.

To use this idea for matrices, we define the exponential function of matrices by adapting the series expansion $e^x = \sum_{k=0}^\infty \frac{x^k}{k!}$ to a finite square matrix A:

$$e^A = \sum_{k=0}^\infty \frac{A^k}{k!},$$

with the usual convention $A^0 = I$. For finite square matrices this series is well defined. It obeys some familiar rules, for example the addition of exponents:

$$\begin{aligned} e^{A+B} &= \sum_{k=0}^\infty \frac{1}{k!}(A+B)^k = \sum_{k=0}^\infty \frac{1}{k!}\sum_{j=0}^k \binom{k}{j} A^j B^{k-j} \\ &= \sum_{j=0}^\infty \frac{1}{j!}A^j \sum_{k=j}^\infty \frac{1}{(k-j)!}B^{k-j} = e^A\, e^B. \end{aligned} \tag{6.42}$$

We can state the theorem.

Theorem 6.25. *Let Q be the generator matrix of a continuous-time Markov chain with finite state space $\mathcal{S}$. Then both the Kolmogorov backward equation (6.38) and the Kolmogorov forward equation (6.40) have the unique solution*

$$\mathbf{P}_t = e^{tQ} = \sum_{k=0}^\infty \frac{t^k}{k!}Q^k. \tag{6.43}$$

Proof. Convergent power series can be differentiated term by term. This applies also to a matrix-valued series. After differentiation we can take a Q factor as a common factor both on the left and on the right, resulting in both the backward and the forward equation.

$$\begin{aligned}\frac{d}{dt}e^{tQ} &= \frac{d}{dt}\sum_{k=0}^{\infty}\frac{t^k}{k!}Q^k = \sum_{k=0}^{\infty}\frac{d}{dt}\left(Q^k\frac{t^k}{k!}\right) = \sum_{k=1}^{\infty}Q^k\frac{kt^{k-1}}{k!}\\ &= \sum_{k=1}^{\infty}Q^k\frac{t^{k-1}}{(k-1)!} = Q\bigg(\sum_{j=0}^{\infty}Q^j\frac{t^j}{j!}\bigg) = \bigg(\sum_{j=0}^{\infty}Q^j\frac{t^j}{j!}\bigg)Q\\ &= Qe^{tQ} = e^{tQ}Q.\end{aligned}$$

The initial condition is also satisfied: at $t = 0$, the definition of the matrix exponential gives

$$e^{0Q} = \sum_{k=0}^{\infty}\frac{0^k}{k!}Q^k = I$$

because by convention $0^0 = 0! = 1$ and $Q^0 = I$, while $0^k = 0$ for $k \geq 1$ and therefore all these terms vanish. This shows that e^{tQ} solves both equations (6.38) and (6.40).

The solution to either differential equation is unique by Theorem A.5 because the functions $f(A) = QA$ and $f(A) = AQ$ defined on square matrices are both Lipschitz continuous. Hence it must be that $\mathbf{P}_t = e^{tQ}$. □

Example 6.26 (Continuation of Example 6.23)**.** Let $X_t = Y_{N_t}$ where Y_n is a discrete-time Markov chain with transition matrix $U = \{u(x,y)\}_{x,y\in\mathcal{S}}$ and N_t is an independent rate α Poisson process. Assume now that $\mathcal{S}$ is finite. In Example 6.23 we found the generator matrix $Q = \alpha(U - I)$. We compute the exponential e^{tQ}. First note that, since $I^k = I$ for all $k \geq 0$, then for any scalar γ,

$$e^{\gamma I} = \sum_{k=0}^{\infty}\frac{\gamma^k}{k!}I^k = \bigg(\sum_{k=0}^{\infty}\frac{\gamma^k}{k!}\bigg)I = e^{\gamma}I. \tag{6.44}$$

Then, applying (6.42) and the above, the transition probability matrix is

$$\begin{aligned}\mathbf{P}_t = e^{tQ} = e^{\alpha t(U-I)} = e^{\alpha tU}e^{-\alpha tI} &= \bigg(\sum_{k=0}^{\infty}\frac{(\alpha t)^k}{k!}U^k\bigg)e^{-\alpha t}I\\ &= \sum_{k=0}^{\infty}e^{-\alpha t}\frac{(\alpha t)^k}{k!}U^k.\end{aligned}$$

This is exactly the matrix version of equation (6.7). △

Example 6.27. Fix two positive constants λ, μ. We derive the transition probability function $\mathbf{P}_t$ of the continuous-time Markov chain on the state space $\mathcal{S} = \{0, 1\}$ with generator matrix

$$Q = \begin{bmatrix}-\lambda & \lambda\\ \mu & -\mu\end{bmatrix}.$$

To this end, we solve the forward equations

$$\mathbf{P}'_t = \mathbf{P}_tQ \quad\Longleftrightarrow\quad \begin{bmatrix}p'_t(1,1) & p'_t(1,2)\\ p'_t(2,1) & p'_t(2,2)\end{bmatrix} = \begin{bmatrix}p_t(1,1) & p_t(1,2)\\ p_t(2,1) & p_t(2,2)\end{bmatrix}\begin{bmatrix}-\lambda & \lambda\\ \mu & -\mu\end{bmatrix}.$$

Start with the equation for the upper left corner entry. Use $p_t(1,2) = 1 - p_t(1,1)$, rearrange, multiply by an integrating factor, integrate over $[0,t]$, and use the initial condition $p_0(1,1) = 1$:

$$\begin{aligned}
&p_t'(1,1) = -\lambda p_t(1,1) + \mu p_t(1,2) = \mu - (\lambda+\mu)p_t(1,1)\\
\Longrightarrow\;& p_t'(1,1) + (\lambda+\mu)p_t(1,1) = \mu\\
\Longrightarrow\;& \frac{d}{dt}\big[e^{(\lambda+\mu)t}p_t(1,1)\big] = \mu\, e^{(\lambda+\mu)t}\\
\Longrightarrow\;& e^{(\lambda+\mu)t}p_t(1,1) - p_0(1,1) = \int_0^t \mu\, e^{(\lambda+\mu)s}\,ds = \frac{\mu}{\lambda+\mu}e^{(\lambda+\mu)t} - \frac{\mu}{\lambda+\mu}\\
\Longrightarrow\;& p_t(1,1) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}.
\end{aligned}$$

From $p_t(1,2) = 1 - p_t(1,1)$ we get

$$p_t(1,2) = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}.$$

The same differential equation develops for $p_t(2,1)$ but with initial condition $p_0(1,2) = 0$. Hence the solution is

$$p_t(2,1) = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}$$

and the complementary probability is

$$p_t(2,2) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}.$$

Summary of the result in matrix form:

$$\mathbf{P}_t = \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{bmatrix} + e^{-(\lambda+\mu)t}\begin{bmatrix} \frac{\lambda}{\lambda+\mu} & -\frac{\lambda}{\lambda+\mu} \\ -\frac{\mu}{\lambda+\mu} & \frac{\mu}{\lambda+\mu} \end{bmatrix}.$$

The backward equations can also be solved explicitly, with a tiny bit more work than the forward equations required.

From the result follows the limit

$$\lim_{t\to\infty}\mathbf{P}_t = \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{bmatrix},$$

foreshadowing the convergence theorem presented in the next section. △

6.4. Invariant distributions and asymptotic behavior

For a continuous-time Markov chain the probability distribution of the time t state X_t is computed from the initial distribution μ and the transition probability matrix $\mathbf{P}_t$ in the same way as in discrete time. If we represent μ as a row vector then the row vector $\mu\mathbf{P}_t$ gives the probability mass function of X_t:

$$\begin{aligned} P_\mu(X_t = y) &= \sum_x P_\mu(X_0 = x) P_\mu(X_t = y \,|\, X_0 = x) \\ &= \sum_x \mu(x)\, p_t(x,y) = (\mu\mathbf{P}_t)_y. \end{aligned} \tag{6.45}$$

Probability distributions that are invariant for the evolution are important again, exactly as for discrete-time Markov chains.

Definition 6.28. A probability measure π on the state space $\mathcal{S}$ is an **invariant distribution** (also **stationary distribution**) for the continuous-time Markov chain with transition probability function $\mathbf{P}_t$ if

$$\pi\mathbf{P}_t = \pi \qquad \text{for all } t \geq 0. \tag{6.46}$$

△

Thus if we start the Markov chain with an invariant distribution π then at any fixed time $t > 0$ the state X_t has the same distribution π: $P_\pi(X_t = y) = \pi(y)$. Furthermore, as in discrete time, the entire process $\{X_t : t \geq 0\}$ is a *stationary process* which means that its probability distribution does not change under time shifts: for all time points $0 \leq s_0 < \ldots < s_n$, $t > 0$, and all states $x_0, \ldots, x_n \in \mathcal{S}$,

$$\begin{aligned} &P_\pi\big(X_{s_0+t} = x_0,\, X_{s_1+t} = x_1, \ldots, X_{s_n+t} = x_n\big) \\ &\qquad = P_\pi\big(X_{s_0} = x_0,\, X_{s_1} = x_1, \ldots, X_{s_n} = x_n\big). \end{aligned} \tag{6.47}$$

The reader should be immediately warned that an invariant distribution of the jump chain is not necessarily an invariant distribution of the corresponding continuous-time Markov chain. The reason is that the jump chain does not take into account the lengths of time spent in different states. This is illustrated in Example 6.31 below and made precise in Theorem 6.32.

Example 6.29 (Continuation of Example 6.26)**.** Let $X_t = Y_{N_t}$ where Y_n is a discrete-time Markov chain with transition matrix $U = \{u(x,y)\}_{x,y\in\mathcal{S}}$ and N_t is an independent rate α Poisson process. Since the clock N_t ticks at a constant average rate, the switch between discrete and continuous time does not involve a time distortion. As a result X_t and Y_n have the same invariant distributions:

$$\pi \text{ is invariant for } X_t \text{ iff } \pi U = \pi. \tag{6.48}$$

For the proof we utilize formula (6.7) for the transition probability $p_t(x,y)$:

$$\begin{aligned} &\pi(y) = \sum_x \pi(x) p_t(x,y) = \sum_x \pi(x) \sum_{k=0}^\infty e^{-\alpha t} \frac{(\alpha t)^k}{k!} u^{(k)}(x,y) \\ \iff\quad & e^{\alpha t}\pi(y) = \sum_{k=0}^\infty \Big(\sum_x \pi(x) u^{(k)}(x,y) \Big) \frac{(\alpha t)^k}{k!} \\ \iff\quad & \sum_{k=0}^\infty \pi(y) \frac{(\alpha t)^k}{k!} = \sum_{k=0}^\infty \Big(\sum_x \pi(x) u^{(k)}(x,y) \Big) \frac{(\alpha t)^k}{k!}. \end{aligned} \tag{6.49}$$

On the last line on the left we replaced $e^{\alpha t}$ with its series expansion. Convergent power series agree on a nontrivial interval iff their coefficients coincide. Thus having

$\pi(y) = \sum_x \pi(x)p_t(x,y)$ for an interval of t-values is equivalent to having $\pi(y) = \sum_x \pi(x)u^{(k)}(x,y)$ for all $k \geq 0$, which is equivalent to the single condition $\pi(y) = \sum_x \pi(x)u(x,y)$. △

Definition (6.46) is not in general very practical, for all the same reasons that transition probabilities are not always useful. There is a quick check of invariance in terms of the generator matrix.

Theorem 6.30. *The probability distribution π is invariant for a (non-explosive) continuous-time Markov chain if and only if*

$$\pi Q = 0. \tag{6.50}$$

(Here 0 is a row vector of zeros.) This is equivalent to the system of equations

$$\sum_{x:\, x\neq y} \pi(x)q(x,y) = \pi(y)\lambda(y) \qquad \text{for all } y \in \mathcal{S}. \tag{6.51}$$

Proof. Again we give a proof that is valid for the finite state space case. Applying both Kolmogorov's backward and forward equations gives us these equalities:

$$\frac{d}{dt}\big(\pi\mathbf{P}_t\big) = \pi\mathbf{P}_t' = \pi Q\mathbf{P}_t = \pi\mathbf{P}_t Q. \tag{6.52}$$

To be precise, the time derivates are taken entry by entry. For example, the first derivative above represents the vector of derivatives

$$\frac{d}{dt}\big(\pi\mathbf{P}_t\big) = \Big(\frac{d}{dt}(\pi\mathbf{P}_t)_y : y \in \mathcal{S}\Big),$$

indexed by the state space $\mathcal{S}$. The finite $\mathcal{S}$ assumption allows the differentiation operation to move freely through sums: for example, the first equality of (6.52) consists of the following equalities for each state y:

$$\frac{d}{dt}(\pi\mathbf{P}_t)_y = \frac{d}{dt}\Big(\sum_x \pi(x)p_t(x,y)\Big) = \sum_x \pi(x)p_t'(x,y) = (\pi\mathbf{P'}_t)_y.$$

Now we derive the theorem from (6.52). Assume π is invariant. Then $\pi\mathbf{P}_t = \pi$ is constant in time and (6.52) leads to

$$0 = \frac{d}{dt}(\pi\mathbf{P}_t) = \pi\mathbf{P}_t Q = \pi Q.$$

Next assume that $\pi Q = 0$. Then (6.52) gives

$$\frac{d}{dt}\big(\pi\mathbf{P}_t\big) = \pi Q\mathbf{P}_t = 0\mathbf{P}_t = 0.$$

Thus $\pi\mathbf{P}_t$ is constant in time, which gives $\pi\mathbf{P}_t = \pi\mathbf{P}_0 = \pi I = \pi$. □

Equations (6.51) are insightful: they say that at each state y, the incoming probability flow (on the left) equals the outgoing probability flow (on the right).

Example 6.31. Consider the 2-state continuous-time Markov chain with state space $\mathcal{S} = \{0, 1\}$ and generator matrix

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

The equation $\pi Q = 0$ gives the unique invariant distribution

$$\pi = \left[\frac{\mu}{\lambda+\mu} \;\; \frac{\lambda}{\lambda+\mu} \right]$$

The routing matrix is

$$R = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The unique invariant distribution of the routing matrix is $[\,\frac{1}{2}\;\frac{1}{2}\,]$. Hence the two invariant distributions coincide iff $\lambda = \mu$. This is the case where we can represent the continuous-time Markov chain as $X_t = Z_{N_t}$, where Z_n is the jump chain and N_t is an independent rate $\lambda = \mu$ Poisson process. This is a special case of Example 6.29. △

With the help of the generator criterion, we can clarify the distinction between invariance for the continuous-time Markov chain and invariance for the jump chain.

Theorem 6.32. *Let π be a probability measure on the state space $\mathcal{S}$, and define $\nu(x) = \lambda(x)\pi(x)$. Then π is an invariant distribution for the continuous-time Markov chain iff ν is an invariant measure for the routing matrix R.*

Proof. Recall that $q(x,y) = \lambda(x)r(x,y)$ for $x \neq y$ and $\lambda(x)r(x,x) = 0$ for all states because either x is absorbing and $\lambda(x) = 0$ or x is nonabsorbing and $r(x,x) = 0$. Beginning from $\pi Q = 0$ in the form (6.51),

$$\sum_{x:\,x\neq y} \pi(x)q(x,y) = \pi(y)\lambda(y) \quad\Longleftrightarrow\quad \sum_{x:\,x\neq y} \pi(x)\lambda(x)r(x,y) = \pi(y)\lambda(y)$$
$$\Longleftrightarrow\quad \sum_{x\in\mathcal{S}} \pi(x)\lambda(x)r(x,y) = \pi(y)\lambda(y) \quad\Longleftrightarrow\quad \sum_{x\in\mathcal{S}} \nu(x)r(x,y) = \nu(y).$$

□

The above theorem allows us to produce some surprising phenomena.

Example 6.33 (Symmetric random walk with varying rates)**.** In Example 2.78 of Section 2.5 we discovered that any constant measure on $\mathbb{Z}$ is an invariant measure for discrete-time symmetric simple random walk on the integers. Furthermore, SSRW is null recurrent and has no invariant distributions.

Now consider a continuous-time symmetric random walk X_t with routing matrix $r(x, x\pm 1) = \frac{1}{2}$ and holding parameters $\lambda(0) = 2$ and $\lambda(x) = 2^{|x|+2}$ for $x \neq 0$. Define the probability measure π on $\mathbb{Z}$ by

$$\pi(0) = \tfrac{1}{2} \quad \text{and} \quad \pi(x) = \left(\tfrac{1}{2}\right)^{|x|+2} \quad \text{for} \quad x \neq 0.$$

Then $\lambda(x)\pi(x) = 1$ for all $x \in \mathbb{Z}$ and Theorem 6.32 implies that π is an invariant distribution for X_t.

What explains that this continuous-time walk has an invariant distribution? The discrete-time walk is null-recurrent because over time it takes longer and longer excursions away from the origin, though recurrence always forces it to return eventually. For the continuous-time walk the large holding parameters $\lambda(x) = 2^{|x|+2}$ force the walk to run through its far-away excursions very fast so that the bulk of the time is spent around the origin.

The asymptotic frequency version (6.56) of the law of large numbers stated below allows us to quantify this last statement: the fraction of time spent in the interval $\{-m, -m+1, \dots, m-1, m\}$ is asymptotically

$$\sum_{|x| \le m} \pi(x) = 1 - \left(\tfrac{1}{2}\right)^{m+1}.$$

△

Reversibility. As for discrete-time Markov chains, the notion of reversibility is an important sharpening of stationarity.

Definition 6.34. A probability measure π on $\mathcal{S}$ is **reversible** for a continuous-time Markov chain $\{X_t : t \ge 0\}$ with transition probability function $\mathbf{P}_t$ if

$$\pi(x)p_t(x,y) = \pi(y)p_t(y,x) \qquad \text{for all states } x \ne y \text{ and } t \ge 0. \tag{6.53}$$

△

The next theorem collects the properties of reversibility, including the simple criterion in terms of the generator matrix. The condition (6.55) below is also called *detailed balance.*

Theorem 6.35.

(a) *A reversible probability measure is invariant.*

(b) *If the process X_t is started with a reversible initial distribution π, it has the same distribution forward and backward in time: for all time points $0 \le s_0 < \dots < s_n \le t$ and all states $x_0, \dots, x_n \in \mathcal{S}$,*

$$\begin{aligned} &P_\pi\big(X_{s_0} = x_0,\, X_{s_1} = x_1, \dots, X_{s_n} = x_n\big) \\ &\qquad\qquad = P_\pi\big(X_{t-s_0} = x_0,\, X_{t-s_1} = x_1, \dots, X_{t-s_n} = x_n\big). \end{aligned} \tag{6.54}$$

(c) *A probability measure π is reversible iff*

$$\pi(x)q(x,y) = \pi(y)q(y,x) \qquad \textit{for all states } x \ne y. \tag{6.55}$$

Proof. Parts (a) and (b) are proved similarly to their counterparts in discrete theory. A proof of part (c) in the case of bounded holding parameters can be patterned after the calculation in (6.49). We leave these proofs as exercises. □

Section 6.5 will discuss a number of examples. We close this section with a discussion of the limit theorems.

Limit theorems. For continuous-time Markov chains, irreducibility means the same as for discrete-time Markov chains, namely, that it is possible to go from any state to any other state. Since the moves are encoded by the jump chain, we can state equivalently that a continuous-time Markov chain is irreducible iff the jump chain is irreducible. An important consequence of irreducibility is that there can be at most one invariant distribution, as was the case for discrete-time chains. The next theorem combines the important limit theorems.

Theorem 6.36. *An irreducible continuous-time Markov chain with an invariant distribution π satisfies the following limit theorems.*

(a) Limit distribution. *For any initial distribution μ and any state y*

$$\lim_{t\to\infty} P_\mu(X_t = y) = \pi(y).$$

(b) Strong law of large numbers. *Let $f : \mathcal{S} \to \mathbb{R}$ be a function such that $\sum_x \pi(x)|f(x)| < \infty$. Then for any initial distribution μ, the limit below holds with probability one:*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X_s)\,ds = \sum_x \pi(x)f(x).$$

In particular, the asymptotic fraction of time spent at a given state $x \in \mathcal{S}$ is $\pi(x)$:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t I(X_s = x)\,ds = \pi(x) \qquad \textit{with probability one.} \tag{6.56}$$

As in discrete-time, the notions of recurrence and transience of a state capture whether return is certain or not. The definition of the return time has to take into consideration the fact that the process always remains for a while at each state. Thus for a continuous-time Markov chain we define the *first return time* T_x of a state x as the first time t that $X_t = x$ *after* a visit to another state. Precisely,

$$T_x = \inf\{t \ge 0 : X_t = x, \text{ there exists } s \in [0,t) \text{ such that } X_s \ne x\}. \tag{6.57}$$

So if $X_0 \ne x$ then T_x is the first visit to x, but if $X_0 = x$ then T_x is the first return to x after a jump away from x. The convention is that $\inf \varnothing = \infty$ so definition (6.57) is not good for absorbing states. Hence in the definition of recurrence below, absorbing states have to be mentioned separately.

Definition 6.37. The state x is **recurrent** if $P_x(T_x < \infty) = 1$ or if x is absorbing ($\lambda(x) = 0$), otherwise x is **transient**. A recurrent state $x \in \mathcal{S}$ is **positive recurrent** if $E_x[T_x] < \infty$ or if x is absorbing, and **null recurrent** if x is not absorbing but $E_x[T_x] = \infty$. △

Recurrence and transience can be determined from the jump chain because they only involve checking whether return is certain. Not so for positive and null recurrence because their distinction rests on the time it takes to return. Recall Example 6.33.

The next theorem summarizes the facts on recurrence, transience and invariant distributions. They are identical to those of the discrete theory.

Theorem 6.38. *Consider an irreducible continuous-time Markov chain.*

(a) *If there is a recurrent state then all states are recurrent.*

(b) *If there is a positive recurrent state then all states are positive recurrent.*

(c) *There exists a stationary distribution if and only if all states are positive recurrent.*

(d) *If the state space is finite, then all states are positive recurrent and there is a unique stationary distribution.*

As the last theoretical item, we state the connection between the expected return time and the stationary distribution.

Theorem 6.39. *Suppose π is an invariant distribution for an irreducible continuous-time Markov chain. Then*

$$\pi(x) = \frac{1}{\lambda(x)E_x[T_x]} \qquad \textit{for all states } x. \tag{6.58}$$

Proof. This formula can be explained by the limit (6.56). Fix a state x and consider the renewal process whose arrival times are the times when the process jumps to x from another state. By the strong Markov property, the process starts anew at x independently of the past, and consequently the cycle lengths are i.i.d. random variables.

The mean cycle length is $E_x[T_x]$, by the definition of T_x. In a single cycle the amount of time spent in state x has $\text{Exp}(\lambda(x))$ distribution and hence expectation $1/\lambda(x)$. By the renewal-reward limit of Theorem 1.24, the asymptotic fraction of time spent at x is $\frac{1/\lambda(x)}{E_x(T_x)}$. By (6.56) this limit is also equal to $\pi(x)$, and this match gives (6.58). □

Example 6.40 (Modified from Example 4.13 from Durrett)**.** Consider the following weather model for LA. There are three states:1 = sunny, 2 = smoggy, and 3 = rainy. The weather stays sunny for an exponentially distributed number of days with mean 3, then becomes smoggy. It stays smoggy for an exponentially distributed number of days with mean 4, then with probability 1/2 rain comes and with probability 1/2 it will be sunny. The rain lasts for an exponentially distributed number of days with mean 1, then sunshine returns. Asymptotically what is the ratio of sunny days? What is the expected wait time until the weather is sunny if it just turned smoggy?

The holding rates are $\lambda(1) = 1/3$, $\lambda(2) = 1/4$, $\lambda(3) = 1$. The transition probabilities for the embedded chain are

$$r(1,2) = 1, \quad r(2,1) = r(2,3) = \frac{1}{2}, \quad r(3,1) = 1.$$

Hence the Q-matrix is

$$\begin{bmatrix} -1/3 & 1/3 & 0 \\ 1/8 & -1/4 & 1/8 \\ 1 & 0 & -1 \end{bmatrix}$$

This is a finite irreducible MC, so there is a unique stationary distribution π that satisfies $\pi Q = 0$. The equations awe get are

$$\begin{aligned} -1/3\pi(1) + 1/8\pi(2) + \pi(3) &= 0 \\ 1/3\pi(1) - 1/4\pi(2) \qquad &= 0 \\ 1/8\pi(2) - \pi(3) &= 0. \end{aligned}$$

From the second and third equations we get

$$\pi(1) = \frac{3}{4}\pi(2), \qquad \pi(3) = 1/88\pi(2)$$

and $\pi(1)+\pi(2)+\pi(3)=1$ leads to

$$\pi(1)=\tfrac{2}{5}\quad \pi(2)=\tfrac{8}{15}\quad \pi(3)=\tfrac{1}{15}$$

The stationary distribution gives the asymptotic frequency of time spent in the appropriate states. Thus in the long run the weather is sunny 2/5 of the time.

We can answer the second question by considering $E_1[T_1]$ (the expected return time to sunny state), and subtract the expected time we spend in the sunny state:

$$E_1[T_1]-\frac{1}{\lambda(1)}=\frac{1}{\lambda(1)\pi(1)}-\frac{1}{\lambda(1)}=\frac{1}{\frac{1}{3}\cdot\frac{2}{5}}-\frac{1}{1/3}=\frac{9}{2}.$$

Another way to get this is by noticing that the quantity we try to compute is exactly the expected time not spent in the sunny state between two returns to the sunny state. The ratio of time spent not in the sunny state is $1-\pi(1)=\frac{3}{5}$, the expected return time is

$$E_1[T_1]=\frac{1}{\lambda(1)\pi(1)}=\frac{1}{\frac{1}{3}\cdot\frac{2}{5}}=\frac{15}{2},$$

so the expected wait time will be $E_1[T_1](1-\pi(1))=\frac{15}{2}\cdot\frac{3}{5}=\frac{9}{2}$.
Note that by the memoryless property it does not matter if the weather just turn smoggy, only that it is smoggy now. Hence our last computation gave $E_2[T_1]$. △

6.5. Birth and death chains and queueing models

This section takes a closer at an important class of reversible examples. A *continuous-time birth and death chain* is a continuous-time Markov chain on the state space $\mathbb{Z}_{\geq 0}$ of nonnegative integers whose admissible jumps are to nearest neighbor points only. This can be a model of the size of a population that loses or gains one individual at a time. This formulation covers the special case that appeared in Example 6.16 and also many queueing models.

For this discussion we abbreviate the jump rates as

$$\lambda_k=q(k,k+1) \text{ for } k\geq 0 \quad\text{and}\quad \mu_k=q(k,k-1) \text{ for } k\geq 1.$$

To have an irreducible process on the entire state space $\mathbb{Z}_{\geq 0}$ we assume that all these rates are strictly positive.

The detailed balance equations for a reversible measure π are

$$\pi(k)\lambda_k=\pi(k+1)\mu_{k+1}\qquad \text{for all } k\geq 0.$$

These can be solved recursively in terms of $\pi(0)$:

$$\pi(k)=\pi(k-1)\frac{\lambda_{k-1}}{\mu_k}=\pi(k-2)\frac{\lambda_{k-2}\lambda_{k-1}}{\mu_{k-1}\mu_k}=\cdots=\pi(0)\prod_{j=0}^{k-1}\frac{\lambda_j}{\mu_{j+1}}.$$

This measure π can be normalized into a probability measure if the total sum is finite. Hence there is a reversible distribution iff

$$\sum_{k=0}^{\infty}\prod_{j=0}^{k-1}\frac{\lambda_j}{\mu_{j+1}}<\infty. \tag{6.59}$$

(For $k = 0$ the product $\prod_{j=0}^{k-1}$ is by convention equal to 1.) When the series above converges, the reversible distribution is given by

$$\pi(m) = \frac{\prod_{j=0}^{m-1} \frac{\lambda_j}{\mu_{j+1}}}{\sum_{k=0}^{\infty} \prod_{j=0}^{k-1} \frac{\lambda_j}{\mu_{j+1}}} \qquad \text{for } m \in \mathbb{Z}_{\geq 0}. \tag{6.60}$$

Formula (6.60) works also in the finite state space case. Fix a finite integer $K > 0$ and restrict the process to the state space $\{0, 1, \ldots, K\}$. Now $\lambda_K = 0$ so that jumps outside this space do not happen. Then for $m \in \{0, 1, \ldots, K\}$ the reversible distribution $\pi(m)$ is given by the same formula (6.60), once we restrict the sum in the denominator to $\sum_{k=0}^{K}$.

Example 6.41 (Revisit of Example 6.10, the M/M/1 queue)**.** The state space is $\mathbb{Z}_{\geq 0}$, the arrival rate is $\lambda_k = \lambda > 0$ for $k \geq 0$ and the service rate is $\mu_k = \mu > 0$ for $k \geq 1$. Condition (6.59) for the existence of a reversible distribution becomes

$$\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k < \infty \tag{6.61}$$

which is equivalent to $\lambda < \mu$. In this case the reversible distribution is a shifted geometric distribution:

$$\pi(k) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k \qquad \text{for } k \geq 0.$$

The reader can check (Exercise 6.3) that the invariance condition $\pi Q = 0$ leads to exactly the same result: an invariant distribution exists iff $\lambda < \mu$.

From random walk considerations we can understand why the case $\lambda \geq \mu$ does not have an invariant distribution. The holding parameters are constant away from the origin: for $k \geq 1$, $\lambda(k) = \lambda + \mu$ and the routing matrix is

$$r(k, k+1) = \frac{\lambda}{\lambda + \mu} \quad \text{and} \quad r(k, k-1) = \frac{\mu}{\lambda + \mu}.$$

From $r(k, k \pm 1)$ we see that if $\lambda = \mu$ then the process behaves as a symmetric random walk restricted to the positive half-line. This process is recurrent but not positive recurrent. If $\lambda > \mu$ then the process behaves like a transient random walk that diverges to ∞.

For the queuing application, the condition $\lambda < \mu$ means that the service rate is strictly larger than the arrival rate, which means that the server can handle the flow of customers. A queue that satisfies this condition is called *stable*. △

Example 6.42 (M/M/∞ queue)**.** Suppose that we have infinitely many servers. Then each customer goes immediately into service. We assume that customers arrive at rate λ and that each server serves at rate μ. Then the jump rates are $\lambda_k = \lambda$ and $\mu_k = k\mu$. Solving the equations for π yields a Poisson distribution.

$$\pi(k) = e^{-\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \qquad \text{for } k \geq 0.$$

Now there is no restriction on λ and μ because no matter how large λ is relative to μ, for large enough k we have $k\mu > \lambda$. At that point the system begins to serve customers on average faster than they arrive. △

Example 6.43 (M/M/1/2 queue)**.** A small barbershop with a single barber has one chair for service and one chair for waiting. The arrival rate of customers is λ and the service rate is μ. A customer who comes when both chairs are occupied leaves immediately. What is the long-term fraction of time that the barber is working? What is the long term average number of customers in the barber shop? What is the asymptotic rate at which customers are lost due to the shop being full?

Let X_t be the number of customers in the system. This is a birth and death chain with state space $\{0, 1, 2\}$ and jump rates

$$q(0,1) = q(1,2) = \lambda \quad \text{and} \quad q(1,0) = q(2,1) = \mu.$$

The detailed balance equations are

$$\pi(1) = \frac{\lambda}{\mu}\pi(0) \quad \text{and} \quad \pi(2) = \frac{\lambda}{\mu}\pi(1) = \left(\frac{\lambda}{\mu}\right)^2 \pi(0) \tag{6.62}$$

which yield the reversible distribution

$$\pi(0) = \frac{\mu^2}{\lambda^2 + \lambda\mu + \mu^2}\,, \quad \pi(1) = \frac{\lambda\mu}{\lambda^2 + \lambda\mu + \mu^2}\,, \quad \pi(2) = \frac{\lambda^2}{\lambda^2 + \lambda\mu + \mu^2}.$$

The barber works when the process is in state 1 or 2. Hence by Theorem 6.36 the asymptotic fraction of time that the barber works is

$$\pi(1) + \pi(2) = \frac{\lambda^2 + \lambda\mu}{\lambda^2 + \lambda\mu + \mu^2}.$$

The SLLN gives the long term average number of customers in the barber shop:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t X_s\,ds = \sum_{k=0}^{2} k\pi(k) = \frac{2\lambda^2 + \lambda\mu}{\lambda^2 + \lambda\mu + \mu^2}.$$

Heuristically, the asymptotic rate of lost customers should be $\lambda\pi(2)$ because the long term fraction of time spent in state 2 is $\pi(2)$, and during these periods lost customers come by at rate λ.

Here is more rigorous calculation that combines Markov chain theory and renewal-reward theory. Consider the renewal process whose arrivals are the times when the process jumps into state 2. From (6.58),

$$\text{mean cycle length } = E_2[T_2] = \frac{1}{\pi(2)\lambda(2)} = \frac{1}{\pi(2)\mu}.$$

Let the reward of each cycle be the number of lost customers. In mathematical notation, if X denotes the amount of time spent in state 2 in a cycle and N_t is the rate λ Poisson arrival process of customers, then the number of lost customers is N_X, the number of arrivals in time interval $[0, X]$. A key point is that once the system goes into state 2, the duration X is independent of the arrival process because X depends only on how fast the barber serves the current customer. Thus

$$\text{expected reward in a cycle } = E[N_X] = E\big[E[N_X \,|\, X\,]\big] = E[\lambda X] = \lambda/\mu.$$

$EX = 1/\mu$ because $X \sim \text{Exp}(\mu)$. Calculation of the conditional expectation above goes as follows:

$$E[N_X \,|\, X = t\,] = E[N_t \,|\, X = t\,] = E[N_t] = \lambda t$$

where the second equality uses the independence. Then $E[N_X \mid X] = \lambda X$.

The renewal-reward limit of Theorem 1.24 now gives the rate that was expected above:

$$\begin{aligned}&\text{asymptotic rate of losing customers}\\ &\qquad = \frac{\text{expected number of customers lost per cycle}}{\text{mean cycle length}} = \frac{\lambda/\mu}{\bigl(\frac{1}{\pi(2)\mu}\bigr)} = \lambda\pi(2).\end{aligned}$$

Here is another way to get the same answer. We derive the rate of served customers and subtract it from the total rate λ of customers arriving. Label the served customers as type 0 or type 1 depending on the state of the process immediately before the arrival.

To compute the asymptotic rate of type 0 served customers, consider the renewal process whose arrivals are the times of jumps to state 0. Then in each cycle exactly one type 0 customer is served. Using (6.58), the asymptotic rate is

$$\frac{1}{\text{mean cycle length}} = \frac{1}{E_0[T_0]} = \frac{1}{\bigl(\frac{1}{\pi(0)\lambda(0)}\bigr)} = \pi(0)\lambda.$$

Exactly one type 1 customer is served between two consecutive visits to state 2. Use the renewal process from above of jumps to state 2 to capture the asymptotic rate of type 1 served customers:

$$\frac{1}{\text{mean cycle length}} = \frac{1}{E_2[T_2]} = \frac{1}{\bigl(\frac{1}{\pi(2)\lambda(2)}\bigr)} = \pi(2)\mu = \pi(1)\lambda,$$

where the last equality is from the detailed balance condition (6.62).

Hence the total rate of customers served is

$$\pi(0)\lambda + \pi(1)\lambda = \lambda(\pi(0) + \pi(1)),$$

and the rate of customers turned away is

$$\lambda - \lambda(\pi(0) + \pi(1)) = \lambda\pi(2),$$

as we deduced previously. △

Example 6.44 (Back to Example 6.16)**.** We go back to Example 6.16 and modify the process so that it jumps from state 0 to state 1 with rate 1. Then $\lambda_0 = 1$ and for $k \ge 1$, $\lambda_k = k\lambda$ and $\mu_k = k\mu$. Detailed balance equations yield

$$\pi(k) = \pi(0)\prod_{j=0}^{k-1}\frac{\lambda_j}{\mu_{j+1}} = \pi(0)\frac{1}{\mu}\prod_{j=1}^{k-1}\frac{j\lambda}{(j+1)\mu} = \pi(0)\frac{1}{\lambda}(\lambda/\mu)^k\frac{1}{k}.$$

The sum $1 + \sum_{k=1}^{\infty}\frac{1}{\lambda}\frac{(\lambda/\mu)^k}{k}$ is finite if and only if $\mu > \lambda$. △

Example 6.45 (Based on Problem 4.11 in Durrett)**.** A computer lab has three laser printers in use. A working printer will function for an exponential amount of time with parameter μ. Upon failure it is immediately sent to the repair shop. In the shop machines are fixed by two repairmen, both of whom repair one printer in an exponential amount of time with parameter λ. Only one repairman at a time works on one printer. What is the asymptotic fraction of time when both repairmen are busy? What is the average number of machines in use in the long run?

Let X_t denote the number of working machines. The states of this Markov chain are $\{0,1,2,3\}$, and the state X_t makes only jumps of size ± 1 (either a functioning printer breaks down, or one of the printers is fixed). Hence this is a birth and death chain. The 'birth' rates are

$$\lambda_0 = q(0,1) = 2\lambda, \qquad \lambda_1 = q(1,2) = 2\lambda, \qquad \lambda_2 = q(2,3) = \lambda$$

(if we are at 0 or 1 then both repairmen are working, but in 2 only one of them). The 'death' rates are proportional to the number of working printers:

$$\mu_1 = q(1,0) = \mu, \qquad \mu_2 = q(2,1) = 2\mu, \qquad \mu_3 = q(3,2) = 3\mu.$$

There is a stationary reversible distribution satisfying

$$\begin{aligned}
\pi(1) &= \pi(0)\frac{\lambda_0}{\mu_1} = \pi(0)\frac{2\lambda}{\mu},\\
\pi(2) &= \pi(0)\frac{\lambda_0\lambda_1}{\mu_1\mu_2} = \pi(0)\frac{4\lambda^2}{2\mu^2} = \pi(0)\frac{2\lambda^2}{\mu^2}\\
\pi(3) &= \pi(0)\frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} = \pi(0)\frac{4\lambda^3}{6\mu^3} = \pi(0)\frac{2\lambda^3}{3\mu^3}.
\end{aligned}$$

Then

$$1 = \sum_{k=0}^{3} \pi(k) = \pi(0)\Big(1 + \frac{2\lambda}{\mu} + \frac{2\lambda^2}{\mu^2} + \frac{2\lambda^3}{3\mu^3}\Big)$$

and the equations above give

$$\pi(0) = \frac{3\mu^3}{2\lambda^3 + 6\lambda^2\mu + 6\lambda\mu^2 + 3\mu^3}, \qquad \pi(1) = \frac{6\lambda\mu^2}{2\lambda^3 + 6\lambda^2\mu + 6\lambda\mu^2 + 3\mu^3},$$
$$\pi(2) = \frac{6\lambda^2\mu}{2\lambda^3 + 6\lambda^2\mu + 6\lambda\mu^2 + 3\mu^3} \qquad \pi(3) = \frac{2\lambda^3}{2\lambda^3 + 6\lambda^2\mu + 6\lambda\mu^2 + 3\mu^3}.$$

Both repairmen are busy in states 0 and 1, and hence the long term fraction of time for this is $\pi(0) + \pi(1)$.

The average number of functioning machines in the long run is the time average of the function $f(k) = k$. By the SLLN, this is the average of f with respect to π:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t X_s\,ds = \sum_{k=0}^{3} k\pi(k) = \frac{6\lambda\mu^2 + 12\lambda^2\mu + 6\lambda^3}{2\lambda^3 + 6\lambda^2\mu + 6\lambda\mu^2 + 3\mu^3}.$$

△

Exercises

Theoretical exercises.

Exercise 6.1. Prove identity (6.3).

Hint. Make the necessary changes in the proof of Theorem 2.12.

Exercise 6.2. Let $\mathbf{P}_t = \{p_t(x,y)\}_{x,y\in\mathcal{S}}$ be the function defined in (6.7).

(a) Show that $\mathbf{P}_t$ satisfies the Definition 6.1 of a transition probability function.

(b) Show that $\mathbf{P}_t$ satisfies the Chapman-Kolmogorov equations.

Exercise 6.3. Solve the invariance equation $\pi Q = 0$ for the M/M/1 queue of Example 6.10. **Hint.** An invariant distirbution will not exist for all parameter values λ, μ.

Exercise 6.4. Consider a continuous-time symmetric random walk X_t on the integers, with routing matrix $r(x, x \pm 1) = \frac{1}{2}$ and holding parameters $\lambda(x)$ for $x \in \mathbb{Z}$. Fix a state z and a positive real number c. Show that the holding parameters can be chosen so that $E_z[T_z] = c$, that is, so that c equals the mean time to return to z after starting at z and moving away.

Calculus and analysis

A.1. Series

Double sums $\sum_i \sum_j a_{i,j}$ can often be evaluated better one way than the other, and hence it is convenient to be able to switch around the summation symbols. This is not always possible, as illustrated by the following example.

Example A.1. Define the doubly-indexed array of numbers $\{a_{i,j}\}_{i,j\geq 0}$ by

$$a_{i,i} = 1, \quad a_{i,i+1} = -1 \quad \text{and} \quad a_{i,j} = 0 \ \text{ for } j \geq i+2.$$

Then

$$\sum_{i=0}^{\infty}\Bigl(\sum_{j=0}^{\infty} a_{i,j}\Bigr) = \sum_{i=0}^{\infty}(a_{i,i} + a_{i,i+1}) = \sum_{i=0}^{\infty} 0 = 0$$

while

$$\sum_{j=0}^{\infty}\Bigl(\sum_{i=0}^{\infty} a_{i,j}\Bigr) = a_{0,0} + \sum_{j=1}^{\infty}(a_{j,j} + a_{j-1,j}) = 1 + \sum_{j=1}^{\infty} 0 = 1.$$

△

The next lemma gives sufficient conditions for the switch. This lemma is a special case of a much more general theorem, named after Italian mathematicians *Fubini* and *Tonelli*, that one typically encounters in graduate analysis.

Lemma A.2. *Let $\{a_{i,j}\}$ be a doubly-indexed array of extended real numbers, that is, each $a_{i,j}$ is either a real number or $\pm\infty$. Then*

$$\sum_i \sum_j a_{i,j} = \sum_j \sum_i a_{i,j} \tag{A.1}$$

is true under any one of these four conditions:

(a) $a_{i,j} \in [0,\infty]$ *for all i, j.*

(b) $a_{i,j} \in [-\infty, 0]$ *for all i, j.*

(c) *The absolute values are summable:*

$$\sum_i \sum_j |a_{i,j}| < \infty.$$

(d) *The index i takes only finitely many values, and for each i the series $\sum_j a_{i,j}$ is convergent to a real value.*

Remark A.3 (Switching order of summation with indicators)**.** The switch in (A.1) is simple because the summation limits in the inner sum do not depend on the index of the outer sum. Here is a version where such dependence is present.

$$\sum_{i=1}^{\infty} \sum_{j=1}^{i} a_{i,j} = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} a_{i,j}. \tag{A.2}$$

Identity (A.2) can be justified with the help of indicators. The notation for an indicator is

$$I_A = \begin{cases} 1, & \text{if condition } A \text{ is true} \\ 0, & \text{if condition } A \text{ is false.} \end{cases}$$

Suppose $a_{i,j} \ge 0$ so that the interchange is valid.

$$\sum_{i=1}^{\infty} \sum_{j=1}^{i} a_{i,j} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} I_{\{i \ge j\}}\, a_{i,j} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} I_{\{i \ge j\}}\, a_{i,j} = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} a_{i,j}.$$

In the middle step above the switch of $\sum_{i=1}^{\infty}$ and $\sum_{j=1}^{\infty}$ is valid because the summation limits of the inner sum no longer depend on the index of the outer sum. △

Theorem A.4 (Monotone convergence theorem)**.** *Let $\{a_{i,n} : i, n \ge 0\}$ be a doubly infinite array of numbers in $[0, \infty]$ such that, for each $i \ge 1$, the sequence $\{a_{i,n}\}$ is monotone nondecreasing: $a_{i,1} \le a_{i,2} \le \cdots \le a_{i,n} \le \cdots$. Define $b_i = \lim_{n\to\infty} a_{i,n}$, a limit in $[0, \infty]$. Then we have convergence of infinite series:*

$$\lim_{n\to\infty} \sum_{i=1}^{\infty} a_{i,n} = \sum_{i=1}^{\infty} b_i.$$

Proof. The sequence $A_n = \sum_{i=1}^{\infty} a_{i,n}$ is monotone nondecreasing, and hence its limit $A = \lim_{n\to\infty} A_n$ exists in $[0, \infty]$. The term by term inequalities $a_{i,n} \le b_i$ imply that each $A_n \le B = \sum_{i=1}^{\infty} b_i$. Consequently also $A \le B$. It remains to show $A \ge B$.

Suppose first that $b_j = \infty$ for some j. Then $\lim_{n\to\infty} a_{j,n} = \infty$. By dropping all the other terms from the series we get this lower bound on A:

$$A = \lim_{n\to\infty} \sum_{i=1}^{\infty} a_{i,n} \ge \lim_{n\to\infty} a_{j,n} = \infty.$$

Thus A is also infinite, and in particular then $A = B = \infty$.

The remaining case is the one where all $b_i < \infty$. (However, $B = \infty$ is still possible.) Let c be any (finite) real number such that $c < B$. Pick m so that

$c < \sum_{i=1}^{m} b_i$. Since a finite sum of limits is the limit of the sum,

$$c < \sum_{i=1}^{m} b_i = \sum_{i=1}^{m} \lim_{n\to\infty} a_{i,n} = \lim_{n\to\infty} \sum_{i=1}^{m} a_{i,n} \le \lim_{n\to\infty} \sum_{i=1}^{\infty} a_{i,n} = A.$$

The conclusion is that for any number $c < B$, we also have $A \ge c$. This last inequality persists as we let c tend to B, even if $B = \infty$, and gives us $A \ge B$. The proof is complete. □

A.2. Differential equations

For use in Chapter 6 we state the basic existence and uniqueness theorem for ordinary differential equations in Euclidean space. The setting is the following. Fix a dimension $d \in \mathbb{Z}_{>0}$. Let $|\mathbf{x}| = (x_1^2 + \cdots + x_d^2)^{1/2}$ denote the Euclidean norm of vectors $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. The velocity function $f : \mathbb{R}^d \to \mathbb{R}^d$ is assumed to be Lipschitz continuous: there is a constant C such that

$$|f(\mathbf{y}) - f(\mathbf{x})| \le C|\mathbf{y} - \mathbf{x}| \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A.3}$$

The goal is to solve the following ordinary differential equation (ODE) for a given time point $t_0 \in \mathbb{R}$ and spatial point $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{x}'(t) = f(\mathbf{x}(t)), \quad \mathbf{x}(t_0) = \mathbf{z}. \tag{A.4}$$

A solution to this equation is an $\mathbb{R}^d$-valued function $\mathbf{x}(t)$ that is (i) defined on some open interval (a, b) of time points that contains t_0 and that (ii) satisfies (A.4), the first part at each $t \in (a, b)$. Uniqueness of solution means that there is only one such function on (a, b).

Theorem A.5 (Existence and uniqueness for ODEs)**.** *Assume $f : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the Lipschitz condition (A.3). Then there exists a continuously differentiable solution $\mathbf{x}(t)$ to equation (A.4) defined for all time $t \in \mathbb{R}$. On any open interval (a, b) that contains t_0 there is only one solution function.*

Appendix B

Probability

B.1. Probability measures and random variables

We prove the subadditivity that was claimed in equation (1.4). The term *sub*additivity refers to the feature that the probability of the union is *less* than or equal to the sum of probabilities.

Lemma B.1. *For any sequence of events* $\{A_k\}_{1\le k<\infty}$,

$$P\Bigl(\bigcup_{k=1}^{\infty} A_k\Bigr) \le \sum_{k=1}^{\infty} P(A_k). \tag{B.1}$$

Proof. The gist of the proof is to introduce pairwise disjoint events $\{B_k\}_{1\le k<\infty}$ such that $B_k \subset A_k$ and $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$, and then use countable additivity of probability. Set $B_1 = A_1$, and then successively

$$B_k = A_k \setminus (A_1 \cup \dots \cup A_{k-1}) \quad \text{for } k \ge 2.$$

By the definition it is clear that each $B_k \subset A_k$. Let $j < k$. Then $B_j \subset A_j$ while B_k is disjoint from A_j. Hence $B_j \cap B_k = \varnothing$. We have shown that events $\{B_k\}_{1\le k<\infty}$ are pairwise disjoint.

From $B_k \subset A_k$ it follows that $\bigcup_{k=1}^{\infty} B_k \subset \bigcup_{k=1}^{\infty} A_k$. To get the opposite inclusion, let $\omega \in \bigcup_{k=1}^{\infty} A_k$. We can identify the *first index* k such that $\omega \in A_k$. This means that $\omega \notin A_1 \cup \dots \cup A_{k-1}$, and consequently $\omega \in B_k$. We have shown that every point of $\bigcup_{k=1}^{\infty} A_k$ lies also in some B_k, and hence $\bigcup_{k=1}^{\infty} A_k \subset \bigcup_{k=1}^{\infty} B_k$. The two inclusions together give the equality $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$. Now the final calculation. The pairwise disjointness of the events B_k justifies the second equality.

$$P\Bigl(\bigcup_{k=1}^{\infty} A_k\Bigr) = P\Bigl(\bigcup_{k=1}^{\infty} B_k\Bigr) = \sum_{k=1}^{\infty} P(B_k) \le \sum_{k=1}^{\infty} P(A_k). \qquad \square$$

Random variables with infinite values arise naturally in the subject of stochastic processes.

Lemma B.2. *Let X be a random variable whose values lie in the set $\mathbb{Z}_{\geq 0} \cup \{\infty\}$. Then*

$$\lim_{n\to\infty} P(X \geq n) = P(X = \infty).$$

In particular, if X is finite with probability one, that is, $P(X < \infty) = 1$, then

$$\lim_{n\to\infty} P(X \geq n) = 0.$$

Proof. Decompose the probability according to all the distinct values of X:

$$P(X \geq n) = \sum_{k:\, n \leq k < \infty} P(X = k) \;+\; P(X = \infty). \tag{B.2}$$

A fact from calculus is that a series $\sum_{1 \leq i < \infty} a_i$ converges if and only if its tail converges to zero:

$$\lim_{n\to\infty} \biggl| \sum_{n \leq i < \infty} a_i \biggr| = 0.$$

Since the terms $P(X = k)$ are probabilities of mutually disjoint events, axioms of probability give the bounds $0 \leq \sum_k P(X = k) \leq 1$ that imply convergence of the series. Thus as $n \to \infty$ the first series on the right-hand side of (B.2) converges to zero. □

B.2. Expectations

We go over properties of expectations that are used frequently.

The *indicator random variable of an event* A is by definition

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

Despite its simple-minded definition, it is surprisingly useful. The expectation of the indicator of A is the probability of the event A:

$$E[I_A] = 0 \cdot P(I_A = 0) + 1 \cdot P(I_A = 1) = P(A).$$

Thus every probability of an event is actually a special case of an expectation.

There is a frequent need to calculate expectations of nonnegative integer-valued random variables may take the value ∞. In that case the familiar formula for the expectation of a discrete random variable is extended in a natural way: if X is $\mathbb{Z}_{\geq 0} \cup \{\infty\}$-valued, then

$$EX = \sum_{0 \leq n < \infty} n \cdot P(X = n) \;+\; \infty \cdot P(X = \infty). \tag{B.3}$$

The algebraic conventions regarding ∞ go as follows, where a is any real number and c a strictly positive real number:

$$a + \infty = \infty, \quad \infty + \infty = \infty, \quad c \cdot \infty = \infty, \quad 0 \cdot \infty = 0, \quad \frac{a}{\infty} = 0. \tag{B.4}$$

The important point is that $\infty - \infty$ is *not defined* and cannot be used.

Occasionally the need arises to determine whether a $[0, \infty]$-valued random variable takes the value ∞. A negative answer can be established by checking that the expectation is finite, as stated in the next lemma.

Lemma B.3. *Let X be a $[0,\infty]$-valued random variable. Then $EX < \infty$ implies that $P(X < \infty) = 1$.*

Proof. The expectation is monotone: that is, if $X \ge Y$ then $EX \ge EY$. Apply this to X and $\infty \cdot I_{\{X=\infty\}}$. Since $X = \infty$ on the event $\{X = \infty\}$ and $X \ge 0$ on the complement $\{X = \infty\}^c$, we have the inequality $X \ge \infty \cdot I_{\{X=\infty\}}$. Hence the expectations satisfy

$$EX \ge E\big[\infty \cdot I_{\{X=\infty\}}\big] = \infty \cdot P(X = \infty).$$

From (B.4) we see that if $P(X = \infty) > 0$ then $\infty \cdot P(X = \infty) = \infty$. Thus $P(X = \infty) > 0$ forces $EX = \infty$. The equivalent contrapositive is that $EX < \infty$ forces $P(X = \infty) = 0$. □

Here is a formula that helps calculate expectations.

Lemma B.4. *Let X be a random variable whose values lie in the set $\mathbb{Z}_{\ge 0} \cup \{\infty\}$. That is, X is nonnegative integer valued and it can also take on the value infinity. Then*

$$E(X) = \sum_{1 \le k < \infty} P(X \ge k). \tag{B.5}$$

The summation limits in (B.5) are expressed as $\sum_{1\le k<\infty}$ to emphasize that the sum ranges over finite k and that a term for $k = \infty$ is not part of the sum. This is for emphasis only. The conventional notation $\sum_{k=1}^{\infty}$ does have the same meaning.

Proof. Suppose first that $P(X = \infty) > 0$. Then

$$EX = \sum_{1\le k<\infty} k\, P(X = k) + \infty \cdot P(X = \infty) \ge \infty \cdot P(X = \infty) = \infty.$$

Also $P(X \ge k) \ge P(X = \infty) > 0$ and so

$$\sum_{1\le k<\infty} P(X \ge k) \ge \sum_{1\le k<\infty} P(X = \infty) = \infty.$$

Thus (B.5) holds.

Now suppose that $P(X = \infty) = 0$. Then

$$\begin{aligned} EX &= \sum_{j=1}^{\infty} jP(X = j) = \sum_{j=1}^{\infty}\sum_{k=1}^{j} P(X = j) = \sum_{k=1}^{\infty}\sum_{j=k}^{\infty} P(X = j) \\ &= \sum_{k=1}^{\infty} P(X \ge k) \end{aligned} \tag{B.6}$$

and again (B.5) holds. □

The interchange in the order of summation used in (B.6) above is discussed in Remark A.3.

Next we state two useful versions of the monotone convergence theorem for expectations.

Theorem B.5.

(a) *Let* $\{X_n\}_{n\geq 1}$ *be a monotone non-decreasing sequence of* $[0,\infty]$*-valued random variables, in other words*

$$0 \leq X_1(\omega) \leq X_2(\omega) \leq X_3(\omega) \leq \cdots$$

Denote the $[0,\infty]$*-valued limiting random variable by* $X(\omega) = \lim_{n\to\infty} X_n(\omega)$*. Then the expectations converge:* $E[X] = \lim_{n\to\infty} E[X_n]$.

(b) *Let* $\{Y_k\}_{k\geq 1}$ *be a sequence of* $[0,\infty]$*-valued random variables. Then*

$$E\biggl[\,\sum_{k=1}^{\infty} Y_k\biggr] = \sum_{k=1}^{\infty} E[Y_k].$$

Proof of part (a) can be found in graduate texts on real analysis or probability. Part (b) is a corollary of part (a) by taking $X_n = \sum_{k=1}^{n} Y_k$ so that $X = \sum_{k=1}^{\infty} Y_k$.

Appendix C

Linear algebra

The *inverse* of an $n \times n$ matrix A is an $n \times n$ matrix A^{-1} that satisfies

$$AA^{-1} = A^{-1}A = I.$$

There is only one inverse: for if B also satisfies $AB = BA = I$, then

$$B = BI = B(AA^{-1}) = (BA)A^{-1} = IA^{-1} = A^{-1}.$$

The definition of an inverse has two equations, but in fact it is enough to check one of them.

Lemma C.1. *Let A and B be $n \times n$ matrices such that $AB = I$. Then $B = A^{-1}$.*

Proof. $AB = I$ implies that $ABx = x$ for all vectors $x \in \mathbb{R}^n$. Hence as a linear transformation on $\mathbb{R}^n$, A is onto. This implies that it has an inverse A^{-1}. Then we have

$$B = IB = (A^{-1}A)B = A^{-1}(AB) = A^{-1}I = A^{-1} \qquad \square$$

Bibliography

[ASV18] David F. Anderson, Timo Seppäläinen, and Benedek Valkó, Introduction to Probability, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 2018. MR 3753683

[Dur19] Rick Durrett, Probability: theory and examples, vol. 49, Cambridge university press, 2019.

[KS66] Harry Kesten and Bernt P Stigum, A limit theorem for multidimensional galton-watson processes, The Annals of Mathematical Statistics **37** (1966), no. 5, 1211–1223.

[Res92] Sidney Resnick, Adventures in stochastic processes, Birkhäuser Boston, Inc., Boston, MA, 1992. MR 1181423