Probability and Its Applications

Published in association with the Applied Probability Trust

Editors: J. Gani, C.C. Heyde, P. Jagers, T.G. Kurtz

Probability and Its Applications

Azencott et al.: Series of Irregular Observations. Forecasting and Model Building. 1986
Bass: Diffusions and Elliptic Operators. 1997
Bass: Probabilistic Techniques in Analysis. 1995
Berglund/Gentz: Noise-Induced Phenomena in Slow-Fast Dynamical Systems: A Sample-Paths Approach. 2006
Biagini/Hu/Øksendal/Zhang: Stochastic Calculus for Fractional Brownian Motion and Applications. 2008
Chen: Eigenvalues, Inequalities and Ergodic Theory. 2005
Costa/Fragoso/Marques: Discrete-Time Markov Jump Linear Systems. 2005
Daley/Vere-Jones: An Introduction to the Theory of Point Processes I: Elementary Theory and Methods. 2nd ed. 2003, corr. 2nd printing 2005
Daley/Vere-Jones: An Introduction to the Theory of Point Processes II: General Theory and Structure. 2nd ed. 2008
de la Peña/Gine: Decoupling: From Dependence to Independence, Randomly Stopped Processes U-Statistics and Processes Martingales and Beyond. 1999
Del Moral: Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications. 2004
Durrett: Probability Models for DNA Sequence Evolution. 2002, 2nd ed. 2008
Galambos/Simonelli: Bonferroni-Type Inequalities with Equations. 1996
Gani (ed.): The Craft of Probabilistic Modelling. A Collection of Personal Accounts. 1986
Gut: Stopped Random Walks. Limit Theorems and Applications. 1987
Guyon: Random Fields on a Network. Modeling, Statistics and Applications. 1995
Kallenberg: Foundations of Modern Probability. 1997, 2nd ed. 2002
Kallenberg: Probabilistic Symmetries and Invariance Principles. 2005
Lai/de la Peña/Shao: Self-Normalized Processes. Limit Theory and Statistical Applications. 2009
Last/Brandt: Marked Point Processes on the Real Line. 1995
Molchanov: Theory of Random Sets. 2005
Nualart: The Malliavin Calculus and Related Topics, 1995, 2nd ed. 2006
Rachev/Rueschendorf: Mass Transportation Problems. Volume I: Theory and Volume II: Applications. 1998
Resnick: Extreme Values, Regular Variation and Point Processes. 1987
Schmidli: Stochastic Control in Insurance. 2008
Schneider/Weil: Stochastic and Integral Geometry. 2008
Serfozo: Basics of Applied Stochastic Processes. 2009
Shedler: Regeneration and Networks of Queues. 1986
Silvestrov: Limit Theorems for Randomly Stopped Stochastic Processes. 2004
Thorisson: Coupling, Stationarity and Regeneration. 2000

Richard Serfozo

Basics of Applied Stochastic Processes

Springer

Richard Serfozo
Georgia Institute of Technology
School of Industrial & Systems Engineering
765 Ferst Drive NW.,
Atlanta GA 30332-0205
USA
rserfozo@isye.gatech.edu

Series Editors:

Joe Gani
Chris Heyde
Centre for Mathematics and its Applications
Mathematical Sciences Institute
Australian National University
Canberra, ACT 0200
Australia
gani@maths.anu.edu.au

Thomas G. Kurtz
Department of Mathematics
University of Wisconsin - Madison
480 Lincoln Drive
Madison, WI 53706-1388
USA
kurtz@math.wisc.edu

Peter Jagers
Mathematical Statistics
Chalmers University of Technology
and Göteborg (Gothenburg) University
412 96 Göteborg
Sweden
Jagers@chalmers.se

ISBN: 978-3-540-89331-8 e-ISBN: 978-3-540-89332-5
DOI: 10.1007/978-3-540-89332-5

Probability and Its Applications ISSN print edition: 1431-7028

Library of Congress Control Number: 2008939432

Mathematics Subject Classification (2000): 60-02, 60-J10, 60-J27, 60-K05, 60-J25

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

I dedicate this book to

Joan, Kip, Shannyn, Jillian and Max Serfozo

Preface

The phrase "Applied Stochastic Processes" refers to stochastic processes that are commonly used as mathematical models of random phenomena that evolve over time or space. Since randomness is ubiquitous in our universe, and maybe beyond, the application areas have been very diverse. Here are some examples.

Telecommunications: Sizing networks, antenna coverage, traffic control, alternate routing, and voice recognition.

Computers: Network design, parallel processing, artificial intelligence, pattern recognition, and performance optimization.

Manufacturing: Forecasting, planning, scheduling, facility location, and resource management.

Finance: Portfolios, option pricing, pension funds, and forecasting.

Insurance: Risk analysis, demographics, investments, and diversification.

Internet: Design, control, optimal searching, parallel processing, advertising, and pattern recognition.

Call-centers: Forecasting, staffing, alternate routing, and optimal design.

Airlines: Scheduling, maintenance, ticket pricing, and overbooking.

Supply chains: Network design, inventory control, transhipping, alternate sources, and contracting.

Military: Logistics, scheduling, maintenance, targeting, intelligence, purchasing, and war games.

Infrastructure: Reliability and maintenance of roads, buildings, bridges, dams, levees, and utilities.

Airports: Traffic control, emergencies, security, and runway design.

Inventory control: Retail and rental items, blood, oil, water, and food.

Security: Computers, homeland, banks, phones, and data files.

Medicine: DNA sequencing, diagnoses, epidemics, and vaccines.

Energy: Planning, control, sharing, storage, and disasters.

Other major applications have been in academic disciplines (e.g., Statistics, Mathematics, Engineering, Physics, Biology, Social Sciences and Business), and in subjects related to government (e.g., NASA, NIH and NIST).

The documentation of stochastic applications is in company and government technical reports, academic conference proceedings, and journals. Journals that publish research on applied stochastic processes include *Advances in Applied Probability*, *Annals of Applied Probability*, *Journal of Applied Probability*, *Probability in the Engineering and Informational Sciences*, *Queueing Systems: Theory and Applications*, and *Stochastic Processes and their Applications*.

The focus of this book is on the principal stochastic processes used in applications that are as follows. This list corresponds to the chapter titles.

1. Markov Chains in Discrete Time
2. Renewal and Regenerative Processes
3. Poisson Processes
4. Continuous-time Markov Chains
5. Brownian Motion.

The book describes basic properties of these stochastic processes and illustrates how to use the processes to model systems and solve problems. The presentation is at an introductory level for readers familiar with random variables, distribution functions, manipulations with expectations, and elementary real analysis. Knowledge of stochastic processes or measure theory is not required. A review of conditional probabilities is in the first chapter, and additional background material on probability and real analysis is summarized in the appendix.

The book has two aims. One aim is to present theorems and examples of applied stochastic processes as in most introductory textbooks. So the book would be suitable for one or two courses on applied stochastic processes.

The second aim is to go beyond an introduction and provide a comprehensive description of the processes in the first four chapters mentioned above, and a considerable coverage of Brownian motion (not including stochastic integration). In this regard, the book emphasizes the following.

• Careful and complete proofs that illustrate stochastic reasoning and the algebra and calculus of probabilities and expectations.

• The use of point processes as a vehicle to represent special transition times in Markov chains, space-time Poisson processes, Brownian/Poisson particle systems, and regeneration times in complex systems.

• Techniques for constructing or formulating processes (e.g., clock times, sample-process representations for Poisson processes, marking and transforming of processes, and subordination of processes).

• Mathematical tools and techniques for stochastic analysis including Laplace functionals and Palm probabilities for point processes, coupling, Lévy

formulas for functionals of Markov chains, martingales, stopping times, functional central limit theorems, and convergence concepts.

• Poisson processes in space as well as time, marked Poisson processes, and Poisson limits of sparse point processes.

• Regenerative phenomena (e.g., crude regenerations in key renewal theorem, and regenerate-increment processes as a framework for various strong laws of large numbers and central limit theorems).

A major theme of the book, and of applied stochastic processes in general, is the establishment of limiting distributions and averages for quantities of interest. Accordingly, there is an extensive coverage of characterizations of limiting distributions of the principal processes, strong laws of large numbers for evaluating limiting averages for the processes, central limit theorems that describe deviations of the averages, limit theorems for approximating sparse point processes by Poisson processes, and functional central limit theorems for approximating various processes by functions of Brownian motion.

Each chapter contains numerous examples and exercises that illustrate applications or extensions of the theorems. Several sections are devoted to stochastic networks, queueing systems, branching populations, reversible processes, Markov chain Monte Carlo models, compound Poisson processes, Gaussian processes, and Brownian bridge.

Important topics in applied stochastic processes that the book does not cover include diffusion processes, stationary processes, stochastic integrals and differential equations, interacting particle systems, simulation, Gibbs fields, finance models, large deviations, and stochastic control (Markov decision models). Most of these topics are in more advanced texts, and the rest are broad enough to be subjects of specialized monographs.

I will close with a few acknowledgements. First, I am grateful to William Feller for writing his 1950 book *Introduction to Probability and its Applications*. It opened my eyes to the notion that "One can make sense out of the nonsense of randomness", which sparked my interest in probability. I cannot give enough thanks to Erhan Cinlar, my Ph.D. advisor, who has been a kind friend as well as a mentor. I am also very appreciative to those who have developed the knowledge of stochastic processes — I made extensive use of their works in writing the book, especially the work of Olav Kallenberg 2004.

My loving wife Joan contributed to the clarity of the exposition by advising me to "reach the reader" by adopting a writing style that is not overly terse and easy to read. My colleague Steve Hackman prodded me along similar lines on editorial issues. Careful readings by Brian Fralix, Anton Kleywegt, Evsey Morozov, Christian Rau, and Georgia Tech students were very helpful in catching many typos and errors. I thank all of you for helping me on this project.

Richard Serfozo

Contents

Chapter 1
Markov Chains

A sequence of random variables $X_0, X_1, \ldots$ with values in a countable set S is a Markov chain if at any time n, the future states (or values) $X_{n+1}, X_{n+2}, \ldots$ depend on the history $X_0, \ldots, X_n$ only through the present state X_n. Markov chains are fundamental stochastic processes that have many diverse applications. This is because a Markov chain represents any dynamical system whose states satisfy the recursion $X_n = f(X_{n-1}, Y_n)$, $n \geq 1$, where $Y_1, Y_2 \ldots$ are independent and identically distributed (i.i.d.) and f is a deterministic function. That is, the new state X_n is simply a function of the last state and an auxiliary random variable. Such system dynamics are typical of those for queue lengths in call centers, stresses on materials, waiting times in production and service facilities, inventories in supply chains, parallel-processing software, water levels in dams, insurance funds, stock prices, etc.

This chapter begins by describing the basic structure of a Markov chain and how its single-step transition probabilities determine its evolution. For instance, what is the probability of reaching a certain state, and how long does it take to reach it? The next and main part of the chapter characterizes the stationary or equilibrium distribution of Markov chains. These distributions are the basis of limiting averages of various cost and performance parameters associated with Markov chains. Considerable discussion is devoted to branching phenomena, stochastic networks, and time-reversible chains. Included are examples of Markov chains that represent queueing, production systems, inventory control, reliability, and Monte Carlo simulations.

Before getting into the main text, a reader would benefit by a brief review of conditional probabilities in Section 1.22 of this chapter and related material on random variables and distributions in Sections 1–4 in the Appendix. The rest of the Appendix, which provides more background on probability, would be appropriate for later reading.

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

1.1 Introduction

This section introduces Markov chains and describes a few examples.

A discrete-time *stochastic process* $\{X_n : n \geq 0\}$ on a countable set S is a collection of S-valued random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. The P is a probability measure on a family of events $\mathcal{F}$ (a σ-field) in an event-space Ω.[1] The set S is the *state space* of the process, and the value $X_n \in S$ is the *state* of the process at *time* n. The n may represent a parameter other than time such as a length or a job number.

The *finite-dimensional* distributions of the process are

$$P\{X_0 = i_0, \ldots, X_n = i_n\}, \qquad i_0, \ldots, i_n \in S,\ n \geq 0.$$

These probabilities uniquely determine the probabilities of all events of the process. Consequently, two stochastic processes (defined on different probability spaces or the same one) are equal in distribution if their finite-dimensional distributions are equal. Various types of stochastic processes are defined by specifying the dependency among the variables that determine the finite-dimensional distributions, or by specifying the manner in which the process evolves over time (the system dynamics).

A Markov chain is defined as follows.

Definition 1. A stochastic process $X = \{X_n : n \geq 0\}$ on a countable set S is a *Markov Chain* if, for any $i, j \in S$ and $n \geq 0$,

$$P\{X_{n+1} = j | X_0, \ldots, X_n\} = P\{X_{n+1} = j | X_n\}, \tag{1.1}$$

$$P\{X_{n+1} = j | X_n = i\} = p_{ij}. \tag{1.2}$$

The p_{ij} is the probability that the Markov chain jumps from state i to state j. These *transition probabilities* satisfy $\sum_{j \in S} p_{ij} = 1$, $i \in S$, and the matrix $\mathsf{P} = (p_{ij})$ is the *transition matrix* of the chain.

Condition (1.1), called the *Markov property*, says that, at any time n, the next state X_{n+1} is conditionally independent of the past $X_0, \ldots, X_{n-1}$ given the present state X_n. In other words, the next state is dependent on the past and present only through the present state. The Markov property is an elementary condition that is satisfied by the state of many stochastic phenomena. Consequently, Markov chains, and related continuous-time Markov processes, are natural models or building blocks for applications.

Condition (1.2) simply says the transition probabilities do not depend on the time parameter n; the Markov chain is therefore "time-homogeneous". If the transition probabilities were functions of time, the process X_n would be a non-time-homogeneous Markov chain. Such chains are like time-homogeneous

[1] Further details on probability spaces are in the Appendix. We follow the convention of not displaying the space $(\Omega, \mathcal{F}, P)$ every time random variables or processes are introduced; it is mentioned only when needed for clarity.

chains, but the time dependency introduces added accounting details that we will not address here. See Exercises 12 and 13 for further insights.

Since the state space S is countable, we will sometimes label the states by integers, such as $S = \{0, 1, 2, \ldots\}$ (or $S = \{1, \ldots, m\}$). Under this labeling, the transition matrix has the form

$$\mathsf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

We end this section with a few preliminary examples.

Example 2. Binomial Markov Chain. A *Bernoulli process* is a sequence of independent trials in which each trial results in a success or failure with respective probabilities p and $q = 1-p$. Let X_n denote the number of successes in n trials, for $n \geq 1$. By direct reasoning, it follows that X_n has a binomial distribution with parameters n and p:

$$P\{X_n = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 \leq k \leq n.$$

Now, suppose at the nth trial that $X_n = i$. Then at the next trial, X_{n+1} will equal $i + 1$ or i with probabilities p and $1 - p$, respectively, regardless of the values of $X_1, \ldots, X_{n-1}$. Thus X_n is a Markov chain with transition probabilities $p_{i,i+1} = p$, $p_{ii} = 1 - p$ and $p_{ij} = 0$ otherwise. This binomial Markov chain is a special case of the following random walk.

Example 3. Random Walk. Suppose $Y_1, Y_2, \ldots$ are i.i.d. integer-valued random variables, and define $X_0 = 0$ and

$$X_n = \sum_{m=1}^{n} Y_m, \qquad n \geq 1.$$

The process X_n is a *random walk* on the set of integers S, where Y_n is the step size at time n. A random walk represents a quantity that changes over time (e.g., a stock price, an inventory level, or a gambler's fortune) such that its increments (step sizes) are i.i.d. Since $X_{n+1} = X_n + Y_{n+1}$, and Y_{n+1} is independent of $X_0, \ldots, X_n$, it follows that, for any $i, j \in S$ and $n \geq 0$,

$$\begin{aligned} &P\{X_{n+1} = j | X_0, \ldots, X_{n-1}, X_n = i\} \\ &= P\{X_n + Y_{n+1} = j | X_n = i\} = P\{Y_1 = j - i\}. \end{aligned}$$

Therefore, the random walk X_n is a Markov chain on the nonnegative integers S with transition probabilities $p_{ij} = P\{Y_1 = j - i\}$.

When the step sizes Y_n take values 1 or -1 with $p = P\{Y_1 = 1\}$ and $q = P\{Y_1 = -1\}$, the chain X_n is a *simple random walk*. Its transition probabilities, for each i, are

$$p_{i,i+1} = p, \qquad p_{i,i-1} = q, \qquad p_{ij} = 0, \quad \text{for } j \neq i+1 \text{ or } i-1.$$

This type of walk restricted to a finite state space is described next.

Example 4. Gambler's Ruin. Consider a Markov chain on $S = \{0, 1, \ldots, m\}$ with transition matrix

$$\mathsf{P} = \begin{bmatrix} 1 & 0 & 0 & . & . & . & . & . \\ q & 0 & p & 0 & . & . & . & \\ 0 & q & 0 & p & 0 & . & . & \\ . & . & . & . & . & . & . & . \\ . & . & . & 0 & q & 0 & p & \\ . & . & . & . & . & . & 0 & 1 \end{bmatrix}$$

One can interpret the state of the Markov chain as the fortune of a Gambler who repeatedly plays a game in which the Gambler wins or loses \$1 with respective probabilities p and $q = 1 - p$. If the fortune reaches state 0, the Gambler is ruined since $p_{00} = 1$ (state 0 is absorbing — the chain stays there forever). On the other hand, if the fortune reaches m, the Gambler retires with the fortune m since $p_{mm} = 1$ (m is another absorbing state).

A versatile generalization to state-dependent gambles (and other applications as well) is with a transition matrix

$$\mathsf{P} = \begin{bmatrix} r_0 & p_0 & 0 & . & . & . & . & . & . & . & . & . \\ q_1 & r_1 & p_1 & 0 & . & . & . & . & . & . & . & . \\ 0 & q_2 & r_2 & p_2 & 0 & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & 0 & q_{m-1} & r_{m-1} & p_{m-1} & . & . & . & . \\ . & . & . & . & . & . & q_m & r_m & . & . & . & . \end{bmatrix}$$

In this case, the outcome of the game depends on the Gambler's fortune. When the fortune is i, the Gambler either wins or loses \$1 with respective probabilities p_i or q_i, or breaks even (the fortune does not change) with probability r_i. Another interpretation is that the state of the chain is the location of a random walk with state-dependent steps of size -1, 0, or 1.

Markov chains are common models for a variety of systems and phenomena, such as the following, in which the Markov property is "reasonable".

Example 5. Flexible Manufacturing System. Consider a machine that is capable of producing three types of parts. The state of the machine at time period n is denoted by a random variable X_n that takes values in $S = \{0, 1, 2, 3\}$, where 0 means the machine is idle and $i = 1, 2$ or 3 means the machine produces a type i in the time period. Suppose the machine's production schedule is Markovian in the sense that the next type of part it produces, or a possible

idle period, depends only on its current state, and the probabilities of these changes do not depend on time. Then X_n is a Markov chain. For instance, its transition matrix might be

$$\mathsf{P} = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 2/5 \\ 1/10 & 1/2 & 1/10 & 3/10 \\ 1/5 & 0 & 1/5 & 3/5 \\ 1/5 & 0 & 2/5 & 2/5 \end{bmatrix}$$

Such probabilities can be estimated as in Exercise 65, provided one can observe the evolution of the system. Otherwise, the probabilities can be determined by subjective reasoning or other techniques.

Note that at any period in which the machine produces a type 1 part, in the next period it produces a type 1, 2 or 3 part with respective probabilities 1/2, 1/10, and 3/10. Also, whenever the machine is idle, it remains idle in the next period with probability 1/5. Consequently, the probability the machine is idle for m periods is $(4/5)(1/5)^{m-1}$, which is a geometric distribution with parameter 4/5 and mean 1.25 periods; see Exercise 6.

1.2 Probabilities of Sample Paths

A basic issue in analyzing the structure of a stochastic process is to describe its finite-dimensional distributions. This section shows that these distributions for a Markov chain X_n are simply products of its transition probabilities and the probability distribution of the *initial state* X_0. Finite-dimensional distributions and more general properties of sample paths[2] are conveniently expressed by n-step transition probabilities, which are obtained as the nth product of the transition matrix.

Proposition 6. *Suppose X_n is a Markov chain on S with transition probabilities p_{ij} and initial distribution $\alpha_i = P\{X_0 = i\}$. Then, for any $i_0, \ldots, i_n \in S$ and $n \geq 0$,*

$$P\{X_0 = i_0, \ldots, X_n = i_n\} = \alpha_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}.$$

Proof. Proceeding by induction, this statement is obvious for $n = 0$. Now, assume it is true for some n, and let $A_n = \{X_0 = i_0, \ldots, X_n = i_n\}$. Then the statement is true for $n + 1$, since

$$P(A_{n+1}) = P(A_n) P\{X_{n+1} = x_{n+1} | A_n\} = \alpha_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n} p_{i_n,i_{n+1}},$$

[2] A *sample path* of a stochastic process X_n is a realization of it as a function of time. For instance, if $X_0(\omega) = i_0, \ldots, X_n(\omega) = i_n$, then $i_0, \ldots, i_n$ is the sample path associated with the outcome ω.

where $P\{X_{n+1} = x_{n+1}|A_n\} = p_{i_n,i_{n+1}}$ by the Markov property.

Proposition 6 says that the probability the Markov chain traverses a path $i_0, i_1, \ldots, i_n$ is just the multiplication $p_{i_0,i_1} \cdots p_{i_{n-1},i_n}$ of the probabilities of these transitions. Therefore, the probability that the Markov chain up to time n has a sample path in a subset $\mathcal{P}$ of S^{n+1} is

$$P\{(X_0, \ldots, X_n) \in \mathcal{P}\} = \sum_{(i_0,\ldots,i_n)\in\mathcal{P}} P\{X_0 = i_0\} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}. \tag{1.3}$$

For instance, if the X_n are the monthly profits of a company, then the probability its profits will increase throughout n months is

$$P\{X_0 \le X_1 \le \ldots \le X_n | X_0 = i_0\} = \sum_{i_1=i_0}^{\infty} p_{i_0,i_1} \cdots \sum_{i_n=i_{n-1}}^{\infty} p_{i_{n-1},i_n}.$$

Also, the profit in the third month has the distribution

$$P\{X_3 = j|X_0 = i_0\} = \sum_{i_1\in S} p_{i_0 i_1} \sum_{i_2\in S} p_{i_1 i_2} p_{i_2 j}.$$

Many probabilities like these for the Markov chain can be expressed conveniently in terms of the transition matrix $\mathsf{P} = (p_{ij})$ and its nth product P^n, $n \ge 0$. By definition, $\mathsf{P}^0 = I$ (the identity matrix), and $\mathsf{P}^n = \mathsf{P}^{n-1}\mathsf{P}$, for $n \ge 1$. Let p_{ij}^n denote the (i, j)th entry of P^n (so $\mathsf{P}^n = (p_{ij}^n)$).[3] Then by the definition of matrix multiplication,

$$p_{ij}^n = \sum_{i_1,\ldots,i_{n-1}\in S^{n-1}} p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j}. \tag{1.4}$$

Remark 7. n-Step Probabilities. The probability $P\{X_n = j|X_0 = i\}$ is the sum of the probabilities of all paths of the form $i, i_1, \ldots, i_{n-1}, j$, which is the sum in (1.4). Consequently,

$$P\{X_n = j|X_0 = i\} = p_{ij}^n.$$

This probability can be obtained upon computing P^n. Furthermore, denoting the initial distribution $\alpha_i = P\{X_0 = i\}$ as a row vector $\alpha = (\alpha_i)$, we have

$$P\{X_n = j\} = (\ \alpha\mathsf{P}^n)_j,$$

which is the jth value of the row-vector $\alpha\mathsf{P}^n$.

Interestingly, the multiplication property of matrices $\mathsf{P}^{m+n} = \mathsf{P}^m\mathsf{P}^n$, for $m, n \ge 1$, yields the *Chapman-Kolmogorov equations*

[3] Keep in mind that n in p_{ij}^n is not the usual multiplication operation, but P^n "is" the nth product of P.

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

This says that the probability the chain moves from i to j in $m+n$ steps is equal to the probability that it moves from i to any $k \in S$ in m steps, and then it moves from k to j in n more steps. The following examples involve n-step probabilities.

Example 8. Expected Costs or Utilities. Suppose there is a value $f_n(i) \in \mathbb{R}$ associated with the Markov chain being in state i at time n. The value could be a cost, reward or some utility parameter. Then the mean value at time n, assuming it exists, is

$$\begin{aligned} E[f_n(X_n)] &= \sum_{j \in S} P\{X_n = j\} f_n(j) \\ &= \sum_{j \in S} (\ \alpha \mathsf{P}^n)_j f_n(j) = \ \alpha \mathsf{P}^n f_n, \end{aligned}$$

where $f_n = (f_n(i))$ is a column vector of the values. Furthermore, the mean value up to time n is

$$E[\sum_{m=1}^{n} f_m(X_m)] = \sum_{m=1}^{n} \ \alpha \mathsf{P}^m f_m.$$

Example 9. Taboo Probabilities. There are many probabilities for a Markov chain involving paths that avoid a specified region in the state space. In particular, consider the *taboo probability*

$${}_A p_{ij}^n = P\{X_1, \ldots, X_{n-1} \notin A, X_n = j | X_0 = i\},$$

that the chain X_n moves from state i to state j in n steps without entering a taboo set $A \subset S$. Then as in (1.3) and Remark 7,

$${}_A p_{ij}^n = \sum_{i_1, \ldots, i_{n-1} \in A^c} p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j} = q_{ij}^n, \quad i, j \in A^c,$$

where $Q^n = (q_{ij}^n)$ is the nth product of $Q = (p_{ij}; i, j \in A^c)$ (the matrix P restricted to $A^c = S \backslash A$).

Example 10. Maxima of a Markov Chain. Suppose the Markov chain X_n has the state space $S = \{1, 2, \ldots\}$ and consider the maximum process

$$M_n = \max_{0 \le m \le n} X_m, \qquad n \ge 0.$$

For instance, if X_n is the stress on a part at time n and $X_0 = i$, then the probability the stress does not exceed a level $\ell > i$ up to time n is

$$P\{M_n \le \ell | X_0 = i\} = P\{X_1 \le \ell, \ldots, X_n \le \ell | X_0 = i\} = \sum_{j=1}^{\ell} q_{ij}^n,$$

where $Q^n = (q_{ij}^n)$ is the nth product of $Q = (p_{ij}; i, j \le \ell)$. The M_n is generally not a Markov chain, but it is when the X_n are i.i.d. (see Exercise 17).

1.3 Construction of Markov Chains

This section addresses the following questions. Is there a general framework for constructing or identifying Markov chains? Is there a Markov chain associated with any transition matrix? If so, how is it constructed? How can one simulate a Markov chain? These questions are answered by the first result that shows how to formulate a Markov chain as a function of i.i.d. random variables.

Recall that the random walk in Example 3 is "constructed" with i.i.d. random variables. That is, its evolution is represented by the recursive equation $X_n = X_{n-1} + Y_n$, $n \ge 1$, where $X_0 = 0$, and Y_n are i.i.d. random variables. Here is an analogous and more general construction of a Markov chain via a general recursive equation.

Proposition 11. *Suppose $\{X_n : n \ge 0\}$ is a stochastic process on S of the form*

$$X_n = f(X_{n-1}, Y_n), \qquad n \ge 1, \tag{1.5}$$

where $f : S \times S' \to S$ and $Y_1, Y_2, \ldots$ are i.i.d. random variables with values in a general space S' that are independent of X_0. Then X_n is a Markov chain with transition probabilities $p_{ij} = P\{f(i, Y_1) = j\}$.

Proof. The result will follow upon showing that, for any i, j and n,

$$\begin{aligned} &P\{X_{n+1} = j | X_0, \ldots, X_{n-1}, X_n = i\} \\ &\quad = P\{f(i, Y_{n+1}) = j | X_0, \ldots, X_{n-1}, X_n = i\} \\ &\quad = P\{f(i, Y_{n+1}) = j\} = p_{ij}. \end{aligned}$$

The first equality follows since $X_{n+1} = f(i, Y_{n+1})$ given $X_n = i$. The second equality is due to the fact that Y_{n+1} is independent of $(X_0, \ldots, X_n)$, because this vector, by (1.5), is a function of $(X_0, Y_1, \ldots, Y_n)$, which is independent of Y_{n+1} by assumption. The last equality follows by the definition of p_{ij} and fact that Y_{n+1} and Y_1 have the same distribution.

Proposition 11 is useful for identifying stochastic processes that are Markov chains. This approach is often easier than verifying the Markov property directly; illustrations are in the next section.

We now establish that any Markov chain can be constructed as in Proposition 11. The proof uses the following fact (Exercise 11 gives a similar result

for general random variables and discusses how the result is used to generate random samples from a distribution).

Remark 12. Uniform Representation of a Random Variable. Let α be a probability measure on $S = \{0, 1, \ldots\}$, and let U be a random variable that has a uniform distribution on $[0, 1]$. Define $X = h(U)$, where

$$h(u) = j \quad \text{if } u \in I_j, \text{ for some } j \in S, \tag{1.6}$$

and $I_j = [\sum_{k=0}^{j-1} \alpha_k, \sum_{k=0}^{j} \alpha_k)$. Then $P\{X = j\} = \alpha_j$. This follows since

$$P\{h(U) = j\} = P\{U \in I_j\} = \alpha_j.$$

Theorem 13. (Construction of Markov Chains) *Let p_{ij} be Markovian transition probabilities, and let α be a probability measure on S. Label the elements of S such that $S = \{0, 1, \ldots\}$. Suppose $U_0, U_1, \ldots$ are i.i.d. with a uniform distribution on $[0, 1]$. Assume $X_0 = h(U_0)$, where h is given by (1.6). Define $X_n = f(X_{n-1}, U_n)$, $n \geq 1$, where, for each i,*

$$f(i, u) = j \quad \text{if } u \in I_{ij}, \text{ for some } j \in S, \tag{1.7}$$

and $I_{ij} = [\sum_{k=0}^{j-1} p_{ik}, \sum_{k=0}^{j} p_{ik})$. Then $\{X_n : n \geq 0\}$ is a Markov chain with initial distribution α and transition probabilities p_{ij}.

Proof. By Remark 12, X_0 has the distribution α_j. Furthermore, by Proposition 11, X_n is a Markov chain with transition probabilities

$$P\{f(i, U_1) = j\} = P\{U_1 \in I_{ij}\} = p_{ij}.$$

We are now ready to establish that there exists a Markov chain associated with any transition matrix.

Corollary 14. (Existence of Markov Chains) *For any Markovian transition probabilities p_{ij} and probability measure α on S, there exists a Markov chain $\{X_n : n \geq 0\}$ on S with transition probabilities p_{ij} and initial distribution α.*

Proof. Corollary 6 in the Appendix says that, for any specified countable collection of distributions on $\mathbb{R}$, there exist a probability space and independent random variables on it that have the specified distributions. This justifies the existence of the random variables U_n in Theorem 13, which in turn justifies the existence of a Markov chain $\{X_n : n \geq 0\}$ on S with transition probabilities p_{ij} and initial distribution α.

Theorem 13 also yields the following simulation procedure.

Remark 15. Simulation of a Markov Chain. One can generate an n-step path $i_0, i_1, \ldots, i_n$ of a Markov chain with transition probabilities p_{ij} and initial distribution α as follows. First generate values $u_0, u_1, \ldots, u_n$ from a uniform

distribution on $[0,1]$. Then set $i_0 = h(u_0)$ and $i_m = f(i_{m-1}, u_m)$, $1 \le m \le n-1$, where h and f are defined respectively by (1.6) and (1.7). The resulting $i_0, i_1, \ldots, i_n$ is the desired n-step path of the Markov chain.

1.4 Examples

This section contains more examples of Markov chains. Some of these are justified by verifying the Markov property, while others are justified by the recursive-equation framework in Proposition 11.

Example 16. Machine Deterioration Model. A machine is continuously used to perform a certain job (e.g., a fork-lift truck in a bottling plant, or a metal cutting tool), and X_n denotes its state of deterioration at time n, where the set of states is $S = \{0, 1, \ldots, \ell\}$. When the machine's deterioration is in state $i < \ell$, in the next time period it either remains at that level with probability p_{ii} or it increases to a level $j > i$ with probability $p_{ij} > 0$. When the machine is in state ℓ, in the next time period it either remains there with probability $p_{\ell\ell}$ or it enters state 0 with probability $p_{\ell 0} = 1 - p_{\ell\ell}$. Entering state 0 means the machine is replaced with a new one (or is repaired to be like new). These movements are independent of the past history of the machine.

Under these assumptions, X_n is a Markov chain with transition matrix

$$\mathsf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & \cdots & \cdots & \cdots \\ 0 & p_{11} & p_{12} & p_{13} & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & p_{22} & p_{23} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{\ell 0} & \cdots & \cdots & 0 & 0 & p_{\ell\ell} \end{bmatrix}$$

Example 17. (s, S) *Inventory Model.* A commodity is stocked in a warehouse to satisfy continuing demands, and the demands $D_1, D_2, \ldots$ in time periods $1, 2, \ldots$ are i.i.d. nonnegative integer-valued random variables. The inventory is controlled by the following (s, S) inventory-control policy, where $\mathsf{s} < \mathsf{S}$ are predetermined integers. At the end of period $n-1$, the inventory level X_{n-1} is observed, and one of the following actions is taken:

- Replenish the stock (instantaneously) up to the level S if $X_{n-1} \le \mathsf{s}$.
- Do not replenish the stock if $X_{n-1} > \mathsf{s}$.

Assume $X_0 \le \mathsf{S}$ for simplicity and that X_0 is independent of the D_n.

Under this control policy, the inventory level satisfies the recursion

$$X_n = \begin{cases} \mathsf{S} - D_n & \text{if } X_{n-1} \le \mathsf{s} \\ X_{n-1} - D_n & \text{if } \mathsf{s} < X_{n-1} \le \mathsf{S}, \end{cases} \quad n \ge 1.$$

Therefore, by Proposition 11, X_n is a Markov chain with

$$p_{ij} = \begin{cases} P\{D_1 = \mathsf{S} - j\} \text{ if } i \le \mathsf{s} \\ P\{D_1 = i - j\} \text{ if } \mathsf{s} < i \le \mathsf{S}. \end{cases}$$

Example 18. Movement on a Graph. Suppose that an item moves on the directed graph shown below, where p_{ij} is the probability the item moves from node i to node j, independent of its past history. The item spends exactly one time period at each node it visits, and so $p_{ii} = 0$ for each i. Then the location of the item X_n at time n is a Markov chain on the set of nodes $S = \{1, \ldots, 7\}$. For instance, the nodes could represent machines in a manufacturing facility, where a job moves from machine 1 to 7. It spends one time period at each machine it visits, and its random path through the machines, which requires 4 time periods, is determined by its type or other factors. When a job is finished at machine 7, another job immediately begins at machine 1, and this 4-period manufacturing cycle is repeated indefinitely, with exactly one job in the system at any time. In this setting, the Markov chain X_n records the machine location of a typical job.

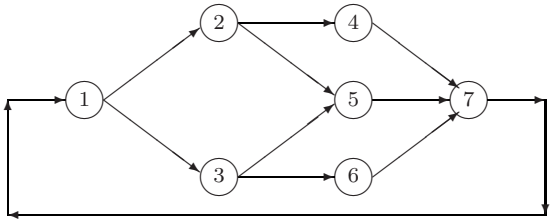

Fig. 1.1 Markov Chain on a Graph

Example 19. Success Runs. Suppose X_n is a Markov chain with transition matrix

$$\mathsf{P} = \begin{bmatrix} 1-p_0 & p_0 & 0 & . & . & . & . \\ 1-p_1 & 0 & p_1 & 0 & . & . & . \\ 1-p_2 & 0 & 0 & p_2 & 0 & . & . \\ . & . & . & . & . & . & . \\ 1-p_m & . & . & . & . & 0 & p_m \\ . & . & . & . & . & . & . \end{bmatrix}$$

This is a model for success runs in a sequence of independent events that result in a success or failure. Namely, X_n denotes the number of successes since the last failure prior to the nth event, and upon having i successes, the next event results in a success with probability p_i, or a failure with probability $1 - p_i$, where $0 < p_i < 1$. For instance, as a model of accidents in a manufacturing plant, X_n could denote the number of weeks without an accident up to week n.

Example 20. Age and Residual Processes for Discrete-Time Renewals. Suppose $\xi_1, \xi_2, \ldots$ are i.i.d. positive integer-valued random variables with $r_i =$

$P\{\xi_1 = i\}$. The times $T_n = \sum_{m=1}^{n} \xi_m$ form a *discrete-time renewal process*, where T_n is the time of the nth renewal (general renewal processes are studied in the next chapter). Consider the process

$$X_n = n - T_{k-1} \quad \text{if } n \in [T_{k-1}, T_k).$$

This is the time since the last renewal prior to or at time n, and X_n is called the *age process* for the renewals. Clearly, X_n is a success-runs Markov chain as above, where

$$p_{i0} = 1 - p_i = r_{i+1}/(1 - \sum_{k=1}^{i} r_k)$$

is the probability of a renewal in the next time period conditioned that the renewal time was greater than i.

A related process is

$$X'_n = T_k - n \quad \text{if } n \in [T_{k-1}, T_k),$$

which is the time to the next renewal after time n. The X'_n is the *residual process* for the renewals. It is a Markov chain on $S = \{0, 1, \ldots\}$ that decreases by one unit in each time period until it reaches 0, and then it jumps to state j with probability r_{j+1}. Thus, its transition probabilities are

$$p_{i,i-1} = 1,\ i \geq 1; \qquad p_{0j} = r_{j+1},\ j \geq 0; \qquad p_{ij} = 0 \text{ otherwise.}$$

As an example, suppose that jobs are processed one at a time and the processing times are $\xi_1, \xi_2, \ldots$ Then X_n denotes the age of the job being processed at time n; age meaning the length of time the job has been in process counting time n. Here $1 - p_i$ is the probability that a job of age i is finished in the next period. The other random variable X'_n denotes the time needed to finish the job being processed at time n.

A variety of queueing and inventory models for systems such as computer systems may be more realistic in discrete time rather than continuous time. The next examples are abstract models for such systems. They have the extraordinary property that their system state is a tractable function of i.i.d. random variables based on the fundamental recursion in Proposition 11.

Example 21. Discrete-Time $M/M/1$ Queueing System. Consider a single-server processing system in which items arrive according to a Bernoulli process as in Example 2, where p is the probability of an arrival (a success) at any discrete time. The service times of the items are i.i.d. and independent of the arrival process, and each service time has a geometric distribution $q(1-q)^{n-1}$, $n \geq 1$, with parameter q and mean $1/q$, where q is the probability of a service completion in any discrete time.

Let X_n denote the number of items in the system at time n. Under these assumptions, the X_n satisfy the recursion $X_n = (X_{n-1} + V_n - U_n)^+$, where

V_n is the number of (potential) arrivals and U_n is the number of (potential) service completions at time n. Under these assumptions, U_n and V_n take values in $\{0,1\}$ and (U_n, V_n), $n \geq 1$, are i.i.d. We also assume they are independent of X_0.

Then by Proposition 11, X_n is a Markov chain on the nonnegative integers S with transition probabilities $p_{ij} = P\{(i + V_1 - U_1)^+ = j\}$, which are, for $i \geq 1$,

$$p_{01} = p, \quad p_{00} = 1 - p,$$
$$p_{i,i+1} = p(1-q), \quad p_{i,i-1} = q(1-p), \quad p_{i,i} = pq + (1-p)(1-q).$$

That is, at each time when at least one item is in service, there may be an arrival and no service completion with probability $p(1-q)$, or a service completion and no arrival with probability $q(1-p)$, or no change in the system with probability $pq + (1-p)(1-q)$ (an arrival and a departure occur, or neither occurs). The X_n is called a discrete-time $M/M/1$ queueing process[4] (analogous queueing processes in continuous time are studied in the next chapters).

One can also view X_n as a random walk, or the size of a population in which a single birth or death occurs with the preceding probabilities. One can model other variations of this queueing system (e.g., with limited waiting space for items, or with batch arrivals) by the input-output model in the next example.

Hereafter, we will use the following standard shorthand notation.

Definition 22. Random variables X and Y are *equal in distribution*, denoted by $X \stackrel{d}{=} Y$, if they have the same distribution. The same notation applies if X and Y are random vectors or more general random elements.

We will now describe a general Markov chain for analyzing queueing systems, including the preceding $M/M/1$ system.

Example 23. Input-Output Process as a Reflected Random Walk. Consider an input-output system in which the quantity of items X_n in the system (possibly negative) at the end of time period n is in a set S of all integers in a fixed interval $[a, b]$. The a and b are integers that may be infinite.

In each period n, the system has a "potential" input (or increase) V_n and "potential" output (or decrease) U_n. The pairs (U_n, V_n), $n \geq 1$, are i.i.d. non-negative integer-valued vectors, and are independent of X_0. Part or all of an arriving quantity is rejected (or disregarded) to the extent that its admittance would force the system state to exceed b. Similarly, outputs are disregarded to the extent that they would move the system state below a. Inputs and outputs occurring at the same time cancel each other.

[4] In $M/M/1$, the M stands for *memoryless* or Markovian when referring to the memoryless geometric distribution (Exercise 7) of inter-arrival times and service times, and 1 refers to the number of servers.

Then the quantity in the system at time n satisfies the recursion[5]

$$X_n = a \vee [b \wedge (X_{n-1} + V_n - U_n)], \quad n \geq 1. \tag{1.8}$$

By Proposition 11, X_n is a Markov chain on S with transition probabilities

$$p_{ij} = P\{a \vee [b \wedge (i + V_1 - U_1)] = j\}. \tag{1.9}$$

This Markov chain has the extraordinary property that its recursive equation (1.8) has a closed form solution for X_n. To see this, consider the random walk or *netput process*

$$Z_n = \sum_{m=1}^{n} (V_m - U_m).$$

Note that

$$X_n = X_0 + Z_n - \sum_{m=1}^{n} \Big[(X_{m-1} + V_m - U_m - b)^+ \\ -(a - (X_{m-1} + V_m - U_m))^+ \Big].$$

This says X_n is the netput Z_n compensated by the quantities in the sum that are disregarded to keep the chain in its state space S. In other words, X_n is the random walk Z_n "reflected" at the boundaries a and b.

One can show by induction, or direct substitution (Exercise 19) that the X_n satisfying (1.8) has the form

$$X_n = \Big[(X_0 + Z_n) \wedge [b + \min_{1 \leq m \leq n} (Z_n - Z_m)] \Big] \\ \bigvee \max_{1 \leq m \leq n} \Big[(a + Z_n - Z_m) \wedge [b + \min_{m+1 \leq \ell \leq n} (Z_n - Z_\ell)] \Big]. \tag{1.10}$$

Because $V_n - U_n$ are i.i.d., the distribution of the increments of Z_n in the preceding expression has the simplification

$$(Z_n - Z_1, Z_n - Z_2, \ldots, Z_n - Z_{n-1}) \stackrel{d}{=} (Z_{n-1}, Z_{n-2}, \ldots, Z_1).$$

Consequently, X_n in (1.10) has the simpler form (in distribution)

$$X_n \stackrel{d}{=} \Big[(X_0 + Z_n) \wedge (b + \min_{0 \leq m \leq n-1} Z_m) \Big] \\ \bigvee \max_{0 \leq m \leq n-1} \Big[(a + Z_m) \wedge (b + \min_{0 \leq \ell \leq n-m-1} Z_\ell) \Big]. \tag{1.11}$$

Here $Z_0 = 0$. This formula describes the distribution of X_n as a function of the distribution of the netput process Z_n (the basic "system data").

[5] $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$; we also use $x^+ = 0 \vee x$.

In particular,

$$X_n \stackrel{d}{=} \begin{cases} (X_0 + Z_n) \vee \Big(a + \max_{0\le m\le n-1} Z_m\Big), & \text{if } b = \infty, \\ (X_0 + Z_n) \wedge \Big(b + \min_{0\le m\le n-1} Z_m\Big), & \text{if } a = -\infty. \end{cases} \tag{1.12}$$

For instance, when $a = 0$, $b = \infty$ and $X_0 = 0$,

$$X_n \stackrel{d}{=} \max_{0\le m\le n} Z_m. \tag{1.13}$$

Note that the $M/M/1$ queue in Example 21 has this nice representation.

The preceding input-output model applies to a variety of contexts. For instance, the input-output variables V_n and U_n need not be inputs or outputs in the usual sense. Here is an example.

Example 24. Waiting Times in a $G/G/1$ Queue.[6] Suppose that items arrive to a processing system at integer-valued times $0 < T_1 < T_2 < \dots$ such that the inter-arrival times $U_n = T_n - T_{n-1}$ are i.i.d., where $T_0 = 0$. The arrival at time T_n has an integer-valued service time V_n, and the V_n are i.i.d. and independent of the arrival times. The service discipline is first-come-first-served with no preemptions.

Of paramount interest are the times that items wait in the queue before receiving service. Let W_n denote the time in queue of the item that arrives at time T_n. Then $D_n = T_n + W_n + V_n$ is the departure time of the nth item. We define W_n by induction. For simplicity, assume the system is empty at time 0. Then clearly $W_1 = 0$ and, assuming $W_1, \dots, W_{n-1}$ are defined,

$$W_n = \begin{cases} 0 & \text{if } D_{n-1} < T_n \\ D_{n-1} - T_n & \text{otherwise.} \end{cases}$$

Substituting $D_{n-1} = T_{n-1} + W_{n-1} + V_{n-1}$ in this expression yields the Lindley recursion

$$W_n = (W_{n-1} + V_{n-1} - U_n)^+, \qquad n \ge 1.$$

Note that W_n is a reflected random walk as in Example 23, where the potential input and output quantities are now the service times and inter-arrival times (here V_{n-1} can be replaced by V_n under the assumptions). Therefore, as in (1.13), W_n is a Markov chain with $W_n \stackrel{d}{=} \max_{0\le m\le n} Z_m$.

The framework in Example 23 also covers the variation in which the waiting times W_n are restricted to not exceed a finite level b. For instance, a system controller may ensure that if an item's waiting time reaches b, then the item is served by an auxiliary server (possibly at a higher cost, but outside of the model).

[6] The $G/G/1$ stands for i.i.d. inter-arrival times with a general distribution, i.i.d. service times with a general distribution, and processing by 1 server.

1.5 Stopping Times and Strong Markov Property

We will now begin a detailed study of the evolution of Markov chains. This section describes the strong Markov property, which is a generalization of the one-step look-ahead Markov property (1.1). It is used for evaluating conditional probabilities conditioned at certain "random times" called stopping times. One important consequence is a regenerative property that the times between entrances of a Markov chain to a fixed state are i.i.d.

We start with preliminaries on stopping times.

Definition 25. A random variable τ that takes values in $\{0, 1, \ldots, \infty\}$ is a *stopping time* for a process $\{X_n : n \geq 0\}$ if, for any finite n, the event $\{\tau = n\}$ is a function of the history $X_0, \ldots, X_n$ up to time n. Exercise 23 gives additional characterizations.

Stopping times are also called *optional times*, or *Markov times* when X_n is a Markov chain. Important examples are hitting times. A *hitting time* of a subset $A \subset S$ by a process X_n is defined by

$$\tau = \min\{n \geq 1 : X_n \in A\}.$$

This is infinite, by convention, when no such n exists. The τ is sometimes called a first entrance or return time[7] to A. It is a stopping time since

$$\{\tau = n\} = \{X_1, \ldots, X_{n-1} \notin A, X_n \in A\}.$$

Constants, of course, are stopping times. There are many ad hoc examples of stopping times such as

$$\tau = \min\{n \geq 2 : X_n = X_{n-1} = X_{n-2}\},$$

the first time the chain remains in the same state for 3 periods.

On the other hand, many random times τ for which $\{\tau = n\}$ involves information about the future $X_{n+1}, X_{n+2}, \ldots$ are not stopping times. An example is the last exit time from a set A defined by

$$\tau = \sup\{n \geq 1 : X_n \in A\},$$

which is infinite when the set is empty.

The Markov property (1.1) says that, at any time n, knowing the present state X_n, the next state X_{n+1} is conditionally independent of the past $X_0, \ldots, X_{n-1}$. We will now establish the strong Markov property that at any finite stopping time (or any deterministic time), the future of the process is conditionally independent of the past given the present state, and the distribution of the future is equal to that of the original chain.

[7] A variation of τ is $\tau' = \min\{n \geq 0 : X_n \in A\}$. Clearly $\tau' = \tau$ if $X_0 \notin A$ and $\tau' = 0$ if $X_0 \in A$.

Theorem 26. (Strong Markov Property) *Suppose that τ is a finite-valued stopping time for a Markov chain X_n on S. Then, for any $i \in S$ and $i_1, i_2, \ldots, j_1, \ldots, j_m \in S$ and $m \geq 1$,*

$$\begin{aligned} P\{X_{\tau+1} = j_1, \ldots, X_{\tau+m} = j_m \Big| X_0 = i_0, \ldots, X_{\tau-1} = i_{\tau-1}, X_\tau = i\} \\ = P\{X_1 = j_1, \ldots, X_m = j_m | X_0 = i\}. \end{aligned} \tag{1.14}$$

Proof. For simplicity, let us write (1.14) as

$$P(A_\tau | B_\tau) = P(A_0 | X_0 = i), \tag{1.15}$$

where $A_n = \{X_{n+1} = j_1, \ldots, X_{n+m} = j_m\}$ and

$$B_n = \{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\}.$$

To prove (1.15), first note that by conditioning on τ,

$$P(A_\tau | B_\tau) = \sum_{n=0}^{\infty} P(A_\tau | B_\tau, \tau = n) P\{\tau = n | B_\tau\}. \tag{1.16}$$

Next, observe that[8]

$$P(A_\tau | B_\tau, \tau = n) = \frac{P(A_n, B_n, \tau = n)}{P(B_n, \tau = n)}.$$

Since τ is a stopping time, $\{\tau = n\}$ is determined by $X_0, \ldots, X_n$, and so $B_n = B_n \cap \{\tau = n\}$ when $P(B_n, \tau = n) > 0$. Using this fact in the preceding display, we have

$$P(A_\tau | B_\tau, \tau = n) = \frac{P(A_n, B_n)}{P(B_n)}.$$

Furthermore, expressing the last two probabilities as multiplications of transition probabilities as in Proposition 6 and canceling terms, we obtain

$$\frac{P(A_n, B_n)}{P(B_n)} = p_{i,j_1} \cdots p_{j_{m-1}, j_m} = P(A_0 | X_0 = i).$$

Using the preceding two displays in (1.16) proves (1.15).

The Strong Markov Property stated in (1.14) for only m steps in the future also applies to the entire future of the chain as follows.

Remark 27. Probabilities of the Infinite Future. Property (1.14) is equivalent to the following: For any $i \in S$ and $B \in S^\infty$,

[8] In expressions like these, commas are often used instead of set intersection; e.g. $P(A_\tau, B_n, \tau = n) = P(A_\tau \cap B_\tau \cap \{\tau = n\})$

$$P\{(X_{\tau+1}, X_{\tau+2}, \ldots) \in B | X_0, \ldots, X_{\tau-1}, X_\tau = i\} = P\{(X_1, X_2, \ldots) \in B | X_0 = i\}. \tag{1.17}$$

This equivalence follows since the distribution (or conditional distribution) of an infinite sequence $(Y_1, Y_2, \ldots)$ of random variables is determined by the distributions of its finite parts $(Y_1, \ldots, Y_m)$ for $m \geq 1$. Another equivalent statement is that, for any bounded function $f : S^\infty \to \mathbb{R}_+$,

$$E[f(X_{\tau+1}, X_{\tau+2}, \ldots) | X_0, \ldots, X_{\tau-1}, X_\tau = i] = E[f(X_1, X_2, \ldots) | X_0 = i]. \tag{1.18}$$

The equivalence of (1.17) and (1.18) follows by basics of equality in distribution; see Exercise 18 ((a) and (b) in this exercise have the same form as (1.17) and (1.18)).

The strong Markov property says, loosely speaking, that a Markov chain regenerates, or starts anew, at a stopping time. For instance, if τ is the hitting time of a state i and it is finite, then since $X_\tau = i$,

$$P\{X_{\tau+m} \in A\} = E[P\{X_{\tau+m} \in A | X_0, \ldots, X_\tau\}] = P\{X_m \in A | X_0 = i\}.$$

Let us extend this idea to see what happens at successive entry times of a state. Suppose state i is such that the Markov chain X_n enters i infinitely often. For simplicity, assume $X_0 = i$. The times $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ at which the chain enters (or hits) i are defined recursively by

$$\tau_n = \min\{m > \tau_{n-1} : X_m = i\}, \qquad n \geq 1. \tag{1.19}$$

These are stopping times of X_n since[9]

$$\{\tau_n > \ell\} = \{\sum_{m=1}^{\ell} \mathbf{1}(X_m = i) < n\}.$$

The next result is a special case of the more general regenerative property in Proposition 67 below.

Proposition 28. (Inter-arrival Times at a State) *Under the preceding assumptions, the times $\xi_n = \tau_n - \tau_{n-1}$, $n \geq 1$, between entrances to state i are i.i.d.*

Proof. We will show by induction that $\xi_1, \ldots, \xi_n$ are i.i.d. for $n \geq 1$. The statement is obviously true for $n = 1$. Next, assume it is true for some n. To prove $\xi_1, \ldots, \xi_{n+1}$ are i.i.d., it suffices by Exercise 20 to show that, for any m and n,

[9] We frequently use the indicator function $\mathbf{1}(\cdot)$ which is 1 or 0 according as the "statement" $(\cdot)$ is true or false.

$$P\{\xi_{n+1} = m|\xi_1, \ldots, \xi_n\} = P\{\xi_1 = m|X_0 = i\}.$$

But this follows by the strong Markov property at τ_n where $X_{\tau_n} = i$, and the fact that $(\xi_1, \ldots, \xi_n)$ is a function of $X_1, \ldots, X_{\tau_n}$.

Example 29. Busy Period. Consider the $M/M/1$ queueing chain X_n in Example 21, where p is the probability of an arrival at any discrete time, and q is the probability of a service completion in any discrete time. Suppose $p < q$, which ensures that the system empties out infinitely often (see Exercise 50). Let $0 \le \tau_1 < \tau_2 < \cdots$ denote the times at which the queue becomes empty (i.e., X_n hits 0). Then by Proposition 28, the durations $\xi_n = \tau_n - \tau_{n-1}$, $n \ge 2$, between the successive empty times are i.i.d.

Let us see what else we can glean from this property. We can write $\xi_n = \gamma_n + \beta_n$, where γ_n is the time until the next arrival after time τ_{n-1} when the system becomes empty, and β_n is the length of the busy period starting at time $\tau_{n-1} + \gamma_n$. By the strong Markov property at τ_{n-1}, where $X_{\tau_{n-1}} = 0$,

$$\begin{aligned} P\{\gamma_n > m\} &= E[P\{\gamma_n > m|X_0, \ldots, X_{\tau_{n-1}}\}] \\ &= P\{\gamma_1 > m|X_0 = 0\} \;=\; (1-p)^m. \end{aligned}$$

This is simply the geometric probability of a typical arrival time. Furthermore, it follows by the the more general regenerative property in Proposition 67 below that the durations of the busy periods β_n are i.i.d. for $n \ge 2$. Although there are no known formulas for the distributions of these times, Exercise 51 shows that $E[\beta_1|X(0) = 0] = q(1-p)/(q-p) - 1/p$.

1.6 Classification of States

Depending on its transition probabilities, a Markov chain may visit some states infinitely often and visit other states only a finite number of times over the infinite time horizon. Also, if a state is visited infinitely often, the mean time between visits may be infinite or finite. These properties are the basis of a classification of states of a Markov chain, which we now present.

Throughout this section, X_n will denote a Markov chain on S with transition probabilities p_{ij}. The form of the distribution of X_0 is not important for many results, and so we will often use conditional probabilities and expectations given $X_0 = i$, and express them as

$$P_i(A) = P\{A|X_0 = i\}, \qquad E_i[Z] = E[Z|X_0 = i\}.$$

For instance, $P_i\{X_n = j\} = p_{ij}^n$.

We begin by studying the hitting times

$$\tau_j = \min\{n \ge 1 : X_n = j\}, \qquad j \in S.$$

Consider the probability

$$f_{ij}^n = P_i\{\tau_j = n\}, \qquad n \geq 1,$$

that the chain starting at i enters j for the "first time" at the nth step. These probabilities are expressible in terms of the p_{ij} by the following recursive equations. The proof is a classic use of a "first-step analysis" that involves conditioning on X_1 and using the Markov property.

Proposition 30. *For $i, j \in S$, $f_{ij}^1 = p_{ij}$ and*

$$f_{ij}^n = \sum_{k \neq j} p_{ik} f_{kj}^{n-1}, \qquad n \geq 2. \tag{1.20}$$

Proof. This expression follows, since conditioning on X_1 and using the Markov property,

$$f_{ij}^n = \sum_{k \neq j} P_i\{\tau_j = n | X_1 = k\} P_i\{X_1 = k\} = \sum_{k \neq j} p_{ik} f_{kj}^{n-1}.$$

Another important quantity for the chain is the probability that beginning at i it ever hits j, which is

$$f_{ij} = P_i\{\tau_j < \infty\} = \sum_{n=1}^{\infty} f_{ij}^n.$$

Note that summing (1.20) over all n yields the linear equations

$$f_{ij} = p_{ij} + \sum_{k \neq j} p_{ik} f_{kj}, \qquad i, j \in S. \tag{1.21}$$

Further properties of the passage or hitting probabilities f_{ij} are in Section 1.7.

We are now ready to start classifying states of the Markov chain X_n.

Definition 31. A state i is *recurrent* if $f_{ii} = 1$ (the chain returns to i with probability one), and i is *transient* if it is not recurrent. A recurrent state i is *positive recurrent* if $E_i[\tau_i] < \infty$, and it is *null recurrent* if $E_i[\tau_i] = \infty$.

The recurrent or transient nature of a state j depends on the number of visits the Markov chain X_n makes to that state, which we denote by

$$N_j = \sum_{n=0}^{\infty} \mathbf{1}(X_n = j).$$

These quantities, which may be infinite, are related to the successive times $0 < \tau_1(j) < \tau_2(j) < \dots$ at which the chain enters j (recall (1.19)), where $\tau_1(j) = \tau_j$. Namely,

$$P_i\{N_j \geq n\} = P_i\{\tau_n(j) < \infty\}, \qquad n \geq 1. \tag{1.22}$$

By the definition of N_j and $E_i[\mathbf{1}(X_n = j)] = p_{ij}^n$, we have

$$E_i[N_j] = \sum_{n=0}^{\infty} p_{ij}^n. \tag{1.23}$$

The distribution of N_j in terms of the f_{ij} is as follows.

Proposition 32. *For $i, j \in S$,*

$$P_i\{N_j > n\} = f_{ij}(f_{jj})^n, \qquad n \geq 0, \tag{1.24}$$

and $P_i\{N_j = 0\} = 1 - f_{ij}$. Hence

$$P_i\{N_j = \infty\} = \begin{cases} 0 & \text{if } f_{jj} < 1 \\ f_{ij} & \text{if } f_{jj} = 1. \end{cases}$$

Furthermore, $E_i[N_j] = 0$ if $f_{ij} = 0$; and otherwise,

$$E_i[N_j] = \begin{cases} f_{ij}/(1 - f_{jj}) & \text{if } f_{jj} < 1 \\ \infty & \text{if } f_{jj} = 1. \end{cases} \tag{1.25}$$

In particular, if i is transient, then $P_i\{N_i = n\} = (1 - f_{ii})f_{ii}^{n-1}$, which is a geometric distribution with mean $1/(1 - f_{ii})$.

Proof. Proceeding by induction, (1.24) is true for $n = 0$ by the definition of f_{ij}. Next, assume (1.24) is true for some $n - 1$. Using $\{N_j > n\} \subseteq \{N_j \geq n\}$ and (1.22),

$$\begin{aligned} P_i\{N_j > n\} &= P_i\{N_j \geq n, N_j > n\} \\ &= P_i\{N_j \geq n\} P_i\{N_j > n | \tau_n(j) < \infty\}. \end{aligned} \tag{1.26}$$

From the strong Markov property at $\tau_n(j)$ and $X_{\tau_n(j)} = j$, it follows that

$$P_i\{N_j > n | \tau_n(j) < \infty\} = f_{jj}.$$

Applying this and the induction hypothesis to the last line in (1.26) yields (1.24) for n, which completes the induction.

Next, note that from (1.24), we have

$$P_i\{N_j = \infty\} = \lim_{n \to \infty} P_i\{N_j > n\} = f_{ij}\mathbf{1}(f_{jj} = 1).$$

Finally (1.25) follows from (1.24) and the formula in Exercise 5 for means.

Here are characterizations for a state to be recurrent or transient.

Corollary 33.

$$\textit{State } i \textit{ is recurrent} \iff P_i\{N_i = \infty\} = 1 \iff E_i[N_i] = \sum_{n=0}^{\infty} p_{ii}^n = \infty.$$

Equivalently,

$$\textit{State } i \textit{ is transient} \iff P_i\{N_i < \infty\} = 1 \iff E_i[N_i] = \sum_{n=0}^{\infty} p_{ii}^n < \infty.$$

Proof. In the first assertion, the forward implications follow by Proposition 32 and (1.23). Also, $E_i[N_i] = \infty$ implies i is recurrent by (1.25).

Corollary 34. *If j is transient, then $\lim_{n\to\infty} p_{ij}^n = 0$, for each i.*

Proof. From the definition of N_j, Proposition 32 and Corollary 33,

$$\sum_{n=0}^{\infty} p_{ij}^n = E_i[N_j] = f_{ij} E_j[N_j] < \infty.$$

Since this sum is finite, $p_{ij}^n \to 0$ for each i.

To continue classifying the evolution of the Markov chain X_n on S, we will use the following terminology. State j is *accessible from* state i, denoted by $i \to j$, if $p_{ij}^n > 0$ for some $n \geq 1$ (i.e., $f_{ij} > 0$). States i and j *communicate* with each other, denoted by $i \leftrightarrow j$, if $i \to j$ and $j \to i$. This communication relation $\leftrightarrow$ is an equivalence relation; see Exercise 24. Consequently, there exists a partition of S into disjoint equivalence classes, which we call *communication classes*.

A set of states C in S is said to be *closed* if no state outside of C is accessible from any state in C ($p_{ij} = 0$ for any $i \in C$, $j \notin C$). If $C = \{i\}$ (a singleton set) is closed, then i is an *absorbing state* (i.e., $p_{ii} = 1$). A closed set may contain several communication classes. Note that if C_1 and C_2 are closed, then so is $C_1 \cap C_2$, which is generally not empty unless the C_i are non-identical communication classes. The C is an *irreducible* set if $i \leftrightarrow j$ for any $i, j \in C$. The communication classes are therefore irreducible.

In addition, we say that the set C is *recurrent* if all of its states are recurrent. Similarly, C is *transient* (or positive recurrent or null recurrent, etc.) if all its states are of that type. A communication class need not be closed if it is transient, but the class is closed when it is recurrent.

Proposition 35. *A recurrent communication class is closed.*

Proof. Suppose C is a communication class that is not closed. Then there exist $i \in C$ and $j \notin C$ such that $p_{ij} > 0$; and $j \not\to i$. Since a return to i is not possible if j is entered, $1 - f_{ii} \geq p_{ij} > 0$. But this contradicts $f_{ii} = 1$, which holds because i is recurrent. Thus C is closed.

Another concept for classifying Markov chains is a subtle property concerning the times between visits to a state. In Example 18, the times between visits to each state are multiples of 4 (they are periodic with period 4). In general, the *period* d_i of a state i is the greatest common divisor of all n that satisfy $p_{ii}^n > 0$. In other words, d_i is the largest integer such that $p_{ii}^n > 0$ if and only if n is a multiple of d_i. State i is *aperiodic* if $d_i = 1$, and otherwise it is *periodic*. For instance, in Examples 5 and 16, each state is aperiodic; and the random walk in Example 3 is periodic with period 3 if each step size is a multiple of 3. Exercise 25 shows that within a communication class, each state has the same period. Therefore, if $p_{ii} > 0$ for any i in a communication class, then all the states in the class are aperiodic.

Example 36. Consider a Markov chain on $S = \{1, 2, \ldots, 7\}$ with transition probabilities

$$\mathsf{P} = \begin{bmatrix} .5 & .4 & .1 & 0 & 0 & 0 & 0 \\ 0 & .8 & 0 & 0 & 0 & .1 & .1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ .5 & 0 & 0 & 0 & .5 & 0 & 0 \\ .3 & 0 & 0 & 0 & .7 & 0 & 0 \\ 0 & .3 & .7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .8 & .2 \end{bmatrix}$$

From the transition graph in Figure 1.2, one can see that $C = \{2, 3, 6, 7\}$ and S are the only two closed sets. Also, C is a communication class and $T = \{1, 4, 5\}$ are communication classes. From results below, it follows that C is a class of positive recurrent states and T is a class of transient states. Furthermore, the states in these classes are aperiodic.

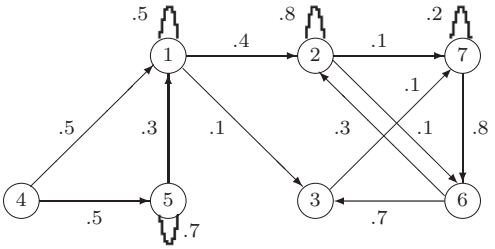

Fig. 1.2 Transition Graph for Example 36

We now establish the major result that all of the states in an irreducible set are of the same type, and they have the same period.

Theorem 37. *Suppose a set of states C is irreducible. Then C is either recurrent or transient. If C is recurrent, then it is either positive recurrent or null recurrent. In addition, each state in C has the same period.*

The proof will use the following characterization of null recurrence, which is proved in Section 1.20 by a coupling argument.

Theorem 38. *For an irreducible Markov chain on S with transition probabilities p_{ij}, a recurrent state i is null-recurrent if and only if*

$$\lim_{n\to\infty} p_{ii}^n = 0.$$

In this case, $\lim_{n\to\infty} p_{ji}^n = 0$*, for* $j \in S$*.*

Proof of Theorem 37. To prove C is either recurrent or transient, it suffices to prove the following statements:
(a) If some $i \in C$ is recurrent and $j \in C$, then j is recurrent.
(b) If some $i \in C$ is transient and $j \in C$, then j is transient.
Clearly (b) follows from (a). Indeed, if $i \in C$ is transient, then no state in C can be recurrent by (a), and hence all states in C must be transient.

To prove (a), suppose $i \in C$ is recurrent, and choose $j \in C$. Since $i \leftrightarrow j$, there exist m and n such that $a = p_{ji}^m p_{ij}^n > 0$. Using $\mathsf{P}^{m+n+\ell} = \mathsf{P}^m \mathsf{P}^\ell \mathsf{P}^n$,

$$p_{jj}^{m+n+\ell} \geq p_{ji}^m p_{ii}^\ell p_{ij}^n = a p_{ii}^\ell. \tag{1.27}$$

Therefore, $\sum_{\ell=0}^\infty p_{jj}^\ell \geq a \sum_{\ell=0}^\infty p_{ii}^\ell$. By Corollary 33, the last sum is infinite since i is recurrent, and hence the first sum is infinite, proving that j is recurrent.

Next, consider the case in which C is recurrent. To prove C is positive recurrent or null recurrent, it suffices to prove the following statements.
(c) If some $i \in C$ is positive recurrent and $j \in C$, then j is positive recurrent.
(d) If some $i \in C$ is null recurrent and $j \in C$, then j is null recurrent.

Arguing as above, it follows that (c) implies (d). It remains to prove (c). We will use the negation of the assertion in Theorem 38, which is that

$$i \text{ is positive recurrent} \iff \limsup_{n\to\infty} p_{ii}^n > 0. \tag{1.28}$$

Now, suppose $i \in C$ is positive recurrent and $j \in C$. From (1.27), which is also valid here, and (1.28), we have

$$\limsup_{n\to\infty} p_{jj}^n \geq a \limsup_{n\to\infty} p_{ii}^n > 0.$$

Thus, the limit superior term for p_{jj}^n is positive, and so j is positive recurrent by (1.28).

The proof that each state in C has the same period is Exercise 25.

An irreducible Markov chain on a finite state space is automatically positive recurrent by the following result.

Corollary 39. *If a closed communication class is finite, then it is positive recurrent.*

Proof. By Theorem 37, a closed communication class C is either transient or positive recurrent or null recurrent. Therefore, it suffices to show that C cannot be transient or null recurrent. Suppose $i \in C$ is transient. Then each $j \in C$ is transient by statement (b) in the proof of Theorem 37, and so $p_{ij}^n \to 0$ by Corollary 34. Therefore, since C is finite, we would have the contradiction

$$1 = P_i\{X_n \in C\} = \sum_{j \in C} p_{ij}^n \to 0.$$

Thus C is not transient. A similar argument shows that C is not null recurrent; here one uses $p_{ij}^n \to 0$ from Theorem 38.

We end this section by describing a canonical decomposition of the Markov chain X_n on S. The starting point is the fact that one can partition S into disjoint communication classes; the classes are irreducible, but not necessarily closed. Let $C_1, C_2, \ldots$ denote the finite or infinite sequence of communication classes that are recurrent, and hence closed by Proposition 35. Then set $T = S \setminus \cup_k C_k$. Clearly, T is transient because it consists of communication classes that are not recurrent. Of course, T is not necessarily closed, and it may be empty, or it may be equal to S.

The transition matrix of X_n, with its rows arranged so the sets of states $C_1, C_2, \ldots, T$ appear in that order, has the form

$$\mathsf{P} = \begin{bmatrix} \mathsf{P}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathsf{P}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathsf{P}_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Q_1 & Q_2 & Q_2 & \cdots & Q \end{bmatrix} \tag{1.29}$$

where

$$\mathsf{P}_k = (p_{ij} : i, j \in C_k), \quad Q_k = (p_{ij} : i \in T, j \in C_k), \quad Q = (p_{ij} : i, j \in T).$$

In summary, these observations establish that the Markov chain X_n on S has the following structure.

Theorem 40. (Decomposition Property) *The state space has the unique representation $S = T \cup C_1 \cup C_2 \cup \ldots$, where T is the set of transient states, and $C_1, C_2, \ldots$ are closed, irreducible recurrent sets. The transition matrix has the form (1.29).*

This decomposition tells us that if the Markov chain starts in a recurrent set C_k, then it moves within that set forever under the transition probabilities in P_k. On the other hand, if the chain starts in the transient set T, then it moves within T and either enters one of the recurrent sets and remains in that set thereafter, or it may remain in T forever, provided T is infinite (a finite T cannot be closed by Corollary 39).

Example 41. Consider a Markov chain on $S = \{1, 2, \ldots, 9\}$ whose transition matrix has the following form, where $\star$ means a positive probability.

$$\mathsf{P} = \begin{bmatrix} 0 & 0 & 0 & \star & \star & \star & 0 & 0 & 0 \\ \star & \star & \star & 0 & 0 & \star & 0 & 0 & 0 \\ 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ \star & \star & 0 & \star & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \star & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 & \star \\ 0 & 0 & 0 & 0 & \star & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 & \star \end{bmatrix}$$

Tracing out the transition graph of this chain, one can see that its closed irreducible sets are $C_1 = \{3\}$, $C_2 = \{5, 8\}$ and $C_3 = \{6, 7, 9\}$; and its set of transient states is $T = \{1, 2, 4\}$. Now, based on these sets and Theorem 40, reordering the states in the order $3, 5, 8, 6, 7, 9, 1, 2, 4$ yields the more informative transition matrix:

$$\mathsf{P} = \left[\begin{array}{c|cc|ccc|ccc} \star & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \star & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \star & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \star & \star & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & \star & 0 & 0 & 0 \\ \hline 0 & \star & 0 & \star & 0 & 0 & 0 & 0 & \star \\ \star & 0 & 0 & \star & 0 & 0 & \star & \star & 0 \\ \star & 0 & 0 & 0 & 0 & 0 & 0 & \star & \star \end{array}\right]$$

We end the classification of Markov chains with a few more terms. The Markov chain X_n is *irreducible* if its state space S is irreducible. In that case, by Theorems 37 and 40, all states of the chain are either positive recurrent, null recurrent or transient, and all states have the same period. The Markov chain is called *ergodic* if it is irreducible, and its states are positive recurrent and aperiodic.

The limiting behavior of ergodic Markov chains is the main topic for the rest of this chapter. Before getting into this, further insight on the movement of a Markov chain between subsets of its state space are given in the next section followed by a section on branching processes.

1.7 Hitting and Absorbtion Probabilities

Basic performance measures for a Markov chain are as follows.

• The probability that a chain will ever hit a specified set of states (e.g., a desired stock price).

• The probability that a chain, beginning in transient states, will be absorbed in a certain recurrent class of states.
• The mean time for a chain to hit a set, or to be absorbed in recurrent states.
This section characterizes these and related quantities.

Throughout this section, X_n will be a Markov chain on S with transition probabilities p_{ij}. Let S_0 and S_1 be disjoint (nonempty) subsets of the state space S. Define

$$\tau = \min\{n \geq 1 : X_0, X_1, \ldots, X_{n-1} \in S_0, X_n \in S_1\}.$$

This is infinite when no such n exists. The τ is the time at which the chain exits S_0 for the first time and enters S_1, or the time at which the chain first hits S_1 without passing through the set $(S_0 \cup S_1)^c$ (which may be empty). Clearly τ is a stopping time of the chain X_n.

We will first consider the hitting or entrance probabilities

$$\gamma_i = P_i\{\tau < \infty\}, \qquad i \in S_0.$$

Clearly γ_i is the probability the chain starting at $i \in S_0$ eventually enters or hits S_1 without passing through $(S_0 \cup S_1)^c$ (or simply the probability the chain hits S_1 from S_0). To avoid degenerate cases, assume the chain does not stay in S_0 forever:

$$P_i\{X_n \in S_0, n \geq 0\} = 0, \qquad i \in S_0. \tag{1.30}$$

This general formulation of hitting-probabilities covers a variety of events, including the following.
(a) A chain hits a single state or set B ($S_0 = B^c$, $S_1 = B$).
(b) A chain beginning in its set of transient states T is absorbed in a recurrent class C ($S_0 = T$, $S_1 = C$).
(c) A chain hits a set B without passing through a set A ($S_0 = (A \cup B)^c$, $S_1 = B$).
(d) A chain beginning in its set of transient states T is absorbed in a recurrent class C at a state $k \in C$ ($S_0 = T$, $S_1 = \{k\}$).

The next result characterizes the hitting probabilities $\gamma = (\gamma_i : i \in S_0)$ in terms of $r = (r_i : i \in S_0)$ and $Q = (q_{ij} : i, j \in S_0)$, where

$$r_i = \sum_{j \in S_1} p_{ij}, \qquad q_{ij} = p_{ij}, \quad i, j \in S_0.$$

Theorem 42. *The probabilities γ_i of hitting S_1 from S_0 satisfy*

$$\gamma_i = r_i + \sum_{j \in S_0} p_{ij}\gamma_j, \qquad i \in S_0. \tag{1.31}$$

This in matrix notation is $\gamma = r + Q\gamma$. *Moreover,* $\gamma = \sum_{n=0}^{\infty} Q^n r$, *and this is the smallest solution to the equation*

$$y = r + Qy, \qquad y \geq 0. \tag{1.32}$$

If S_0 *is finite, then* $\gamma = (I - Q)^{-1} r$ *and this is the unique solution to (1.32).*

Proof. Assertion (1.31) follows since, by conditioning on X_1,

$$\begin{aligned}\gamma_i &= \sum_{j \in S} P_i\{\gamma < \infty | X_1 = j\} p_{ij} \\ &= \sum_{j \in S} [\mathbf{1}(j \in S_1) + \gamma_j \mathbf{1}(j \in S_0)] p_{ij}.\end{aligned}$$

Next, by the definitions of r_j and q_{ij} and taboo-probability reasoning as in Example 9,

$$\begin{aligned}\gamma_i &= \sum_{n=1}^{\infty} P_i\{\tau = n\} \\ &= \sum_{n=1}^{\infty} \sum_{j \in S_0} P_i\{X_1, \ldots, X_{n-2} \in S_0, X_{n-1} = j\} P\{X_n \in S_1 | X_{n-1} = j\} \\ &= \sum_{n=1}^{\infty} \sum_{j \in S_0} q_{ij}^{n-1} r_j.\end{aligned}$$

Thus, in matrix notation $\gamma = \sum_{n=0}^{\infty} Q^n r$.

Now, suppose y is any nonnegative solution to $y = r + Qy$. Then by induction

$$y = \sum_{m=0}^{n} Q^m r + Q^{n+1} y, \qquad n \geq 0.$$

Consequently,

$$y \geq \lim_{n \to \infty} \sum_{m=0}^{n} Q^m r = \sum_{m=0}^{\infty} Q^m r = \gamma.$$

Thus γ is the smallest solution to $y = r + Qy$.

Next, suppose S_0 is finite. To show γ is the unique solution to $y = r + Qy$, let y be any solution and set $x = y - \gamma$. Clearly $x = Qx$ and by induction $x = Q^n x$. Now $Q^n \to 0$ since by (1.30),

$$q_{ij}^n \leq P_i\{X_m \in S_0, m \leq n - 1\} \to 0.$$

Therefore, $x = Q^n x \to 0$, which proves that γ is the unique solution to $y = r + Qy$.

From $\gamma = r + Q\gamma$, it follows that $\gamma = (I - Q)^{-1} r$, provided the inverse $(I - Q)^{-1}$ of $I - Q$ exists. But this exists by a basic property of linear

algebra, since the preceding paragraph showed that the only solution of the finite linear equations $(I - Q)x = 0$ is $x = 0$.

We now consider expected hitting times and related random quantities. Fix a subset B of S and let $\tau_B = \min\{n \geq 1 : X_n \in B\}$. For each $i \in B^c$, assume $P_i\{\tau_B < \infty\} = 1$. For $f : S \to \mathbb{R}_+$, define

$$v_i = E_i[\sum_{n=0}^{\tau_B - 1} f(X_n)], \qquad i \in B^c.$$

The $f(j)$ would typically be a cost or utility for the chain visiting state j. Then v_i would be the expected utility up to the time the chain enters B. An important example is the expected time $v_i = E_i[\tau_B]$ to hit B (here $f(j) \equiv 1$). As another example, v_i is the expected number of visits X_n makes to a fixed state $k \in B^c$ before it enters B, when $f(j) = \mathbf{1}(j = k)$.

From the proof of Theorem 42, it is clear that the following result is true.

Theorem 43. *The assertions and proof of Theorem 42 hold with*

$$\gamma_i = v_i, \quad \textit{and} \quad r_i = \sum_{j \in B} p_{ij} f(j), \quad i \in B^c,$$

where it is assumed that each $r_i < \infty$.

Example 44. Gambler's Ruin Model. Consider the random walk X_n on $S = \{0, 1, \ldots, m\}$ as in Example 4, where X_n is the fortune of a gambler at the nth play of a game. At each play, the gambler wins or looses one dollar with respective probabilities p and $q = 1 - p$. Clearly 0 and m are absorbing states and states $1, \ldots, m - 1$ are transient states. Let γ_i be the probability that the gambler goes broke (enters state 0) when starting with $X_0 = i$ dollars. Theorem 42 tells us that the probabilities γ_i are the unique solution to the difference equation

$$\gamma_i = q\gamma_{i-1} + p\gamma_{i+1}, \qquad 1 \leq i \leq m - 1,$$

with boundary conditions $\gamma_0 = 1$ and $\gamma_m = 0$. We will solve this by a method analogous to that used for differential equations.

Consider a solution of the form $\gamma_i = r^i$, where r is a parameter to be determined. Substituting this in the preceding equation and dividing by r^{i-1}, we obtain $pr^2 - r + q = 0$, which has roots $r_1 = 1$, $r_2 = q/p$. In case $r_1 \neq r_2$ (i.e. $p \neq 1/2$), the general solution is $\gamma_i = a_1 r_1^i + a_2 r_2^i$, where the coefficients are given by the boundary conditions

$$1 = \gamma_0 = a_1 + a_2, \qquad 0 = \gamma_m = a_1 + a_2(p/q)^m.$$

Therefore the solution is

$$\gamma_i = \frac{(q/p)^i - (q/p)^m}{1 - (q/p)^m}, \qquad \text{if } p \neq 1/2.$$

Furthermore, letting $q/p \to 1$ in this expression and using L'Hospital's rule we obtain the solution $\gamma_i = 1 - i/m$, if $p = q = 1/2$.

In addition, the probability that starting at i the gambler reaches m before being ruined is $1 - v_i$, since the fortune eventually is absorbed in 0 or m. The expected time the game lasts before the gambler's fortune reaches 0 or m is the subject of Exercise 45.

1.8 Branching Processes

This section describes a classical Markov chain model for describing the size of a population in which each member of the population independently produces offspring. The main issue is under what conditions does the population explode to infinity or become extinct in a finite time. The principal results are a characterization of the probability of extinction and a procedure for computing it.

Consider a population of identical items that evolves in discrete time as follows. Each item in the population lives for a single time period and at the end of its one-period life it produces k items with probability p_k, $k \geq 0$, independently of the other items. This off-spring probability measure has a finite mean μ, and $p_0 \in (0, 1)$. Consequently, whenever the population size is i, the probability that the population dies out in the next time period is p_0^i. A major quantity of interest is the probability that the population eventually becomes extinct.

Let X_n denote the population size at time n. For simplicity, assume $X_0 = 1$ (the case $X_0 > 1$ is covered in Exercise 61). This process satisfies the recursive formula $X_{n+1} = 0$ when $X_n = 0$, and otherwise,

$$X_{n+1} \stackrel{d}{=} \xi_{n1} + \xi_{n2} \cdots + \xi_{nX_n}, \qquad n \geq 0, \tag{1.33}$$

where $\xi_{n1}, \xi_{n2}, \ldots$ are i.i.d. random variables with probability measure p that are independent of X_n. The interpretation is that the ith item present at time n produces ξ_{ni} items at time $n + 1$.

From this representation it follows by Proposition 11 that X_n is a Markov chain. It is clear that the chain is aperiodic, 0 is an absorbing state, and the other states are transient. Because of the representation (1.33), the analysis in this section is based on properties of sums of i.i.d. random variables, not involving knowledge of Markov chains.

Some indication of the growth of X_n is given by its mean. Conditioning on X_{n-1} and using (1.33), we have

$$E[X_n] = E\Big[E[X_n|X_{n-1}]\Big] = \mu E[X_{n-1}].$$

Then iterating this backward and using $X_0 = 1$ yields

$$E[X_n] = \mu^n. \tag{1.34}$$

From this one can see how the mean population evolves as a function of the mean production μ of each item. In particular, $E[X_n]$ converges to 0, 1 or ∞ according as μ is < 1, $= 1$, or > 1.

We will now discuss the possibility of the population becoming extinct in terms of

$$z_n = P\{X_n = 0\}, \qquad n \geq 1,$$

which is the probability of extinction before or at time n. Clearly z_n is increasing, since $X_n = 0$ implies $X_{n+1} = 0$. This and $z_n \leq 1$ ensure the existence of the limit

$$z = \lim_{n\to\infty} z_n,$$

which is the probability the population eventually becomes extinct.

To characterize the extinction probabilities z_n and z, we will use

$$\phi(s) = \sum_{k=0}^{\infty} p_k s^k, \qquad 0 \leq s \leq 1,$$

which is the generating function of the single-item production. Here are some preliminary insights.

Proposition 45. *The time-dependent extinction probabilities z_n are given by $z_1 = p_0$ and*

$$z_n = \phi(z_{n-1}), \qquad n \geq 2. \tag{1.35}$$

In addition, the extinction probability $z \in (p_0, 1]$ is such that $z = \phi(z)$; so z is a fixed point *of ϕ.*

Proof. Since $X_0 = 1$, we have, upon conditioning on X_1,

$$z_n = \sum_{k=1}^{\infty} P\{X_n = 0|X_1 = k\}p_k, \qquad n \geq 1.$$

When $X_1 = k$, each of these k items generates a separate subpopulation, and so the population at time n is the union of k i.i.d. subpopulations that have evolved for $n-1$ time units. Therefore, $P\{X_n = 0|X_1 = k\} = z_{n-1}^k$. Substituting this in the preceding display proves (1.35). Furthermore, since the generating function $\phi(s)$ is continuous and $z_n \to z$, then letting $n \to \infty$ in (1.35) yields $z = \phi(z)$.

The preceding result says the extinction probability z is a fixed point of ϕ and z is obtainable by the recursion (1.35). Here is a characterization of z.

Theorem 46. *If $\mu \leq 1$, then $z = 1$. If $\mu > 1$, then z is the unique s in $(0,1)$ that satisfies $s = \phi(s)$.*

Proof. If $\mu < 1$, then by (1.34),

$$P\{X_n \geq 1\} \leq E[X_n] = \mu^n \to 0.$$

Therefore $z = \lim_{n\to\infty}(1 - P\{X_n \geq 1\}) = 1$. If $\mu > 1$, then Lemma 47 below establishes that ϕ has a unique fixed point in $(0,1)$, which is necessarily the extinction probability z by Proposition 45. Finally, if $\mu = 1$, the only case of interest is $p_0 + p_1 < 1$, since $p_0 > 0$. Then another application of Lemma 47 yields $z = 1$.

Lemma 47. *In the context of Theorem 46, if $\mu > 1$ or if $\mu = 1$ and $p_0 + p_1 < 1$, then the generating function $\phi(s)$ is strictly convex. Moreover, $\phi(s)$ has a unique fixed point z in $(p_0, 1)$ if $\mu > 1$, and $z = 1$ if $\mu = 1$.*

Proof. Under the assumptions, ϕ is strictly convex since

$$\phi''(s) = \sum_{k=2}^{\infty} k(k-1)p_k s^{k-1} > 0, \qquad s \in [0,1].$$

Also, the graph of $\phi(s)$ goes through the point $(1,1)$ and its slope there is $\phi'(1) = \mu \geq 1$. Therefore, if $\mu > 1$, there is exactly one $s < 1$ for which $s = \phi(s)$. That is, ϕ has a unique fixed point z in $(p_0, 1)$. The graphs in Figure 1.3 show the two possibilities

$$z < 1 \text{ if } \phi'(1) = \mu > 1, \qquad \text{and } z = 1 \text{ if } \phi'(1) = \mu = 1.$$

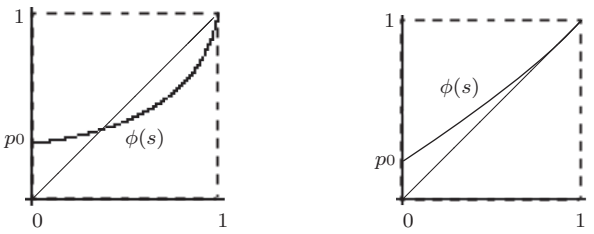

Fig. 1.3 Graph of $\phi(s)$

For the interesting case in which $\mu > 1$, the extinction probability z is the unique solution of the equation $s = \phi(s)$, for $s \in (0,1)$ (which is a polynomial equation when the quantity an item can produce is bounded). The solution can sometimes be obtained when $\phi(s)$ has a nice form such as in Exercise 62.

As another example, consider the elementary case in which each item can produce at most 2 items. Then the extinction probability z is the solution of the quadratic equation $s = p_0 + p_1 s + p_2 s^2$. Thus $z = \min\{p_0/p_2, 1\}$. More

generally, when each item can produce at most m items, then z is the solution of a polynomial equation of order m.

Here is an easy way to find bounds on z.

Remark 48. Bounds. In case $\mu > 1$, suppose $0 < a \le b < 1$ are such that $\phi(a) \ge a$ and $\phi(b) \le b$. Then $a \le z \le b$ by Figure 1.3.

For practical applications, one can compute the extinction probability by the recursion (1.35) as follows.

Remark 49. Computation of Extinction Probability. When $\mu > 1$, the following procedure yields a value z^* that is within ε of z. Let $z_0 = 0$ and compute $z_1, z_2, \ldots$ by $z_n = \phi(z_{n-1})$. Stop at

$$z^* = \min\{z_n : \ \phi(z_n + \varepsilon) \le z_n + \varepsilon\}.$$

Then $z_n \uparrow z^*$ and Remark 48 ensure that $0 \le z - z^* \le \varepsilon$.

1.9 Stationary Distributions

We now turn to the main theme of characterizing the long run or limiting behavior of ergodic Markov chains in terms of their stationary distributions. This section introduces the notion of a stationary distribution for a Markov chain and shows that a Markov chain is stationary if its initial distribution is a stationary distribution. The rest of the section is devoted to establishing the existence of stationary distributions for positive recurrent Markov chains. Included are general formulas for stationary distributions and examples showing how to evaluate such distributions.

Definition 50. A probability measure π on S is a *stationary distribution* for the Markov chain X_n (or for P) if

$$\pi_i = \sum_j \pi_j p_{ji}, \qquad i \in S. \tag{1.36}$$

This equation in matrix notation is $\pi = \pi\mathsf{P}$, where $\pi = (\pi_i : i \in S)$ is a row vector. In general, any measure η with $\sum_i \eta_i \le \infty$ that satisfies $\eta = \eta\mathsf{P}$ is an *invariant measure* for P.

A Markov chain may have an infinite number of stationary distributions or invariant measures; see Exercise 30. We begin by relating stationary distributions to stationary processes.

Definition 51. A stochastic process $\{X_n : n \ge 0\}$ on a general state space is *stationary* if, for any $n \ge 0$,

$$(X_n, \dots, X_{n+k}) \stackrel{d}{=} (X_0, \dots, X_k), \qquad k \geq 1. \tag{1.37}$$

That is, the finite-dimensional distributions of X_n remain the same if the time is shifted by any amount n. A stationary process is sometimes said to be a process that is in *equilibrium* or in *steady state*.

If a process X_n is stationary, then necessarily the distribution of each X_n does not depend on n (i.e., $X_n \stackrel{d}{=} X_0$, $n \geq 1$). This simpler condition is also sufficient for a Markov chain to be stationary as we now justify. Also, a Markov chain is stationary if and only it its initial state has a stationary distribution.

Proposition 52. *The following statements are equivalent for the Markov chain X_n.*
(a) *X_n is stationary.*
(b) *$X_n \stackrel{d}{=} X_0$, $n \geq 1$.*
(c) *The distribution of X_0 is a stationary distribution.*

Proof. Because the finite-dimensional probabilities of the Markov chain are simply products of transition probabilities, the stationarity criterion (1.37) is equivalent to

$$P\{X_n = i_0\} \prod_{m=1}^{k} p_{i_{m-1},i_m} = P\{X_0 = i_0\} \prod_{m=1}^{k} p_{i_{m-1},i_m}, \quad i_0, \dots, i_k \in S.$$

In light of this, (a) is equivalent to (b).

Next, note that (b) is equivalent to

$$P\{X_0 = i\} = \sum_j P\{X_0 = j\} p_{ji}^n, \quad n \geq 1.$$

Then (b) implies (c), since the last equality for $n = 1$ is statement (c). Conversely, if (c) holds, then one can show by induction that the preceding equality holds for each $n \geq 1$, and hence (c) implies (b).

We are now ready to present the major result that any positive recurrent Markov chain has a stationary distribution. This is based on the following preliminary result that an irreducible, recurrent Markov chain has a unique invariant measure up to a multiple by a constant.

Consider the measure η on S defined as follows. For any fixed $i \in S$, let $\tau_i = \min\{n \geq 1 : X_n = i\}$, and define $\eta_i = 1$ and

$$\eta_j = E_i[\sum_{n=0}^{\tau_i - 1} \mathbf{1}(X_n = j)], \qquad j \in S \backslash \{i\}. \tag{1.38}$$

The η_j (a function of i) is the expected number of visits X_n makes to state j in between visits to i. The measure η is *positive* if each η_j is positive. Another

way of writing the preceding expression is

$$\eta_j = \sum_{n=0}^{\infty} P_i\{X_n = j, \tau_i > n\}, \qquad j \in S\backslash\{i\}. \tag{1.39}$$

This follows by taking expectations of the identity[10]

$$\sum_{n=0}^{\tau_i - 1} \mathbf{1}(X_n = j) = \sum_{n=0}^{\infty} \mathbf{1}(X_n = j, \tau_i > n).$$

The proof of the following result is at the end of this section.

Theorem 53. (Invariant Measures) *If the Markov chain X_n is irreducible and recurrent, then η defined by (1.38) is a positive invariant measure for the chain. This invariant measure is unique up to multiplication by a constant.*

In some instances, the invariant measure (1.38) for an irreducible recurrent Markov chain is finite and hence can be normalized to be a stationary distribution. This is true, of course, when the state space is finite. It is also true for infinite state spaces when the chain has the added condition of being positive recurrent. Here is a more general result that an irreducible chain has a positive stationary distribution if and only if it is positive recurrent.

Theorem 54. (Stationary Distributions) *An irreducible Markov chain X_n has a positive stationary distribution if and only if all of its states are positive recurrent. In that case, the stationary distribution is unique and has the following form: For any fixed $i \in S$,*

$$\pi_j = \frac{E_i\Big[\sum_{n=0}^{\tau_i - 1} \mathbf{1}(X_n = j)\Big]}{\mu_i}, \qquad j \in S, \tag{1.40}$$

where $\tau_i = \min\{n \geq 1 : X_n = i\}$ and $\mu_i = E_i[\tau_i]$. Another expression for this distribution is

$$\pi_j = 1/\mu_j, \qquad j \in S. \tag{1.41}$$

Proof. Suppose the irreducible chain X_n has a positive stationary distribution π. We first show that the chain is not transient. Suppose to the contrary that it is transient. Since $\pi = \pi\mathsf{P}$, we have $\pi = \pi\mathsf{P}^n$, for $n \geq 1$. Also, by Corollary 34, $p_{ij}^n \to 0$ as $n \to \infty$ for each j. Therefore, by the dominated convergence theorem for sums,[11]

$$\pi_j = \sum_i \pi_i p_{ij}^n \to 0.$$

But this contradicts $\pi_j > 0$. Thus the chain is not transient.

[10] Identities like $\sum_{n=0}^{\tau-1} Y_n = \sum_{n=0}^{\infty} Y_n \mathbf{1}(\tau > n)$ are convenient for computing expectations.

[11] As an example, for a probability measure p on S and bounded $f_n : S \to \mathbb{R}$ such that $f_n(i) \to f(i)$, it follows that $\sum_i f_n(i)p_i \to \sum_i f(i)p_i$.

Now, since the chain is irreducible but not transient, it must be recurrent by Theorem 37. To prove the chain is positive recurrent, it suffices to show that $E_i[\tau_i]$ is finite for some i. Let η be defined as in (1.38); it is the unique positive invariant measure for the chain by Theorem 53. Because the chain is assumed to have a stationary distribution π, this distribution is a multiple of η, and hence it must have the form $\pi_j = \eta_j / \sum_k \eta_k$, where $\sum_k \eta_k$ is necessarily finite. Now by Exercise 5, an interchange of sums, and expression (1.39) for η_j, we have

$$\begin{aligned} E_i[\tau_i] &= \sum_{n=0}^{\infty} P_i\{\tau_i > n\} = \sum_{n=0}^{\infty} \sum_k P_i\{X_n = k, \tau_i > n\} \\ &= \sum_k \eta_k < \infty. \end{aligned} \tag{1.42}$$

Thus i is positive recurrent, and hence the chain is positive recurrent.

Now consider the converse and assume the chain is positive recurrent. Then $\sum_k \eta_k = \mu_i$ is finite as we saw in (1.42). Hence $\pi_j = \eta_j / \mu_i$, which is (1.40), is the unique positive stationary distribution for the chain.

Finally, note that $\eta_i = 1$ and (1.38) imply $\pi_i = 1/\mu_i$, and this is true for any fixed i. Thus, it follows that (1.41) is an alternative expression for the distribution in (1.40).

The preceding theorem yields the following criterion for ergodicity that is very useful for applications.

Corollary 55. *An irreducible aperiodic Markov chain is ergodic if and only if it has a stationary distribution. In this case, the stationary distribution is positive and has the form shown in Theorem 54.*

Proof. If a Markov chain is ergodic, it has a positive stationary distribution by Theorem 54. Conversely, if an irreducible aperiodic Markov chain has a stationary distribution, then this stationary distribution is positive by Theorem 53, and hence is ergodic by Theorem 54.

Although (1.40) and (1.41) are closed-form expressions for this distribution, they are typically not used to obtain numerical values for π. They come in handy, however, for the analysis or modeling of ergodic Markov chains (e.g., Proposition 69 below).

For an irreducible, aperiodic Markov chain, a common approach for determining whether or not the chain is ergodic is as follows. First, find (if possible) a positive invariant measure η that satisfies the equations $\eta = \eta P$. This is typically done algebraically, or by a computer package or by verifying by substitution that a candidate distribution satisfies the equations. Then find conditions under which η is finite or infinite. When η is finite, it can be normalized to be the stationary distribution of the chain, and hence the

chain is ergodic. On the other hand, when η is infinite, the chain is either null-recurrent or transient. Here are two illustrations.

Example 56. Consider the flexible manufacturing system in Example 5 in which X_n is the state of a machine, which can be idle (state 0) or producing a type i part ($i = 1$ or 2 or 3). Suppose X_n is a Markov chain with transition matrix

$$\mathsf{P} = \begin{bmatrix} .1 & .2 & .2 & .5 \\ .3 & .4 & 0 & .3 \\ .4 & 0 & .4 & .2 \\ .3 & 0 & .2 & .5 \end{bmatrix}$$

This chain is clearly ergodic since it is irreducible, finite and aperiodic. The equations $\pi = \pi\mathsf{P}$ are

$$\begin{aligned} \pi_0 &= .1\pi_0 + .3\pi_1 + .4\pi_2 + .3\pi_3 \\ \pi_1 &= .2\pi_0 + .4\pi_1 \\ \pi_2 &= .2\pi_0 \qquad\quad + .4\pi_2 + .2\pi_3 \\ \pi_3 &= .5\pi_0 + .3\pi_1 + .2\pi_2 + .5\pi_3 \end{aligned}$$

That is,

$$\begin{aligned} -9\pi_0 + 3\pi_1 + 4\pi_2 + 3\pi_3 &= 0 \\ 2\pi_0 - 6\pi_1 \qquad\quad &= 0 \\ 2\pi_0 \qquad - 6\pi_2 + 2\pi_3 &= 0 \\ 5\pi_0 + 3\pi_1 + 2\pi_2 - 5\pi_3 &= 0 \end{aligned}$$

We will fix π_0 and solve the last set of equations for the other variables, and then use $1 = \sum_i \pi_i$ to find π_0. From the second and third equations in the last display, $\pi_1 = \pi_0/3$, and $\pi_2 = \pi_0/3 + \pi_3/3$. Substituting these in the first equation yields $\pi_3 = 20\pi_0/13$, and then $\pi_2 = 11\pi_0/13$. The fourth equation is not needed. Finally, using $1 = \sum_i \pi_i$, we obtain the stationary distribution

$$\pi = \Big(\frac{39}{145}, \frac{13}{145}, \frac{33}{145}, \frac{60}{145}\Big).$$

This distribution tells us, for instance, that the machine works on a type 3 part about 41% of the time, and it is idle more than 25% of the time.

Example 57. Machine Deterioration: Stationary Distribution. Consider Example 16 in which the Markov chain X_n denotes the state of deterioration of a machine over time with transition matrix

$$\mathsf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & \cdots & \cdots & \cdots \\ 0 & p_{11} & p_{12} & p_{13} & \cdots & \cdots & \cdots \\ 0 & 0 & p_{22} & p_{23} & p_{24} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{\ell 0} & \cdots & \cdots & 0 & 0 & p_{\ell\ell} \end{bmatrix}$$

Assume the p_{ij} are positive. Then the chain is ergodic since it is irreducible and aperiodic and finite. The system of equations $\pi = \pi\mathsf{P}$ is

$$\begin{aligned} \pi_0 &= \pi_0 p_{00} + \pi_\ell p_{\ell 0} \\ \pi_i &= \sum_{j=0}^{i} \pi_j p_{ji}, \qquad 1 \le i \le \ell. \end{aligned}$$

Because of the structure of these equations, they have a tractable algebraic solution, which one can see by considering them one at a time in the order $i = 0, 1, \ldots$ Namely, the solution to the last ℓ equations, for a fixed π_0, is $\pi_i = \pi_0 c_i$, where the c_i are given recursively by $c_0 = 1$ and

$$c_i = \frac{1}{1 - p_{ii}} \sum_{j=1}^{i-1} c_j p_{ji}, \qquad 1 \le i \le \ell.$$

Then from $1 = \sum_{i=0}^{\ell} \pi_i$, it follows that $\pi_0 = (\sum_{i=0}^{\ell} c_i)^{-1}$. Thus, the stationary distribution of the chain is

$$\pi_i = \frac{c_i}{\sum_{j=0}^{\ell} c_j}, \qquad 0 \le i \le \ell.$$

The remainder of this section is devoted to the proof of Theorem 53, which is restated here.

Theorem 58. (Invariant Measures) *If the Markov chain X_n is irreducible and recurrent, then η defined by (1.38) is a positive invariant measure for the chain. This invariant measure is unique up to multiplication by a constant.*

Proof. The first task is to verify $\eta = \eta\mathsf{P}$. From (1.39),

$$\eta_j = \sum_{n=0}^{\infty} P_i\{X_n = j, \tau_i > n\},$$

where i is a fixed reference state. For a more convenient expression, let $Q = (q_{k\ell})$, where $q_{k\ell} = p_{k\ell}\mathbf{1}(\ell \neq i)$, and denote its nth product by $Q^n = (q^n_{k\ell})$, where $q_{k\ell} = \mathbf{1}(k = \ell)$. Expressing τ_i in terms of the X_m, and using products of probabilities as in Section 1.2, we have for $n \ge 1$ and $j \neq i$,

$$P_i\{X_n = j, \tau_i > n\} = P_i\{X_m \neq i,\, 1 \le m \le n-1,\, X_n = j\} = q^n_{ij}.$$

This is the taboo probability of moving from i to j in n steps while avoiding i. Thus, in matrix notation, the vector η (including $\eta_i = 1$), has the form

$$\eta = \sum_{n=0}^{\infty} e_i Q^n, \tag{1.43}$$

where e_i is the m-dimensional row vector with 1 in position i and 0 elsewhere. Using a little algebra on this sum and noting that Q is P with zeros in column i, we have

$$\eta = e_i + \sum_{n=1}^{\infty} e_i Q^{n-1} Q = e_i + \eta Q = \eta \mathsf{P}.$$

Thus, η is an invariant measure.

To prove each η_j is finite, suppose to the contrary that $\eta_j = \infty$ for some $j \neq i$. Since $i \leftrightarrow j$, there is an m such that $p_{ji}^m > 0$. Also, $\eta = \eta \mathsf{P}$ implies $\eta = \eta \mathsf{P}^m$ by induction. Then we obtain the contradiction

$$1 = \eta_i = \sum_k \eta_k p_{ki}^m \geq \eta_j p_{ji}^m = \infty.$$

Therefore, each η_j is finite. Furthermore, each η_j is positive because there is an ℓ such that $p_{ij}^\ell > 0$, and so using $\eta_i = 1$,

$$\eta_j = \sum_k \eta_k p_{kj}^\ell \geq \eta_i p_{ij}^\ell = p_{ij}^\ell > 0.$$

To prove η is unique up to multiplication by a constant, let γ be another positive invariant measure. Assume $\gamma_i = 1$, where i is the fixed reference state in the definition of η. There is no loss in generality in this assumption since any such measure can be normalized by dividing by γ_i. Let $\zeta = \gamma - \eta$. The proof will be complete upon showing that $\zeta = 0$.

We first show that $\zeta \geq 0$. Note that $\gamma = \gamma \mathsf{P}$ can be written as

$$\gamma = e_i + \gamma Q. \tag{1.44}$$

Then by induction,

$$\gamma = \sum_{m=0}^{n} e_i Q^m + \gamma Q^{n+1}, \qquad n \geq 0.$$

Since the last term is nonnegative,

$$\gamma \geq \sum_{m=0}^{n} e_i Q^m \to \sum_{m=0}^{\infty} e_i Q^m = \eta.$$

Thus $\zeta = \gamma - \eta \geq 0$.

Next, subtracting $\eta = \eta\mathsf{P}$ from $\gamma = \gamma\mathsf{P}$ yields $\zeta = \zeta\mathsf{P}$. This implies $\zeta = \zeta\mathsf{P}^n$, $n \geq 1$. Using this and $\zeta_i = 1 - 1 = 0$, we have

$$0 = \zeta_i = \sum_{j \neq i} \zeta_j p_{ji}^n, \qquad n \geq 1. \tag{1.45}$$

By the irreducibility of the chain, for each j, there is an n_j such that $p_{ji}^{n_j} > 0$. Then $\zeta \geq 0$ and (1.45) imply that $\zeta_j = 0$ for each j.

1.10 Limiting Distributions

We have seen that a positive recurrent Markov chain has a unique stationary distribution π that satisfies the balance equations $\pi = \pi\mathsf{P}$. This section shows that this stationary distribution is also the limiting distribution when the Markov chain is ergodic.

Suppose X_n is a Markov chain on S with transition probabilities p_{ij}. A probability measure π is the *limiting distribution* of the chain if

$$\lim_{n\to\infty} P\{X_n = i\} = \pi_i, \qquad i \in S.$$

Note that π does not depend on the distribution of X_0. Exercise 31 shows that a limiting distribution for a Markov chain is always a stationary distribution — no additional assumptions on the chain are needed. On the other hand, there are non-ergodic chains with stationary distributions that are not limiting distributions.

Here is the major result concerning limiting distributions for ergodic Markov chains.

Theorem 59. *If a Markov chain is ergodic, then its stationary distribution is its limiting distribution, which is positive.*

Proof. The assertion follows by the coupling described in Theorems 110 and 111 below. Theorem 54 ensures that π is positive. An alternative approach is to prove the assertion by applying the discrete-time version of the renewal theorem; see the proof of Theorem 51 in Chapter 2.

The next result, which includes Corollary 55, says that the existence of a limiting distribution is another criterion for ergodicity.

Corollary 60. *For an irreducible, aperiodic Markov chain, the following statements are equivalent.*
(a) *The chain is ergodic.*
(b) *The chain has a stationary distribution.*
(c) *The chain has a limiting distribution.*
When these statements hold, the stationary and limiting distributions are the same, and it is positive.

Proof. The equivalence of (a) and (b) follows by Corollary 55. Next, (a) $\Rightarrow$ (c) by Theorem 59, and (c) $\Rightarrow$ (b) by Exercise 31. The equality and positiveness of the stationary and limiting distributions follow by Theorem 59 and Corollary 55.

The rest of this section covers properties of limiting distributions. We will often use the notion of convergence in distribution of random variables defined as follows (later, we use this convergence for more general random elements). In particular, if a Markov chain X_n has a limiting distribution π, then X_n converges in distribution to a random variable X with distribution π.

Definition 61. Random variables X_n with values in S *converge in distribution* to a random variable X, denoted by $X_n \xrightarrow{d} X$, if

$$\lim_{n\to\infty} P\{X_n = i\} = P\{X = i\}, \qquad i \in S.$$

The Appendix points out that $X_n \xrightarrow{d} X$ is equivalent to

$$E[f(X_n)] \to E[f(X)], \text{ for any bounded } f : S \to \mathbb{R}. \tag{1.46}$$

This characterization of convergence in distribution yields the following result for Markov chains.

Remark 62. Limits of Expectations. The Markov chain X_n has a limiting distribution π if and only if, for any bounded function $f : S \to R$,

$$\lim_{n\to\infty} E[f(X_n)] = \sum_i f(i)\pi_i.$$

The last limit statement also justifies the limiting average

$$n^{-1} \sum_{m=1}^{n} E[f(X_m)] \to \sum_i f(i)\pi_i.$$

This follows by the property that $n^{-1}\sum_{m=1}^{n} a_m \to a$ if $a_n \to a$. More general averages are in Remark 78 and Exercises 34 and 36.

A useful property of ergodic Markov chains is that many functions of it also have limiting distributions. This is because of the following elementary but important result, which says that the distant future of an ergodic Markov chain behaves as if the chain were stationary.

Proposition 63. (Asymptotic Stationarity). *An ergodic Markov chain X_n is asymptotically stationary in the sense that*

$$(X_n, X_{n+1}, \ldots) \xrightarrow{d} (\bar{X}_0, \bar{X}_1, \ldots), \quad \text{as } n \to \infty, \tag{1.47}$$

where $\bar{X}_n$ is a stationary version of X_n (the chain $\bar{X}_n$ is stationary and it has the same transition probabilities as X_n).

Proof. Since $P\{X_n = i\} \to P\{\bar{X}_0 = i\}$ and X_n and $\bar{X}_n$ have the same transition probabilities, it follows that, for $B \subset S^\infty$,

$$\begin{aligned} P\{(X_n, X_{n+1}, \ldots) \in B\} &= \sum_i P\{X_n = i\} P_i\{(i, \bar{X}_1, \bar{X}_2, \ldots) \in B\} \\ &\to \sum_i P\{\bar{X}_0 = i\} P_i\{(i, \bar{X}_1, \bar{X}_2, \ldots) \in B\} \\ &= P\{(\bar{X}_0, \bar{X}_1, \ldots) \in B\}. \end{aligned}$$

The convergence is of the form (1.46) for $f(i) = P_i\{(i, \bar{X}_1, \bar{X}_2, \ldots) \in B\}$. This proves (1.47).

Remark 64. By properties of convergence in distribution, each of the following statements is equivalent to (1.47). For any bounded $f : S^\infty \to \mathbb{R}$,

$$\begin{aligned} E[f(X_n, X_{n+1}, \ldots)] &\to E[f(\bar{X}_0, \bar{X}_1, \ldots)], \\ f(X_n, X_{n+1}, \ldots) &\xrightarrow{d} f(\bar{X}_0, \bar{X}_1, \ldots). \end{aligned}$$

Example 65. For an ergodic Markov chain X_n on a countable set S of real numbers, we know by Proposition 63 and Remark 64 that

$$\begin{aligned} X_{n+1} - X_n &\xrightarrow{d} \bar{X}_1 - \bar{X}_0, \\ \max\{X_n, X_{n+1}\} &\xrightarrow{d} \max\{\bar{X}_0, \bar{X}_1\}. \end{aligned}$$

In other words,

$$\begin{aligned} \lim_{n\to\infty} P\{X_{n+1} - X_n = k\} &= \sum_i \pi_i p_{i,i+k}, \\ \lim_{n\to\infty} P\{\max\{X_n, X_{n+1}\} \le k\} &= \sum_{i\le k} \sum_{j\le k} \pi_i p_{ij}. \end{aligned}$$

1.11 Regenerative Property and Cycle Costs

The strong Markov property says, loosely speaking, that a Markov chain regenerates, or starts anew, at a stopping time. This section describes a related result that if a Markov chain enters a fixed state infinitely often, then it regenerates each time it enters the state, and its sample-path segments between entrances to the state are i.i.d. This fundamental "regenerative" property is the key to obtaining strong laws of large numbers for ergodic Markov chains that we present in the next section, and for obtaining related central limit theorems that we present in Chapter 2. We also present a formula for the

mean of certain costs or utilities in a regenerative cycle. Basics of more general regenerative processes are covered in Chapter 2.

We begin by identifying a regenerative property that any ergodic Markov chain has. Suppose a Markov chain X_n on S has a state i that it enters infinitely often (this would be true for any state when the chain is irreducible and recurrent). For this fixed state i, let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ denote the times at which the chain hits i. Recall that the times $\tau_n - \tau_{n-1}$ between the entrances to i are i.i.d. by Proposition 28. This is part of the more general regenerative property described as follows.

The nth *segment* (or cycle) of the Markov chain in the time interval $[\tau_{n-1}, \tau_n)$ is defined by

$$\zeta_n = \Big(\tau_n - \tau_{n-1}, \{X_m : m \in [\tau_{n-1}, \tau_n)\}\Big), \qquad n \geq 1.$$

The segment ζ_n takes values in $\{(m, \mathbf{j}) : m \geq 1, \mathbf{j} \in \{i\} \times S^{m-1}\}$, and it contains all the information about the sample path of the chain in the time interval $[\tau_{n-1}, \tau_n)$.

Definition 66. The Markov chain X_n is *regenerative over the times* τ_n if the segments ζ_n, for $n \geq 1$, are i.i.d. In particular, $\tau_n - \tau_{n-1}$ are i.i.d. More general regenerative processes are studied in the next chapter.

Proposition 67. (Regenerative Property) *If the Markov chain X_n starting at $X_0 = i$ enters state i infinitely often, then the chain is regenerative over the times τ_n at which the chain enters i.*

Proof. This follows by the proof of Proposition 28 with ζ_n in place of ξ_n.

Example 68. Consider the (s, S) inventory model in Example 17, where X_n is the inventory level in period n. Assuming the demands are positive, it follows that this Markov chain X_n enters state S, where the inventory is replenished, infinitely often. Then this chain is regenerative over the replenishment times $\tau_1 < \tau_2 < \ldots$. Consequently, the times between replenishment $\tau_n - \tau_{n-1}$ are i.i.d. Also, since the evolutions of the inventory in these cycles are i.i.d., there are many performance parameters that are i.i.d. such as:

• The amounts of time $Y_n = \sum_{m=\tau_{n-1}}^{\tau_n - 1} \mathbf{1}(X_m > i)$ in the cycles during which the inventory exceeds i.

• The numbers of demands $N_n = \sum_{m=\tau_{n-1}}^{\tau_n - 1} \mathbf{1}(X_{m+1} < X_m)$ in the cycles that are satisfied.

The regenerative property plays an important role in determining average costs or performance values for a Markov chain. A variety of costs or utility functions for a Markov chain are of the form $\sum_{m=0}^{n} V_m$, where V_n are values associated with the chain at time n. Then next result is a formula for the mean of such a cost function in a regenerative cycle. The formula is due to the form of a stationary distribution expressed as (1.40). We use the formula in the next sections to describe limiting values in strong laws of large numbers.

Proposition 69. (Cycle Costs) *Let X_n be an irreducible positive recurrent Markov chain with stationary distribution π. Suppose V_n, $n \geq 1$, are real-valued random variables associated with the chain such that*

$$E_i[V_n|X_0, \ldots, X_n] = a_{X_n}, \qquad n \geq 0,$$

where a_j are constants. Then, for the hitting time τ_i of a fixed state i,

$$E_i\Big[\sum_{n=0}^{\tau_i - 1} V_n\Big] = \pi_i^{-1} \sum_j a_j \pi_j, \tag{1.48}$$

provided the last sum is absolutely convergent.[12]

Proof. Noting that $\{\tau_i > n\}$ is a function of $X_0, \ldots, X_n$, and using the pull-through formula for conditional probabilities, the left-hand side of (1.48) is

$$\begin{aligned} E_i\Big[\sum_{n=0}^{\infty} V_n \mathbf{1}(\tau_i > n)\Big] &= E_i\Big[E_i\Big[\sum_{n=0}^{\infty} \mathbf{1}(\tau_i > n) V_n \Big| X_0, \ldots, X_n\Big]\Big] \\ &= E_i\Big[\sum_{n=0}^{\infty} \mathbf{1}(\tau_i > n) a_{X_n}\Big] = \sum_j a_j E_i\Big[\sum_{n=0}^{\infty} \mathbf{1}(X_n = j, \tau_i > n)\Big] \\ &= \pi_i^{-1} \sum_j a_j \pi_j. \end{aligned}$$

The last equality is due to Theorem 54 and (1.39), and the equality before it uses Fubini's theorem in the appendix and the assumption that $\sum_j |a_j|\pi_j$ is finite.

A typical example of (1.48), for $f : S^2 \to \mathbb{R}$, is

$$E_i\Big[\sum_{n=0}^{\tau_i - 1} f(X_n, X_{n+1})\Big] = \pi_i^{-1} \sum_j \Big[\sum_k f(j,k) p_{jk}\Big] \pi_j. \tag{1.49}$$

In this case,

$$E[f(X_n, X_{n+1})|X_0, \ldots, X_n] = \sum_k f(X_n, k) p_{X_n k}.$$

Here is an illustration of the use of (1.49).

Example 70. Let X_n denote the $M/M/1$ queueing chain as in Example 21. Assume that $p < q$, which says the probability of an arrival is less than the probability of a service completion in a time period. Then by Exercise 50, the stationary distribution is $\pi_i = (1-\rho)\rho^i$, $i \geq 0$, where $\rho = p(1-q)/[q(1-p)]$.

[12] A sum $\sum_j c_j$ is absolutely convergent if $\sum_j |c_j|$ is finite.

The ρ is the *traffic intensity*. Let us consider the number of customers N that are served in a busy period (between visits to state 0). Since N is also the number of arrivals in a busy period, assuming $X_0 = 0$,

$$N = \sum_{n=0}^{\tau_0 - 1} \mathbf{1}(X_{n+1} = X_n + 1).$$

From $p_{01} = p$, $p_{j,j+1} = p(1-q)$, $j \geq 1$, and (1.49),

$$E_0[N] = \pi_0^{-1} \sum_j \pi_j p_{j,j+1} = p + \sum_{j=1}^{\infty} \rho^j p(1-q).$$

Thus, we have the result $E_0[N] = p(1 - q\rho)/(1-\rho)$.

1.12 Strong Laws of Large Numbers

Many properties or performance measures of a stochastic process are expressed as limiting averages, e.g., the average amount of time the process spends in a state, or the average cost of running the process. Such an average is usually expressed as a strong law of large numbers (SLLN), where the limit is a function of the limiting distribution of the process. In this section, we present several strong laws of large numbers for ergodic Markov chains.

We begin with a motivating example.

Example 71. Multi-Type Job Processing. A system processes several types of jobs, where S denotes the set of job types. Suppose that a sequence of jobs it processes (labeled by job type) is an ergodic Markov chain X_n on S with stationary distribution π (here the label n refers to the order in a sequence and is not a time parameter in the usual sense). The revenue received from each type-i job is $f(i)$, where $f : S \to \mathbb{R}_+$. Our interest is in the average revenue per job processed over the infinite time horizon.

First, note that the average revenue per job from the first n jobs is $n^{-1} \sum_{m=1}^{n} f(X_m)$. Then by Theorem 74 below, the average revenue per job (over the infinite horizon) is the limiting average[13]

$$\lim_{n \to \infty} n^{-1} \sum_{m=1}^{n} f(X_m) = \sum_i f(i)\pi_i \qquad a.s.$$

In applications like this, rewards or utility values are sometimes "random" functions of the process. For instance, in this job-processing setting, suppose the processing times of a type-i job are i.i.d. random variables with mean

[13] a.s. stands for *almost surely*, which means with probability one. For instance, $Y_n \to Y$ a.s. means $P\{\lim_{n\to\infty} Y_n = Y\} = 1$, and $Y > Z$ a.s. means $P\{Y > Z\} = 1$.

a_i, independent of other jobs and the order in which they are processed. Let $V_1, V_2, \ldots$ denote the processing times of the respective jobs $X_1, X_2, \ldots$ Then by Theorem 75 below, the average processing time per job is

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} V_m = \sum_i a_i \pi_i \quad \text{a.s.}$$

As we will see shortly, the average cost and processing time in the preceding example are examples of a SLLN. Such laws are applications of the following one, which is proved in standard texts on probability theory.

Theorem 72. (Classical SLLN) *If $Y_1, Y_2, \ldots$ are i.i.d. random variables with a mean that may be infinite, then*

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} Y_m = E[Y_1] \quad a.s.$$

We will now describe analogous results for Markov chains. More general SLLNs for regenerative processes are in Chapter 2. For this discussion, suppose X_n is an ergodic Markov chain on S with transition probabilities p_{ij} and stationary distribution π.

We first consider limiting averages associated with hitting times of a state. Fix a state i, and let $0 < \tau_1 < \tau_1 < \cdots$ denote the successive times at which the chain hits or enters state i. Define

$$N_i(n) = \sum_{m=1}^{n} \mathbf{1}(X_m = i),$$

which is the number of times the chain hits i in the first n time periods.

Proposition 73. *The average time between visits to state i is*

$$\lim_{n\to\infty} n^{-1}\tau_n = 1/\pi_i \quad a.s. \tag{1.50}$$

The average number of visits to state i is

$$\lim_{n\to\infty} n^{-1} N_i(n) = \pi_i \quad a.s. \tag{1.51}$$

Proof. We can write

$$n^{-1}\tau_n = n^{-1}\tau_1 + (1 - 1/n)\Big[(n-1)^{-1} \sum_{m=2}^{n} (\tau_m - \tau_{m-1})\Big].$$

Exercise 33 justifies $n^{-1}\tau_1 \to 0$ a.s. Also, by Proposition 28, the times $\tau_m - \tau_{m-1}$ between entrances to state i are i.i.d. for $m \geq 2$, and their mean is $\mu_i = 1/\pi_i$ by Theorem 54. Then the classical SLLN ensures that the bracketed

term in the preceding expression converges to $1/\pi_i$ a.s. Combining these observations[14] proves (1.50).

In addition, (1.51) follows from (1.50) by Theorem 10 in Chapter 2.

We are now ready to present two results that establish limiting averages like those in Example 71. The first result is a SLLN, also called an *ergodic theorem*, for ergodic Markov chains.

Theorem 74. *For the ergodic Markov chain X_n with stationary distribution π and any $f : S \to \mathbb{R}$,*

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} f(X_m) = \sum_i f(i)\pi_i \quad a.s.,$$

provided the sum is absolutely convergent.

This SLLN justifies the approximation

$$\sum_{m=1}^{n} f(X_m) \approx n \sum_i f(i)\pi_i, \qquad \text{for large } n.$$

In addition, the distribution of the sum $\sum_{m=1}^{n} f(X_m)$ can be approximated by a normal distribution with mean $n \sum_i f(i)\pi_i$ as shown by the central limit theorem for Markov chains in Example 68 in Chapter 2.

Theorem 74 is a special case of the next SLLN for "random" functions of ergodic Markov chains. Here we use the following conditional probability terminology. Random variables $Y_1, Y_2, \ldots$ are *conditionally independent given* the Markov chain $X = \{X_n\}$ if, for any $y_1, \ldots, y_m$ and m,

$$P\{Y_1 \le y_1, \ldots, Y_m \le y_m | X\} = \prod_{k=1}^{m} P\{Y_k \le y_k | X\}.$$

If in addition, $P\{Y_k \le y|X\} = P\{Y_k \le y|X_k\}$, independent of k, then we say $Y_1, Y_2, \ldots$ are *conditionally independent given $\{X_n\}$ with distributions* $P\{Y_1 \le y|X_1\}$. For instance, in Example 71 the processing times $V_1, V_2, \ldots$ of the respective jobs $X_1, X_2, \ldots$ are conditionally independent given $\{X_n\}$ with distributions $P\{V_1 \le v|X_1 = i\}$ that have means a_i, $i \in S$.

Theorem 75. *Associated with the ergodic Markov chain X_n, suppose V_n, $n \ge 1$ are random variables that are conditionally independent given $\{X_n\}$ with distributions $P\{V_1 \le v|X_1 = i\}$ that have means a_i, $i \in S$. Then*

[14] One often establishes convergence by exploiting the following properties, which follow automatically from the analogous limit statements for real numbers. If $Y_n \to Y$ a.s. and $Z_n \to Z$ a.s., then $Y_n + Z_n \to Y + Z$ a.s. and $Y_n Z_n \to YZ$ a.s.. Also, $Y_{\nu_n} \to Y$ a.s. for integer-valued $\nu_n \to \infty$ a.s.

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} V_m = \sum_i a_i \pi_i \quad a.s., \tag{1.52}$$

provided the sum is absolutely convergent.

Proof. We will use the notation $N_i(n)$ and τ_n in Proposition 73 for a fixed state i, and assume $X_0 = i$. Consider the sequence

$$Z_n = \Big(\tau_n - \tau_{n-1}, \{(X_m, V_m) : m \in [\tau_{n-1}, \tau_n)\}\Big), \qquad n \geq 1.$$

This Z_n contains all the information about the process $\{(X_m, V_m) : m \geq 0\}$ in the time interval $[\tau_{n-1}, \tau_n)$, where $\tau_0 = 0$. It follows as in Proposition 67 that the Z_n are i.i.d. Consequently, the values

$$Y_k = \sum_{m=1}^{\infty} V_m \mathbf{1}(\tau_{k-1} < m \leq \tau_k), \qquad k \geq 1,$$

accumulated in the respective intervals $(\tau_{k-1}, \tau_k]$ are i.i.d., since they are deterministic functions of the Z_k. Then by the classical SLLN and Proposition 69,

$$n^{-1} \sum_{k=1}^{n} Y_k \to E[Y_1] = \pi_i^{-1} \sum_j a_j \pi_j \quad \text{a.s.} \tag{1.53}$$

Next, we show that this limit leads to (1.52). First assume the V_m are nonnegative. Note that[15]

$$n^{-1} \sum_{k=1}^{N_i(n)} Y_k \leq n^{-1} \sum_{m=1}^{n} V_m \leq n^{-1} \sum_{k=1}^{N_i(n)+1} Y_k. \tag{1.54}$$

Applying (1.51) and (1.53) to the left-hand side of (1.54), we have

$$\Big[N_i(n)/n\Big]\Big[N_i(n)^{-1} \sum_{k=1}^{N_i(n)} Y_k\Big] \to \pi_i E[Y_1] = \sum_j a_j \pi_j \quad a.s.$$

A similar argument shows that the right-hand side of (1.54) has this same limit. Consequently, $n^{-1} \sum_{m=1}^{n} V_m$ also has the limit $\sum_j a_j \pi_j$, since it is sandwiched between these two terms.

For general V_m, we can write $V_m = V_m^+ - V_m^-$, where $V_m^+ = \max\{0, V_m\}$ and $V_m^- = -\min\{V_n, 0\}$. Then by what we just proved,

[15] Here and below, we use the convention that $\sum_{n=1}^{0}(\cdot) = 0$.

$$n^{-1}\sum_{m=1}^{n} V_m = n^{-1}\sum_{m=1}^{n} V_m^+ - n^{-1}\sum_{m=1}^{n} V_m^- \\ \to \sum_j a_j^+\pi_j - \sum_j a_j^-\pi_j = \sum_j a_j\pi_j \quad \text{a.s.}$$

Example 76. Functions Involving Auxiliary Variables. In the setting of Theorem 75, a typical random function or value associated with the Markov chain has the form $V_n = g(X_n, Y_n)$, where Y_n are auxiliary variables in S' and $g : S \times S' \to \mathbb{R}$. For instance, Y_n might represent the state of an environment or cost that affects the value at time n, and $g(i, y)$ is the value whenever the chain is in state i and the auxiliary variable is in state y. Assume the Y_n are i.i.d. and independent of the chain. Then according to Theorem 75, the average value is

$$n^{-1}\sum_{m=1}^{n} g(X_m, Y_m) \to \sum_i E[g(i, Y_1)]\pi_i \quad \text{a.s.},$$

provided the last sum is absolutely convergent.

This section ends with a few observations about limiting averages. Intuition suggests that a limiting average for a stochastic process should be the mean value associated with a stationary version of the process. Although this is not generally true, it is for the Markov averages above.

Remark 77. Stationary Chains. Suppose $\bar{X}_n$ is a stationary version of the ergodic Markov chain X_n in the preceding results (both chains have the same transition probabilities, but $\bar{X}_n$ is stationary). By Proposition 52, each $\bar{X}_n$ has the distribution π, and so in the context of the preceding results,

$$E[f(\bar{X}_n)] = \sum_i f(i)\pi_i,$$
$$E[V_n] = \sum_j E[V_n|\bar{X}_n = j]\pi_j \;=\; \sum_j a_j\pi_j.$$

That is, the limiting averages in Theorems 74 and 75 are the means of single-period values when the chain is in equilibrium.

Do limiting averages exist for "expected values" analogous to those in the preceding SLLN?

Remark 78. Limit Laws for Expectations. The limiting averages in Proposition 73 and Theorems 74 and 75 are also true for expected values. For instance, $n^{-1}E[\tau_n] \to 1/\pi_i$ and

$$n^{-1}\sum_{m=1}^{n} E[V_m] \to \sum_i a_i\pi_i.$$

Such results for expected values follow by a discrete-time version of the key renewal theorem in Chapter 2 for functions of regenerative processes.

Although these limits for means hold for regenerative processes, they generally do not hold for any non-regenerative sequence that obeys a SLLN.

1.13 Examples of Limiting Averages

This section contains several corollaries and illustrations of the strong laws of large numbers in the preceding section. For this discussion, suppose X_n is an ergodic Markov chain on S with transition probabilities p_{ij} and stationary distribution π.

Costs or utilities associated with "jumps" of the Markov chain X_n are often modeled by sums of the form $\sum_{m=1}^{n} f(X_{m-1}, X_m)$, where $f(i,j)$ is the cost of a jump from i to j. Averages of such sums are described by the following extension of Theorem 74.

Corollary 79. *For a fixed integer ℓ, the process $\tilde{X}_n = (X_n, \ldots, X_{n+\ell})$ is an ergodic Markov chain on $S^{\ell+1}$ with stationary distribution*

$$\pi(\mathbf{i}) = \pi_{i_0} p_{i_0,i_1} \cdots p_{i_{\ell-1},i_\ell}.$$

Hence, for $f : S^{\ell+1} \to \mathbb{R}$,

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} f(\tilde{X}_m) = \sum_{\mathbf{i}\in S^{\ell+1}} f(\mathbf{i})\pi(\mathbf{i}) \quad a.s.,$$

provided the sum is absolutely convergent.

Proof. The first assertion follows by Exercise 28, and the second assertion is an example of Theorem 74.

The following are basic properties of ergodic Markov chains that are manifestations of limiting averages.

Example 80. Movements Between Sets. The average number of transitions the chain makes from a set A to a set B per unit time is

$$\lambda(A,B) = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} \mathbf{1}(X_{m-1} \in A, X_m \in B) \quad \text{a.s.}$$

This limit exists by Corollary 79 with $\ell = 1$ and $f(i,j) = \mathbf{1}(i \in A, j \in B)$, and it is

$$\lambda(A,B) = \sum_{i\in A} \pi_i \sum_{j\in B} p_{ij}.$$

This quantity is often called the *rate* at which the chain moves from A to B. It is also the mean number of transitions the chain makes from A to B per

unit time when the chain is in equilibrium; recall Remark 77. In particular, the rate at which the chain moves from state i to state j is $\lambda(i,j) = \pi_i p_{ij}$, and therefore

$$\lambda(A,B) = \sum_{i\in A}\sum_{j\in B}\lambda(i,j).$$

Example 81. Balance Equations. In terms of rates of movements, the total balance equations for the Markov chain have the following interpretation: For $i \in S$,

$$\pi_i = \sum_j \pi_j p_{ji} \quad \text{is equivalent to} \quad \lambda(i,S) = \lambda(S,i).$$

This says that the rate at which the chain moves out of i is equal to the rate at which it moves into state i. More generally,

$$\lambda(A,A^c) = \lambda(A^c,A), \tag{1.55}$$

which says that the rate at which the chain moves out of a set A is the rate at which it moves into A. To prove (1.55), first observe that summing the balance property $\pi_i = \sum_j \pi_j p_{ji}$ on $i \in A$, we have $\lambda(A,S) = \lambda(S,A)$. Then subtracting $\lambda(A,A)$ from these terms yields (1.55).

The next example is an extension of Proposition 73.

Example 82. Entrance Rates into Sets. The number of entrances the Markov chain makes into a set $A \subset S$ up to time n is defined by

$$\nu_A(n) = \sum_{m=1}^{n} \mathbf{1}(X_{m-1} \notin A, X_m \in A).$$

By Example 80, the average number of entrances per unit time into A is

$$\lambda(A) = \lim_{n\to\infty} n^{-1}\nu_A(n) = \sum_{i\in A^c, j\in A} \pi_i p_{ij}. \tag{1.56}$$

This is the same as the rate $\lambda(A^c,A)$ at which the chain enters A.

A related quantity is the nth time the chain enters the set A, which is $\tau_A(n) = \min\{m : \nu_A(m) = n\}$. The limiting averages of these times is

$$\lim_{n\to\infty} n^{-1}\tau_A(n) = 1/\lambda(A) \quad \text{a.s.}$$

This follows since $\nu_A(\tau_A(n)) = n$ and $\tau_A(n) \to \infty$; and so by (1.56),

$$n^{-1}\tau_A(n) = \Big(\frac{\nu_A(\tau_A(n))}{\tau_A(n)}\Big)^{-1} \to 1/\lambda(A) \quad a.s.$$

Example 83. Average Sojourn Time in a Set. Let $W_n(A)$ denote the amount of time the Markov chain spends in a set A on its nth sojourn in that set. Then the average sojourn or waiting time in A is

$$W(A) = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} W_m(A) \quad \text{a.s.}$$

We will prove that this limit exists and

$$W(A) = \frac{1}{\lambda(A)} \sum_{i\in A} \pi_i. \tag{1.57}$$

First note that by the definition of the waiting times and the $\tau_A(n)$ from the preceding example,

$$\sum_{m=1}^{n} W_m(A) = \sum_{i\in A} N_i(\tau_A(n)),$$

where $N_i(n)$ is the cumulative sojourn time in state i up to time n. Then (1.57) follows, since by the preceding example and $n^{-1}N_i(n) \to \pi_i$ from Proposition 73, we have

$$\begin{aligned} n^{-1} \sum_{m=1}^{n} W_m(A) \;&=\; (\tau_A(n)/n)\Big[\sum_{i\in A} \tau_A(n)^{-1} N_i(\tau_A(n))\Big] \\ &\to \frac{1}{\lambda(A)} \sum_{i\in A} \pi_i. \end{aligned}$$

Example 84. Machine Deterioration: Average Time Machine is Good. Consider the ergodic Markov chain X_n in Example 57 that denotes the state of deterioration of a machine. Its stationary distribution is

$$\pi_i = \frac{c_i}{\sum_{k=0}^{\ell} c_k}, \qquad 0 \le i \le \ell,$$

where $c_0 = 1$ and $c_i = (1 - p_{ii})^{-1} \sum_{j=1}^{i-1} c_j p_{ji}$.

Suppose the sets of "acceptable" states and "bad" states of the machine are $A = \{0, 1, 2\}$ and $B = \{\ell - 2, \ell - 1, \ell\}$, respectively (assume $\ell \ge 5$). Then the preceding examples yield the following properties. The rate at which the machine jumps from being acceptable to being bad is

$$\lambda(A, B) = \sum_{i\in A, j\in B} \pi_i p_{ij} = \sum_{i=0}^{2} \frac{c_i}{\sum_{k=0}^{\ell} c_k} \sum_{j=\ell-2}^{\ell} p_{ij}.$$

The rate at which the machine enters an acceptable state is

$$\lambda(A) = \sum_{i=3}^{\ell} \frac{c_i}{\sum_{k=0}^{\ell} c_k} \sum_{j=0}^{2} p_{ij}.$$

The average time the machine is in an acceptable state is

$$W(A) = \frac{1}{\lambda(A)} \sum_{k \in A} \pi_k = \frac{1 + c_1 + c_2}{\sum_{i=3}^{\ell} c_i \sum_{j=0}^{2} p_{ij}}.$$

1.14 Optimal Design of Markovian Systems

This section explains how one can use average cost formulas for Markov chains to determine optimal design parameters for systems.

Consider a Markov chain on a space S whose transition probabilities $p_{ij}(x)$ are a function of a variable x in a set. The x is a design parameter (a vector or any type of variable) that can be selected to minimize the cost of running the Markov chain. In particular, suppose, for each x that $p_{ij}(x)$ determines an ergodic Markov chain and denote its stationary distribution by $\pi_i(x)$, $i \in S$. Assume there is a cost $f(i, x)$ whenever the chain visits state i, and a cost $g(i, j, x)$ whenever the chain jumps from i to j. Then by Theorem 74 above, the average cost under the design variable x is

$$\phi(x) = \sum_i \pi_i(x)[f(i, x) + \sum_j p_{ij}(x) g(i, j, x)], \tag{1.58}$$

provided the sum is absolutely convergent. The aim is to find a value of x that minimizes this cost.

For instance, in the (s, S) inventory model in Example 17, one may want to find values of s and S that minimize the average cost.

In some cases, it might be convenient to first determine the stationary distributions $\pi(x)$ for each x and then minimize $\phi(x)$ by an appropriate algorithm. When there are only a small number of x-values, a total enumeration approach might be feasible.

Another more general approach is to determine the stationary distributions simultaneously with minimizing the cost by the following non-linear mathematical program:

$$\min_x \sum_i \pi_i(x)[f(i, x) + \sum_j p_{ij}(x) g(i, j, x)]$$

subject to

$$\pi_i(x) = \sum_j \pi_j(x) p_{ji}(x), \qquad i \in S,$$

$$\sum_i \pi_i(x) = 1, \qquad \pi_i(x) > 0,\ i \in S.$$

A procedure for solving this would depend on the structure of the problem.

The preceding is a generic design problem for a Markovian system. It is a "static" optimization problem in that an optimal x is selected once (at time 0) and the system is run indefinitely with x set at this value. This is different from a "dynamic" optimization problem in which one can vary a parameter x as the system evolves to minimize the cost. For instance, one might vary the service rate in a queueing system as the queue length varies. Such a problem is a Markov decision problem (or a stochastic control, or dynamic programming problem).

Example 85. Optimal Machine Replacement. Consider the machine deterioration model in Example 16 in which X_n denotes the state of deterioration of a machine (or equipment) at time n. The X_n is a Markov chain on $S = \{0, 1, \ldots, \ell\}$ with transition matrix

$$\mathsf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & . & . & . & . \\ 0 & p_{11} & p_{12} & p_{13} & . & . & . \\ 0 & 0 & 0 & p_{22} & p_{23} & . & \\ . & . & . & . & . & . & . \\ p_{\ell 0} & . & . & . & 0 & 0 & p_{\ell\ell} \end{bmatrix}$$

Now, suppose there is a cost $h(i)$ for operating the machine in state i and a cost C for replacing the machine, where

$$h(0) \le h(1) \le \cdots \le h(\ell) < C.$$

Because of these costs, it may be beneficial to replace the machine before it reaches state ℓ.

Accordingly, let us consider the following *control-limit policy*: Replace the machine in the next time period putting its state to 0 when and only when the machine's state equals or exceeds a state x (the control limit).

Under this policy, the transition probabilities of the chain are

$$p_{ij}(x) = \begin{cases} p_{ij} & \text{if } 0 \le i \le x-1 \\ \mathbf{1}(j=0) & \text{if } x \le i \le \ell. \end{cases}$$

This chain is clearly ergodic. By Example 57, its stationary distribution is

$$\pi_i(x) = \frac{c_i(x)}{\sum_{i=0}^{\ell} c_i(x)}, \qquad 0 \le i \le \ell,$$

but under the control, the $c_i(x)$ are given recursively by $c_0(x) = 1$ and

$$c_i(x) = \begin{cases} (1 - p_{ii})^{-1} \sum_{j=0}^{i-1} c_j(x) p_{ji}, & 1 \le i \le x - 1, \\ \sum_{j=0}^{x-1} c_j(x) p_{ji}, & x \le i \le \ell. \end{cases}$$

Now by (1.58), the average cost of the system is

$$\phi(x) = \sum_{i=0}^{\ell} \pi_i(x)[h(i) + C\mathbf{1}(x \le i \le \ell)].$$

For specific costs and transition probabilities, one can readily compute $\phi(x)$ from the preceding formulas for each $x = 1, \ldots, \ell$, and then select a control limit x that minimizes the cost.

Although this optimal control-limit policy is a static policy, it is often optimal in the dynamic sense. Indeed, assume that, for each k,

$$\sum_{i'=k}^{\ell} p_{ii'} \le \sum_{i'=k}^{\ell} p_{ji'}, \qquad \text{when } i \le j.$$

In other words, the deterioration is higher in worse states, which would generally be true. Under this monotonicity condition, it is known from the theory of Markov decision processes (e.g., see [90] and examples in [33, 76, 109]) that within the class of all dynamic machine replacement policies, there is a control-limit policy that is optimal. Thus, the optimal control-limit policy described above is also optimal in the dynamic sense.

1.15 Closed Network Model

We will now present two Markov chain models of networks in which discrete items move among nodes. This section describes a model for a closed network in which a fixed number of items move indefinitely among the nodes. The next section describes a model for an open network in which items enter from outside and move among the nodes for a while and eventually exit the network.

The notation here is different from above. A typical state i will now be denoted by a vector $x = (x_1, \ldots, x_m)$ with nonnegative integer entries, and transition and stationary probabilities p_{ij} and π_i will now be denoted by $p(x, y)$ and $\pi(x)$.

Consider a network in which ν items move in discrete time among m nodes (or processing stations) in the set $\mathbb{M} = \{1, \ldots, m\}$. The state of the network at time n is denoted by the random vector $X_n = (X_n^1, \ldots, X_n^m)$, where X_n^i denotes the number of items at node i, and $\sum_{i=1}^m X_n^i = \nu$. The state space is the set S of all vectors $x = (x_1, \ldots, x_m)$ with nonnegative integer-valued entries such that $\sum_{i=1}^m x_i = \nu$.

At each time period, exactly one item moves from its current node to another node or returns to the same node according to prescribed probabilities. Specifically, whenever the process is in a state x, one item is selected to move from node i with probability $p_i(x_i)$, and that item moves to a node j with a prescribed *routing probability* p_{ij}, for i, j in $\mathbb{M}$. In other words, one item moves from node i to node j with probability $p_i(x_i)p_{ij}$, and this movement is independent of the past history. These selection and routing probabilities are such that $p_i(0) = 0$, $p_i(k) > 0$ if $k > 0$; and

$$\sum_{i=1}^{m} p_i(x_i) = 1, \qquad \sum_{j=1}^{m} p_{ij} = 1, \qquad x \in S,\ i \in \mathbb{M}.$$

Under these assumptions, X_n is a Markov chain. Now, a typical transition is of the form $x \to x - e_i + e_j$ (one item moves from node i to node j) with probability $p_i(x_i)p_{ij}$. Here e_i is the ith unit vector with 1 in position i and 0 elsewhere. Therefore, the transition probabilities $p(x, y) = P\{X_1 = y | X_0 = x\}$ are

$$p(x,y) = \begin{cases} p_i(x_i)p_{ij} & \text{if } y = x - e_i + e_j \text{ for some } i, j \in \mathbb{M} \\ 0 & \text{otherwise.} \end{cases}$$

The communication properties of this chain are determined by the routing probabilities p_{ij}. Indeed, Exercise 40 shows that the chain is irreducible or aperiodic if and only if p_{ij} has these respective properties. One can envision these routing probabilities as defining a virtual or artificial Markov chain that depicts the state of a single item moving in $\mathbb{M}$ with a service time at each node being exactly one time unit. This virtual chain evolves like the network chain X_n with $\nu = 1$.

For simplicity, we will assume the routing probabilities p_{ij} are ergodic with stationary distribution w_i. That is, w_i satisfies the *traffic equations*

$$w_i = \sum_j w_j p_{ji}, \qquad i \in \mathbb{M}.$$

Exercise 40 ensures that X_n is ergodic when it is aperiodic, since its state space is finite. The stationary distribution of the network chain is as follows.[16]

Theorem 86. *The stationary distribution for the closed-network Markov chain X_n described above is*

$$\pi(x) = c f_1(x_1) \cdots f_m(x_m), \qquad x \in S, \tag{1.59}$$

where $f_i(x_i) = w_i^{x_i} \prod_{k=1}^{x_i} p_i(k)^{-1}$. The normalization constant c is given by

$$c^{-1} = \sum_{x \in S} f_1(x_1) \cdots f_m(x_m).$$

[16] Here and below, we use the convention that $\prod_{k=1}^{0}(\cdot) = 1$.

Proof. It suffices to show that π given by (1.59) satisfies the balance equations $\pi(x) = \sum_y \pi(y)p(y,x)$.

Since any transition into x has the form $x + e_i - e_j \to x$,

$$\sum_y \pi(y)p(y,x) = \sum_{i,j} \pi(x + e_i - e_j)p(x + e_i - e_j, x)\mathbf{1}(x_j > 0).$$

From the definitions of π and the transition probabilities, when $x_j > 0$,

$$\begin{aligned} \pi(x + e_i - e_j) &= \pi(x)\frac{w_i p_j(x_j)}{w_j p_i(x_i + 1)}, \\ p(x + e_i - e_j, x) &= p_i(x_i + 1)p_{ij}. \end{aligned}$$

Substituting these expressions in the preceding display yields

$$\sum_y \pi(y)p(y,x) = \pi(x)\sum_j p_j(x_j)\Big(w_j^{-1}\sum_i w_i p_{ij}\Big) = \pi(x).$$

The last equality follows since $\sum_j p_j(x_j) = 1$, and the stationary distribution w_i of the routing probabilities satisfies $w_j = \sum_i w_i p_{ij}$. Thus π satisfies the balance equations $\pi = \pi\mathsf{P}$, and hence it is the stationary distribution.

In the preceding result, the stationary distribution

$$\pi(x_1, \ldots, x_m) = c f_1(x_1) \cdots f_m(x_m)$$

is the "joint distribution" for the network process in equilibrium. This distribution provides expressions for many performance parameters of the network. To describe c and the marginal distributions of π, we will use the functions

$$f_I(k) = \sum_{\sum_{i\in I} x_i = k} \prod_{i\in I} f_i(x_i), \qquad 0 \le k \le \nu,\ I \subset \mathbb{M}.$$

This f_I is the *convolution*[17] of the functions f_i, $i \in I$.

First note that $c = f_{\mathbb{M}}(\nu)^{-1}$. Next, note that the equilibrium distribution of the number of items in a subset of nodes I is

$$\begin{aligned} \pi_I(k) &= \sum_{x\in S} \pi(x_1, \ldots, x_m)\mathbf{1}(\sum_{i\in I} x_i = k) \\ &= c f_I(k) f_{I^c}(\nu - k), \qquad 0 \le k \le \nu. \end{aligned}$$

In particular, if $I = \{i\}$, the ith "marginal distribution" of the number of items in node i is

$$\pi_i(k) = f_i(k) f_{\mathbb{M}\setminus\{i\}}(\nu - k)/f_{\mathbb{M}}(\nu), \qquad 0 \le k \le \nu.$$

[17] Properties of convolutions, which are not needed here, are in the Appendix.

A standard performance parameter for the network is the average time items wait at a node or in a set of nodes.

Example 87. Rates of Item Movements. We first consider the rate at which items move from node i to node j, which is

$$r(i,j) = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} \mathbf{1}(X_m = X_{m-1} - e_i + e_j). \tag{1.60}$$

By Corollary 79 and the definitions of $\pi(x)$ and $p(x,y)$,

$$\begin{aligned} r(i,j) &= \sum_{x\in S} \pi(x) p(x, x - e_i + e_j) \mathbf{1}(x_i \geq 1) \\ &= w_i p_{ij} \sum_{x\in S} \pi(x - e_i) \mathbf{1}(x_i \geq 1). \end{aligned}$$

Now the last sum equals $f_{\mathbb{M}}(\nu-1)/f_{\mathbb{M}}(\nu)$, the normalization constant for the closed network with ν items divided by the constant for the network with $\nu - 1$ items. Therefore

$$r(i,j) = \frac{f_{\mathbb{M}}(\nu-1)}{f_{\mathbb{M}}(\nu)} w_i p_{ij}.$$

Furthermore, the rate $r(I,J)$ at which items move from a set of nodes I to a set of nodes J is (1.60) summed over $i \in I$ and $j \in J$. Consequently,

$$r(I,J) = \frac{f_{\mathbb{M}}(\nu-1)}{f_{\mathbb{M}}(\nu)} \sum_{i\in I} w_i \sum_{j\in J} p_{ij}.$$

Example 88. Average Empty Times. Next, consider the average length of time that a set of nodes I is empty, which is the average time $W(S_I)$ that X_n spends in $S_I = \{x \in S : \sum_{i\in I} x_i = 0\}$. By Example 83,

$$W(S_I) = \frac{1}{\lambda(S_I^c, S_I)} \sum_{x\in S_I} \pi(x),$$

where

$$\begin{aligned} \lambda(S_I^c, S_I) &= \sum_{y\in S_I^c} \sum_{x\in S_I} \pi(y) p(y,x) \\ &= \sum_{x\in S_I} \sum_{j\in I^c} \sum_{i\in I} \pi(x - e_j + e_i) p(x - e_j + e_i, x) \\ &= \sum_{x\in S_I} \pi(x) \sum_{j\in I^c} p_j(x_j) w_j^{-1} \sum_{i\in I} w_i p_{ij}. \end{aligned}$$

Example 89. Equally-Likely Item Selection. Consider the closed network described above in which at each time period, any one of the ν items is randomly chosen (or is equally-likely) to move. That is, the probability that one of the x_i items at node i is selected is

$$p_i(x_i) = x_i/\nu.$$

In this case, the stationary distribution of the ergodic network Markov chain X_n is the *multinomial distribution*[18]

$$\pi(x) = \frac{\nu!}{x_1! \cdots x_m!} w_1^{x_1} \cdots w_m^{x_m}, \qquad x \in S.$$

This result says that the joint distribution of the quantities of items in the nodes in equilibrium is equal to the multinomial distribution of quantities of items in m boxes, when ν items are independently put in the m boxes with respective probabilities $w_1, \ldots, w_m$.

From the structure of the multinomial distribution, it follows that the equilibrium distribution of the quantity of items in a subset of nodes I is a binomial distribution with parameters ν and $p = \sum_{i \in I} w_i$. More generally, if I, J, K is a partition of $\mathbb{M}$, then the joint equilibrium distribution of the numbers of items in these subsets is the multinomial distribution

$$\pi_{I,J,K}(i,j,k) = \frac{\nu!}{i!j!k!} p_I^i p_J^j p_K^k, \qquad i + j + k = \nu,$$

where $p_I = \sum_{i \in I} w_i$, and p_J and p_K are defined similarly.

1.16 Open Network Model

We will now consider an open m-node network in which items occasionally enter the network from outside and move among the nodes, as in the closed network above, but eventually the items exit the network. Much of the notation will be similar to that in the last section. The state of the network is denoted by the random vector $X_n = (X_n^1, \ldots, X_n^m)$, where X_n^i denotes the number of items at node i at time n, and the state space is $S = \mathbb{Z}_+^m$.

For a typical state $x = (x_1, \ldots, x_m)$, an item at node i is selected to move (as in the closed network) with probability $p_i(x_i)$, and p_{ij} is the probability the item is routed to node j. In addition, $p_0(|x|)$ will denote the probability

[18] Here the normalization constant in (1.59) is $c = \nu!$ because of $w_1 + \cdots + w_m = 1$ and the multinomial formula

$$(a_1 + \cdots + a_m)^\nu = \sum_{x: \sum_i x_i = \nu} \frac{\nu!}{x_1! \cdots x_m!} a_1^{x_1} \cdots a_m^{x_m}.$$

that an item enters the network from outside, as a function of $|x| = \sum_{i=1}^{m} x_i$, the number of items in the network. Also, p_{0j} and p_{i0} will denote the respective probabilities that an item from outside is routed to node j and an item at node i exits the network (think of 0 as the "outside node"). These probabilities are such that

$$p_0(|x|) + \sum_{i=1}^{m} p_i(x_i) = 1, \qquad x \in S,$$
$$\sum_{j=0}^{m} p_{ij} = 1, \qquad 0 \le i \le m.$$

A network transition at each time period is triggered by exactly one of the following events:
• One item moves from outside the network to some node j in the network with probability $p_0(|x|)p_{0j}$.
• One item at some node i moves to a node j in the network with probability $p_i(x_i)p_{ij}$.
• One item at some node i exits the network with probability $p_i(x_i)p_{i0}$.

Under these assumptions, X_n is a Markov chain with transition probabilities

$$p(x,y) = \begin{cases} p_0(|x|)p_{0j} & \text{if } y = x + e_j \text{ for some } 1 \le j \le m \\ p_i(x_i)p_{ij} & \text{if } y = x - e_i + e_j \text{ for some } 1 \le i \le m,\ 0 \le j \le m \\ 0 & \text{otherwise.} \end{cases}$$

Here $e_0 = 0$.

As in the closed network, the single-item routing probabilities p_{ij} determine the communication properties of the open-network chain X_n. Assume the routing probabilities p_{ij} on $\{0, \ldots, m\}$ are irreducible and aperiodic. Let $w = (w_0, \ldots, w_m)$, with $w_0 = 1$, be an invariant measure of p_{ij}; that is,

$$w_i = \sum_{j=0}^{m} w_j p_{ji}, \qquad 0 \le i \le m.$$

Then an argument as in Exercise 40 proves that X_n is irreducible and aperiodic.

The state space may be finite or infinite. It is infinite if $p_0(k) > 0$ for each $k \ge 0$. On the other hand, assume that if $p_0(k) = 0$ for some k, then $p_0(\ell) = 0$, for $\ell \ge k$. Then the total quantity of items in the network cannot exceed $\max\{k : p_0(k) > 0\}$ and the state space is finite (assuming the initial quantity is below this maximum).

Theorem 90. *The open-network Markov chain X_n described above is ergodic if and only if the state space is finite or*

$$c^{-1} = \sum_{x \in S} \prod_{k=0}^{|x|-1} p_0(k) \prod_{i=1}^{m} f_i(x_i) < \infty,$$

where $f_i(x_i) = w_i^{x_i} \prod_{k=1}^{x_i} p_i(k)^{-1}$. *In that case, its stationary distribution is*

$$\pi(x) = c \prod_{k=0}^{|x|-1} p_0(k) \prod_{i=1}^{m} f_i(x_i), \qquad x \in S. \tag{1.61}$$

Proof. Similarly to the proof of Theorem 86, one can show that π given by (1.61) with $c = 1$ is an invariant measure for the chain; this is left as Exercise 41. This observation, in light of Theorem 54, proves the assertions.

Marginal distributions and rates of item movements for open networks are similar to those for closed networks. However, there are a few simplifications.

Example 91. Independent Quantities at Nodes. Suppose the entry probabilities for the open network are $p_0(k) = p$, $k \geq 0$. Then the joint distribution (1.61) is the product $\pi(x) = \prod_{i=1}^{m} \pi_i(x_i)$ of the marginal distributions

$$\pi_i(k) = c_i (p w_i)^k \prod_{\ell=1}^{k} p_i(\ell)^{-1}, \qquad k \geq 0,$$

where $c_i^{-1} = \sum_{k=0}^{\infty} (p w_i)^k \prod_{\ell=1}^{k} p_i(\ell)^{-1}$. So in equilibrium the state components $X_n^1, \ldots, X_n^m$ are independent for each fixed n.

Also, similarly to Example 87, the rate at which items move from node i to node j is

$$r(i,j) = w_i p_{ij}, \qquad 0 \leq i, j \leq m.$$

1.17 Reversible Markov Chains

We will now study an important class of "reversible" Markov chains. An ergodic Markov chain is reversible if its rate of transitions from one state to another is equal to the rate of the transitions in the reverse order. When the ergodic chain is stationary, the reversibility condition is equivalent to a "time-reversibility" property that, at any instant, the future of the process is stochastically indistinguishable from viewing the process in reverse time. A remarkable feature of an ergodic reversible Markov chain is that its equilibrium distribution has a known universal product-form (a product of a ratios of transition probabilities). Further properties of reversible Markov chains in continuous time are covered in Chapter 4.

Throughout this section, X_n will denote an irreducible Markov chain on S with transition probabilities p_{ij}.

Definition 92. The Markov chain X_n is *reversible* if there is a measure η on S that satisfies the *detailed balance equations*

$$\eta_i p_{ij} = \eta_j p_{ji}, \quad i \neq j \text{ in } S. \tag{1.62}$$

We also say that X_n (or p_{ij}) *is reversible with respect to* η.

The η in this definition is an invariant measure for the chain, since summing (1.62) over j yields $\eta_i = \sum_j \eta_j p_{ji}$. When η is finite, it can be normalized to be the stationary distribution of the chain. This definition of reversibility is related to the property of a chain being "reversible in time"; see Exercise 67.

Note that if X_n is reversible, then it has the *two-way communication property*: for each $i \neq j$ in S, the probabilities p_{ij} and p_{ji} are both positive or both equal to 0. This property yields the criterion that a chain is "not" reversible if a transition from some i to j is possible, but the reverse transition is not possible. For instance, the success runs chain in Example 19 is not reversible since a transition from a state $i > 1$ to 0 is possible but the reverse transition is not possible. Because of the two-way communication property, a periodic Markov chain with period greater than 2 cannot be reversible; see Exercise 57.

A distinguishing characteristic of a reversible Markov chain is describable in terms of the rate of its transitions between sectors of its state space as follows; this is an immediate consequence of (1.62).

Remark 93. Suppose X_n has an invariant measure η. Then X_n is reversible with respect to η if and only if

$$\sum_{i\in A, j\in B} \eta_i p_{ij} = \sum_{j\in A, i\in B} \eta_j p_{ji}, \qquad A, B \subset S.$$

That is $\lambda(A, B) = \lambda(B, A)$, using the rate notation of Example 80. Therefore, when the chain is ergodic, the rate of transitions between any two sets is equal to the rate of the reverse transitions.

A quintessential reversible chain is as follows.

Example 94. Random Walk. Consider a random walk X_n on $S = \{0, 1, \ldots\}$ with transition probabilities

$$\mathsf{P} = \begin{bmatrix} r_0 & p_0 & 0 & \cdots\cdots \\ q_1 & r_1 & p_1 & 0 & \cdots \\ 0 & q_2 & r_2 & p_2 & 0 \cdots \\ \cdots\cdots\cdots\cdots\cdots \end{bmatrix}$$

Assume the p_i and q_i are positive. Then clearly this Markov chain is irreducible. Now, its detailed balance equations (1.62) (for the two respective cases $j = i + 1$, and $j = i - 1$) are

$$\eta_i p_i = \eta_{i+1} q_{i+1}, \quad i \geq 0, \qquad \eta_i q_i = \eta_{i-1} p_{i-1}, \quad i \geq 1.$$

Note that these equations are the same. The last one is the recursive equation $\eta_i = \eta_{i-1} p_{i-1}/q_i$. Iterating this backward, we arrive at the solution

$$\eta_i = \eta_0 \prod_{j=1}^{i} p_{j-1}/q_j, \qquad i \geq 1, \tag{1.63}$$

where η_0 is any positive number. Consequently, the chain is reversible and this η is a positive invariant measure. Furthermore, this measure is finite if and only if $\gamma = \sum_{i=1}^{\infty} \prod_{j=1}^{i} p_{j-1}/q_j$ is finite.

Thus, the chain is ergodic if and only if $\gamma < \infty$. In that case, η given by (1.63) is the stationary distribution for the chain, where $\eta_0 = (1+\gamma)^{-1}$.

Now, consider the chain restricted to $S = \{0, 1, \ldots, m\}$ so that its transition matrix is

$$\mathsf{P} = \begin{bmatrix} r_0 & p_0 & 0 & \cdots & \cdots & \cdots \\ q_1 & r_1 & p_1 & 0 & \cdots & \cdots \\ 0 & q_2 & r_2 & p_2 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & q_{m-1} & r_{m-1} & p_{m-1} & \\ \cdots & \cdots & \cdots & q_m & r_m & \end{bmatrix}$$

The detailed balance equations for this chain are the same as those above, and hence it is reversible and its stationary distribution is as above with $\gamma = \sum_{i=1}^{m} \prod_{j=1}^{i} p_{j-1}/q_j$. We will see later in Chapter 4 that such restrictions of reversible Markov chains to smaller subspaces are also reversible and their stationary distributions are simply restrictions of the stationary distributions of the original chains.

Exercises 50 and 52 discuss related queueing models.

The most remarkable property of a reversible Markov chain is that an invariant measure for it is automatically given by expression (1.64) below, which is a product of ratios of the transition probabilities (we saw this in (1.63) above). In this result, a sequence of states $i_0, \ldots, i_n$ in S is a *path* from i_0 to i_n if $p_{i_{k-1}, i_k} > 0$, $k = 1, \ldots, n$.

Theorem 95. *If the Markov chain X_n is reversible, then an invariant measure for it is $\eta_{i_0} = 1$ and*

$$\eta_i = \prod_{k=1}^{\ell} \frac{p_{i_{k-1}, i_k}}{p_{i_k, i_{k-1}}}, \qquad i \in S \backslash \{i_0\}, \tag{1.64}$$

where i_0 is a fixed state and $i_0, i_1, \ldots, i_\ell = i$ is any path from i_0 to i.

A proof of Theorem 95 is contained below in the proof of (c) implies (a) in Theorem 97.

Remark 96. One can construct the invariant measure η in (1.64) by the following recursion. Let $S_0 = \{i_0\}$ and

$$S_{n+1} = \{i \in S \backslash S_n : p_{ij} > 0 \text{ for some } j \in S_n\}.$$

Then set $\eta_{i_0} \equiv 1$ and, for each $n \geq 1$, define

$$\eta_i = \frac{p_{ji}}{p_{ij}} \eta_j, \quad \text{for } i \in S_{n+1} \backslash S_n \text{ and any } j \in S_n \text{ with } p_{ji} > 0.$$

The "universal" invariant measure (1.64) of a reversible Markov chain is based on the following Kolmogorov criterion for reversibility. Statement (c) is a "ratio form" of Kolmogorov's criterion, which justifies that (1.64) is the same for any path from i_0 to i. With no loss in generality, we assume that the Markov chain X_n has the two-way communication property.

Theorem 97. *The following statements are equivalent.*
(a) *The Markov chain X_n is reversible.*
(b) (Kolmogorov Criterion) *For each $i_0, \ldots, i_\ell$ in S with $i_\ell = i_0$,*

$$\prod_{k=1}^{\ell} p_{i_{k-1}, i_k} = \prod_{k=1}^{\ell} p_{i_k, i_{k-1}}.$$

(c) *For each path $i_0, \ldots, i_\ell$ in S, the product $\prod_{k=1}^{\ell} \frac{p_{i_{k-1}, i_k}}{p_{i_k, i_{k-1}}}$ depends on $i_0, \ldots, i_\ell$ and ℓ only through i_0 and i_ℓ.*

Proof. (a) $\Rightarrow$ (b). If X_n is reversible with respect to η, then, for each $i_0, \ldots, i_\ell$ in S with $i_\ell = i_0$,

$$\prod_{k=1}^{\ell} \eta_{i_{k-1}} p_{i_{k-1}, i_k} = \prod_{k=1}^{\ell} \eta_{i_k} p_{i_k, i_{k-1}}.$$

Canceling the η's yields (b).
(b) $\Rightarrow$ (c). To prove (c), it suffices to show

$$\prod_{k=1}^{\ell} \frac{p_{i_{k-1}, i_k}}{p_{i_k, i_{k-1}}} = \prod_{k=1}^{m} \frac{p_{j_{k-1}, j_k}}{p_{j_k, j_{k-1}}}, \tag{1.65}$$

where $i_0, \ldots, i_\ell$ and $j_0, \ldots, j_m$ are two paths with $i_0 = j_0$ and $i_\ell = j_m$. Since $i_0, \ldots, i_\ell, j_{m-1}, \ldots, j_1, j_0$ is a path from i_0 to itself, (b) implies

$$\prod_{k=1}^{\ell} p_{i_{k-1}, i_k} \prod_{k=1}^{m} p_{j_k, j_{k-1}} = \prod_{k=1}^{m} p_{j_{k-1}, j_k} \prod_{k=1}^{\ell} p_{i_k, i_{k-1}}.$$

These quantities are positive, by the definition of a path. Then dividing both sides of this equation by the second and fourth products yields (1.65).
(c) $\Rightarrow$ (a). Suppose (c) holds. We will show that X_n is reversible with respect to η defined by (1.64). For a fixed i, let $i_0, \ldots, i_\ell = i$ be a path from i_0 to i.

Choose any j such that $p_{ij} > 0$. Then

$$\begin{aligned}\eta_i p_{ij} &= p_{ji} \prod_{k=1}^{\ell} \frac{p_{i_{k-1},i_k}}{p_{i_k,i_{k-1}}} \frac{p_{ij}}{p_{ji}} \\ &= p_{ji}\eta_j.\end{aligned}$$

These detailed balance equations also hold trivially for i, j such that $p_{ij} = p_{ji} = 0$. Thus X_n is reversible with respect to η.

To verify the Kolmogorov criterion (b), or its ratio analogue (c), one may not have to consider all possible sequences or paths in S. In many instances, certain properties of p_{ij} and S lead to simpler versions of the Kolmogorov criterion. In particular, for some processes on vector state spaces only a small family of paths generated by the basis vectors need be considered. Here is another type of simplification.

Remark 98. The Kolmogorov criterion holds for all paths, if it holds for paths consisting of distinct states (aside from the same beginning and end states). This is because any path can be partitioned into subpaths of distinct states.

Reversibility of the Markov chain X_n can sometimes be recognized from the form of its *communication graph*. This is an undirected graph whose set of vertices is the state space S, and there is an edge linking a pair i, j if either p_{ij} or p_{ji} is not 0. The graph is connected when X_n is irreducible (which we have assumed).

Theorem 99. *If the Markov chain X_n has an invariant measure η, and its communication graph is a tree, then it is reversible with respect to η.*

Proof. Suppose i, j are vertices in the communication graph that are linked by an edge. Let A_i be the set of all the states in S reachable from i if the edge between i and j were deleted. Since the graph is a tree, it follows by the definition of A_i that

$$\eta_i p_{ij} = \lambda(A_i, A_i^c), \qquad \eta_j p_{ji} = \lambda(A_i^c, A_i),$$

where $\lambda(A, B) = \sum_{k\in A, \ell\in B} \eta_k p_{k\ell}$. Furthermore, we know from Example 81 that $\lambda(A, A^c) = \lambda(A^c, A)$, for any $A \subset S$. Therefore, the terms in the preceding display are equal, and so η satisfies the detailed balance equations, which proves the assertion.

Note that the communication graph of the random walk Example 94 is a tree, but there are many reversible processes whose communication graphs are not trees.

Example 100. Random Walk on a Circle. Suppose the Markov chain X_n takes values on the set of states $S = \{0, 1, \ldots, \ell\}$ arranged clockwise on a circle. From state i the chain can move to state $i + 1$ with probability p_i and move

to state $i-1$ with probability $q_i = 1-p_i$, where $\ell+1 = 0$ and $0-1 = \ell$. This circular random walk may not be reversible like the standard random walk in Example 94. Note that its communication graph is a circle (not a tree). Consequently, a path of distinct states from any state back to itself consists of all the states. In this case, the Kolmogorov criterion for reversibility is

$$p_0 \cdots p_\ell = q_0 \cdots q_\ell. \tag{1.66}$$

In other words, the chain is reversible if and only if (1.66) holds. In this case, the stationary distribution given by (1.64) is

$$\pi_i = \pi_0 \prod_{k=1}^{i} p_{k-1}/q_k, \qquad 1 \le i \le \ell,$$

where $\pi_0^{-1} = 1 + \sum_{i=1}^{\ell} \prod_{k=1}^{i} p_{k-1}/q_k$.

When this chain is not reversible, its stationary distribution is still tractable; see Exercise 56. Even if the p_i are all the same, the chain may not be reversible.

The following is another example in which the Kolmogorov ratio criterion simplifies considerably.

Example 101. McCabe's Library. Consider a finite collection of books (or items or data) labeled $1, \ldots, m$ that are placed in a row on a bookshelf. The successive book selections (one-at-a-time) by users are independent, and each user selects book b with probability p_b. When a book at location 1 is selected, it is returned to that location. When a book at location $k \ge 2$ is selected, it is returned to location $k-1$, and the book there is placed in location k. This rearrangement is done before the next book is selected. The state of the library at any selection is a vector $\mathbf{i} = (i_1, \ldots, i_m)$, where i_k denotes the book at location k. Then the state of the library at successive book selections is a Markov chain X_n on the set S of all $m!$ permutations of the books $(1, \ldots, m)$. Its transition probabilities are

$$p_{\mathbf{ij}} = \begin{cases} p_{i_k} & \text{if the book } i_k \text{ at location } k \text{ is selected, for some } k, \\ 0 & \text{otherwise.} \end{cases}$$

In the first line, $\mathbf{j}$ is the vector obtained from $\mathbf{i}$ after selecting book i_k at location k.

We will show the Markov chain is reversible by applying the Kolmogorov ratio criterion. Consider any path $\mathbf{i}^0, \ldots, \mathbf{i}^\ell$ of distinct states. Let $b_1, \ldots, b_\ell$ denote the book selections that determine this path starting with $\mathbf{i}^0$. Now, for the reverse path $\mathbf{i}^\ell, \ldots, \mathbf{i}^0$, let $b'_1, \ldots, b'_\ell$ denote the book selections that determine this reverse path starting with $\mathbf{i}^\ell$. Then the Kolmogorov ratio is

$$\prod_{n=1}^{\ell} \frac{p_{\mathbf{i}^{n-1},\mathbf{i}^n}}{p_{\mathbf{i}^n,\mathbf{i}^{n-1}}} = \prod_{n=1}^{\ell} \frac{p_{b_n}}{p_{b'_n}}.$$

This product simplifies as follows. To move from $\mathbf{i}^0$ to $\mathbf{i}^\ell$, each book i_k^ℓ with $(i_k^\ell < i_k^0)$ has to be selected at least $i_k^0 - i_k^\ell$ times; and after the $(i_k^0 - i_k^\ell)$-th selection, each subsequent b_n book selection has to be compensated by the associated book b'_n. Similarly, to move in reverse from $\mathbf{i}^\ell$ to $\mathbf{i}^0$, each book i_k^0 with $(i_k^0 < i_k^\ell)$ has to be selected at least $i_k^\ell - i_k^0$ times; and after the $(i_k^\ell - i_k^0)$-th selection, each subsequent b'_n selection has to be compensated by the associated b_n. Consequently,

$$\prod_{n=1}^{\ell} \frac{p_{\mathbf{i}^{n-1},\mathbf{i}^n}}{p_{\mathbf{i}^n,\mathbf{i}^{n-1}}} = \prod_{k=1}^{m} p_{i_k^\ell}^{(i_k^0 - i_k^\ell)}. \tag{1.67}$$

Here is a typical path and its reverse path, with the book selections above the arrows.

$$(1,2,3,4) \xrightarrow{3} (1,3,2,4) \xrightarrow{3} (3,1,2,4) \xrightarrow{2} (3,2,1,4) \xrightarrow{4} (3,2,4,1)$$

$$(1,2,3,4) \xleftarrow{2} (1,3,2,4) \xleftarrow{1} (3,1,2,4) \xleftarrow{1} (3,2,1,4) \xleftarrow{1} (3,2,4,1)$$

Then the product (1.67), with $(i_k^0 - i_k^\ell : 1 \le k \le 4) = (-3, 0, 2, 1)$, is

$$\frac{p_3^2 p_2 p_4}{p_1^3 p_2} = p_1^{-3} p_3^2 p_4.$$

Note that the product (1.67) does not depend on the interior states $\mathbf{i}^1, \ldots, \mathbf{i}^{\ell-1}$ of the path. Then by Theorem 97, the Markov chain X_n is reversible, and an invariant measure for it is the product (1.67) evaluated at $\mathbf{i}^\ell = \mathbf{i}$, where $\mathbf{i}^0$ is fixed. Therefore, setting $\mathbf{i}^0 = (1, \ldots, m)$, the stationary distribution is

$$\pi_{\mathbf{i}} = c \prod_{k=1}^{m} p_{i_k}^{(k-i_k)}, \quad \mathbf{i} \in S, \tag{1.68}$$

where c is the normalization constant under which this is a distribution.

A similar model applies for an infinite book collection. This model assumes the state space S is the set of all permutations of the books $(1, 2, \ldots)$ obtained by a finite number of book selections. The argument above yields the same invariant measure (1.68), where m is replaced by $\min\{k : \mathbf{i}_{k'} = k',\ k' > k\}$, the first location in $\mathbf{i}$ after which all the books are in their starting locations. See Exercise 58 for some details on these models.

1.18 Markov Chain Monte Carlo

This section describes a few standard Markov chains that are used in Monte Carlo simulations. The field of simulation and related statistics (e.g., see [43]) is very important for numerical explorations of systems, but our discussion will not go beyond an introduction to Markov chain procedures for estimation of system parameters.

There are a variety of statistical problems in which one uses Monte Carlo simulations for estimating expectations of the form

$$\mu = \sum_{i \in S} g(i)\pi_i,$$

where π is a specified probability measure and $g : S \to \mathbb{R}$. A standard Markov chain Monte Carlo approach is to construct an ergodic Markov transition matrix whose stationary distribution is π. Then for a sample path $X_1, \ldots, X_n$ of the chain, an estimator for μ is

$$\hat{\mu}_n = n^{-1} \sum_{m=1}^{n} g(X_m).$$

By Theorem 74 for Markov chains, $\hat{\mu}_n \to \mu$ a.s., and so $\hat{\mu}_n$ is a consistent estimator[19] of μ. Other consistent estimators are given in Exercise 59.

In some applications, the target distribution has the form $\pi_i = c\eta_i$, where η is known, but the normalization constant c is unknown and is difficult to compute. In this case, one can obtain consistent estimators for c as well as μ as in Exercise 59.

For a particular application, one would simulate the Markov chain for a large number of steps n, and then take the resulting $\hat{\mu}_n$ as an approximate value of μ. An implementation of this procedure requires the formulation of an ergodic Markov chain whose stationary distribution is π. There are many ways this can be done.

One approach is to use a reversible Markov chain whose transition probabilities have the form

$$p_{ij} = r(i,j)/\pi_i, \qquad i,j \in S, \tag{1.69}$$

where $r(i,j) = r(j,i)$. This chain is clearly reversible with stationary distribution π, provided the $r(i,j)$ are chosen so that the chain is ergodic. Here are two standard examples.

Example 102. Hastings-Metropolis Markov Chain. Let π be a positive probability measure on S, and consider a Markov chain with transition probabilities

[19] A statistic $\widehat{\theta}_n$ that is a function of observed values $X_1, \ldots, X_n$, is a *consistent estimator* of a parameter θ if $\widehat{\theta}_n \to \theta$ a.s. as $n \to \infty$.

$$p_{ij} = \begin{cases} \gamma_{ij} \min\{1, \pi_j \gamma_{ji}/(\pi_i \gamma_{ij})\} & \text{if } j \neq i, \\ 1 - \sum_{k \neq i} p_{ik} & \text{if } j = i. \end{cases} \tag{1.70}$$

Here γ_{ij} are probabilities that one selects such that the chain is ergodic.

These transition probabilities may seem rather artificial, but they are amenable to easy simulations for obtaining estimates discussed above. Note that the Markov chain is reversible with stationary distribution π, since p_{ij} can be written as in (1.69) with $r(i,j) = \min\{\pi_i \gamma_{ij}, \pi_j \gamma_{ji}\}$.

An important feature of the p_{ij} is that they do not depend on the normalization constant c, where $\pi_i = c\eta_i$; it is often difficult to calculate c. Also, the probabilities simplify to $p_{ij} = \gamma_{ij} \min\{1, \pi_j/\pi_i\}$ when $\gamma_{ij} = \gamma_{ji}$, for each i and j.

A Hastings-Metropolis simulation generates transitions according to the probabilities (1.70) as follows. Whenever the chain is in state i, the next state is determined by the following steps.
(1) Select a state j with probability γ_{ij}.
(2) If state j is selected, then choose j as the next state with probability $\min\{1, \pi_j \gamma_{ji}/(\pi_i \gamma_{ij})\}$; and otherwise choose the current state i to be the next state.
These steps are repeated for successive transitions.

The preceding example is for a single- or multi-dimensional state space; the next example is for the latter case.

Example 103. Gibbs Sampler. Let π be a probability measure on a set S of vectors of the form $\mathbf{i} = (i_1, \ldots, i_m)$, and let $(Y_1, \ldots, Y_m)$ denote a random vector with probability measure π. Consider a Markov chain with transition probabilities

$$p_{\mathbf{ij}} = m^{-1} P\{Y_k = j_k | Y_\ell = i_\ell, \ell \neq k\}, \quad \text{if } j_\ell = i_\ell,\ \ell \neq k, \text{ for some } k. \tag{1.71}$$

Also, $p_{ij} = 0$ otherwise. An easy check shows that these probabilities are of the form (1.69), and hence the Markov chain is reversible with stationary distribution π.

A Gibbs simulation generates transitions according to (1.71) as follows. Whenever the chain is in state $\mathbf{i}$, the next state is determined by changing a single component of $\mathbf{i}$ by the following steps.
(1) Randomly select a component of $\mathbf{i}$ that is to be changed: component k is selected with probability $1/m$.
(2) For the component k to be changed, choose a value j_k with probability $P\{Y_k = j_k | Y_\ell = i_\ell, \ell \neq k\}$, and then take the new state $\mathbf{j}$ to be $\mathbf{i}$ with i_k changed to j_k.
These steps are repeated for successive transitions.

A variation of this simulation can be constructed by changing the components one at a time in a specified order, and repeating this indefinitely. The resulting Markov chain has the stationary distribution π, but it may not be reversible (which is not essential for estimations).

Hastings-Metropolis and Gibbs Sampler Markov chains are illustrated by the following example.

Example 104. Hastings-Metropolis Simulation. We will consider a probability measure π on the set of vectors $S = \{0, 1, \ldots, L\}^m$. Let $\mathcal{M}$ denote any family of subsets of $\{1, \ldots, m\}$, and, for $\mathbf{i} = (i_1, \ldots, i_m) \in S$ and $A \in \mathcal{M}$, let $\mathbf{i}_A = \sum_{k \in A} i_k$. Suppose that π has the form

$$\pi_{\mathbf{i}} = c \prod_{A \in \mathcal{M}} f_A(\mathbf{i}_A),$$

where f_A, $A \in \mathcal{M}$, are positive functions that are known. This type of distribution arises in stochastic networks, where $\mathbf{i}_A$ is the number of items at the nodes in A.

To construct a Hastings-Metropolis simulation, we only need to decide which transitions are to be feasible and to choose the transition probabilities $\gamma_{\mathbf{ij}}$. Assume the feasible transitions from a state $\mathbf{i}$ will be into the set

$$S(\mathbf{i}) = \{\mathbf{i} - e_k + e_\ell \in S : \ k \neq \ell \in \{0, 1, \ldots, m\}\}.$$

Recall that $\mathbf{i} - e_k + e_\ell$ is the vector $\mathbf{i}$ with one unit subtracted from i_k and one unit added to i_ℓ. The number of its elements is $|S(\mathbf{i})| \leq m(m+1)$. Next, assume $\gamma_{\mathbf{ij}} = 1/|S(\mathbf{i})|$, which means that each $\mathbf{j} \in S(\mathbf{i})$ is equally likely. Then the transition probabilities (1.70) are

$$p_{\mathbf{ij}} = \frac{1}{|S(\mathbf{i})|} \min\{1, \pi_{\mathbf{j}}/\pi_{\mathbf{i}}\}, \qquad \mathbf{j} \in S(\mathbf{i}).$$

Letting $\mathcal{M}_k = \{A \in \mathcal{M} : k \in A\}$, it follows by the definition of $\pi_{\mathbf{i}}$ that

$$\frac{\pi_{\mathbf{j}}}{\pi_{\mathbf{i}}} = \frac{\prod_{A \in \mathcal{M}_k \setminus \mathcal{M}_\ell} r_A(\mathbf{i}_A)}{\prod_{A' \in \mathcal{M}_\ell \setminus \mathcal{M}_k} r_{A'}(\mathbf{i}_{A'} + 1)}, \quad \text{for } \mathbf{j} = \mathbf{i} - e_k + e_\ell, \tag{1.72}$$

where $r_A(n) = f_A(n-1)/f_A(n)$ and $\mathcal{M}_0$ is the empty set. Assume the functions f_A are such that the Markov chain is ergodic. Then the simulation generates transitions as follows.

Whenever the chain is in state $\mathbf{i}$, the simulation obtains the next state by the following rules.
(1) Randomly select a state $\mathbf{j} \in S(\mathbf{i})$ with probability $1/|S(\mathbf{i})|$.
(2) If $\mathbf{j}$ is selected, then choose it as the next state with probability $\min\{1, \pi_{\mathbf{j}}/\pi_{\mathbf{i}}\}$, and otherwise, choose the current state $\mathbf{i}$ to be the next state.

Example 105. Gibbs Sampler Simulation. Suppose that $(Y_1, \ldots, Y_m)$ has the joint distribution π in the preceding example. Then a Gibbs simulation generates transitions as follows.

Whenever the chain is in state $\mathbf{i}$, the next state is determined by changing a single component of $\mathbf{i}$ by the following steps.

(1) Select a component k of $\mathbf{i}$ that is to be changed with probability $1/m$.
(2) For the component k to be changed, choose a value $j_k \in \{0, 1, \ldots, L\}$ with probability

$$P\{Y_k = j_k | Y_\ell = i_\ell, \ell \neq k\} = \frac{\prod_{A \in \mathcal{M}_k} f_A(\mathbf{i}_A + j_k - i_k)}{\sum_{j'_k=0}^{L} \prod_{A \in \mathcal{M}_k} f_A(\mathbf{i}_A + j'_k - i_k)},$$

where $j_\ell = i_\ell$, $\ell \neq k$, for some k. Then take the new state $\mathbf{j}$ to be $\mathbf{i}$ with i_k changed to j_k.

1.19 Markov Chains on Subspaces

This section discusses two types of Markov chains that are associated with observing a Markov chain on parts of its state space.

The first type of chain in the following example addresses the questions: If one observes a Markov chain only on a certain subset of states, is the observed sequence of states a Markov chain? If so, what are its transition probabilities, and does it inherit the stationary distribution of the parent chain?

Example 106. Markov Chain on a Subspace. Let X_n be an irreducible Markov chain on S. Suppose the hitting time of a fixed subset $S' \subset S$ is finite starting from at any state; then the successive times that it visits S' are finite. Let X'_n denote the state of the chain at its nth visit to S'. By the strong Markov property for X_n at the hitting times of S', it follows that X'_n is an irreducible Markov chain on S'. This chain is called the *restriction of* X_n to S'.

We will show that its transition probabilities are

$$p'_{ij} = p_{ij} + \sum_{k \notin S'} p_{ik} \gamma_{kj}, \tag{1.73}$$

where γ_{kj} is the probability the chain X_n beginning in state k eventually hits S' in state j (such hitting probabilities are described in Section 1.7). In addition, we show that if X_n is recurrent, then so is X'_n, and an invariant measure for it is an invariant measure for X_n restricted to S'.

To derive the transition probabilities, note that a transition of X'_n out of a state i consists of (instantaneously) selecting a sequence of states, say $k_1, k_2, \ldots, k_\ell$, by the probabilities $p_{ik_1}, p_{k_1,k_2}, \ldots, p_{k_\ell,j}$ until some $j \in S'$ is reached. Then, conditioning on the first selection, the transition probabilities of X'_n are

$$p'_{ij} = p_{ij} + \sum_{\ell=1}^{\infty} \sum_{k_1, \ldots, k_\ell \notin S'} p_{ik_1} p_{k_1,k_2} \cdots, p_{k_\ell,j}.$$

The last sum can also be written as in (1.73), which proves (1.73).

Clearly X'_n is recurrent when X_n is. Next, recall from Theorem 53 that an invariant measure for X_n is as follows: For a fixed $i \in S$,

$$\eta_j = E_i\Big[\sum_{n=0}^{\tau_i - 1} \mathbf{1}(X_n = j)\Big], \qquad j \in S. \tag{1.74}$$

The η_j is the expected number of visits X_n makes to state j in between visits to the fixed reference state i. However, for $i, j \in S'$, the η_j is also the expected number of visits X'_n makes to j between visits to i. Thus by Theorem 53, the η_j above for $i, j \in S'$ defines an invariant measure for X'_n.

The next type of Markov chain addresses the questions: Can the stationary distribution of a Markov chain be constructed by determining the stationary distributions of the chain restricted to certain subsets of the state space by pasting these distributions together? If so, is there a procedure for doing the pasting?

Example 107. Star-Like Collage of Subchains. Let X_n be an irreducible Markov chain on S. Suppose $S_0, S_1, S_2, \ldots$ is a countable partition of S such that whenever the chain is in any S_k, it can only take transitions into $S_0 \cup S_k$. The partition is star-like in that to move from one set in the partition to another the chain must go through S_0. Assume that the chain restricted to each set $S_0 \cup S_k$ is ergodic with stationary distribution p_i^k, $i \in S_0 \cup S_k$.

For simplicity, assume S_0 consists of the single state 0; see Exercise 60 for the case when S_0 is not a singleton. Under these conditions, a natural candidate for the stationary distribution of X_n is

$$\pi_i = \pi_0 c_k p_i^k, \qquad \text{if } i \in S_0 \cup S_k \text{ for some } k,$$

where π_0 and c_k are to be determined. This is a *collage* (or pasting together) of the stationary distributions p^k of the subchains.

First note that $c_k = 1/p_0^k$, since $\pi_0 = \pi_0 c_k p_0^k$ because $S_0 = \{0\}$. In addition, π_0 is determined by

$$1 = \sum_i \pi_i = \pi_0 + \pi_0 \sum_k c_k \sum_{i \in S_k} p_i^k.$$

Thus, we have the following result. If the preceding double sum is finite, then X_n is ergodic and its stationary distribution from above would be

$$\pi_i = \pi_0 p_i^k / p_0^k, \qquad \text{if } i \in S_k \text{ for some } k, \tag{1.75}$$

and $\pi_0 = \big(1 + \sum_k \sum_{i \in S_k} p_i^k / p_0^k\big)^{-1}$. An easy check shows that this distribution satisfies $\pi = \pi \mathsf{P}$.

Example 108. Dual Birth-Death Subprocesses. Suppose X_n takes values in the set of integers. Assume that in order for it to move between the positive

and negative integers it must pass through 0 and, it can enter 0 only from states 1 or -1. Furthermore, assume that the chain behaves like an ergodic birth-death process on the nonnegative integers with stationary distribution $p_i^1 = (1-\rho_1)\rho^{i-1}$, where $0 < \rho_1 > 1$. Similarly, the chains behavior on the nonpositive integers is that of a birth-death process with stationary distribution $p_i^2 = (1-\rho_2)\rho^{i-1}$. This process is an example of the model above in which the communication graph of the process is a star with center set $S_0 = \{0\}$ and point sets $S_1 = \{1, 2, \ldots\}$ and $S_2 = \{\ldots, -2, -1\}$. Under the assumptions, the stationary distribution of the process is

$$\pi_i = \begin{cases} \pi_0 \rho_1^{i-1}, & i \geq 1, \\ \pi_0 \rho_2^{i-1}, & i \leq 1, \end{cases}$$

where $\pi_0^{-1} = 1 + 1/(1-\rho_1) + 1/(1-\rho_2)$.

Example 109. Random Walks on Intersecting Circles. The random walk on discrete points on a circle described in Example 100 has a tractable stationary distribution, even when it is not reversible (Exercise 56). Consider the generalization of a random walk on several ellipses that have 0 as a common point. Then the stationary distribution of the random walk is a collage of the stationary distributions on the ellipses.

1.20 Limit Theorems via Coupling

The material in this section was used to classify states of a Markov chain (Theorem 37), and to prove the important property that a stationary distribution is a limiting distribution (Theorem 59).

The first result is that if two independent irreducible and recurrent Markov chains have the same transition probabilities, then the difference between their distributions converges to 0 as time tends to infinity. It is based on constructing Markov chains on a common probability space with certain properties. This is an example of *coupling*, which refers to constructing stochastic processes (not necessarily on the same probability space) in order to prove convergence in distribution, stochastic ordering, rates of convergence of probability measures, etc.

Theorem 110. *Suppose X_n and Y_n are independent, irreducible, aperiodic recurrent Markov chains on S with arbitrary initial distributions, but with the same transition probabilities. Then*

$$\sup_i |P\{X_n = i\} - P\{Y_n = i\}| \to 0, \quad \text{as } n \to \infty. \tag{1.76}$$

Proof. For a fixed state i_0, let $\tau = \min\{n \geq 1 : X_n = Y_n = i_0\}$. Clearly τ is a stopping time of the process $Z_n = (X_n, Y_n)$, $n \geq 0$. Under the hypotheses, Exercise 27 shows that Z_n is an irreducible recurrent Markov chain on S^2. Consequently, $\tau < \infty$ a.s.

Let p_{ij} denote the transition probabilities for the two Markov chains X_n and Y_n. To prove their distributions get close to each other, the key observation is that the two chains are both equal to i_0 at time τ and, at any time thereafter, they have the same (conditional) distribution, since the chains have the same transition probabilities. In particular, since $X_m = Y_m = i_0$ when $\tau = m$, it follows by the strong Markov property that, for $n > m$,

$$P\{X_n = i | \tau = m\} = p_{i_0,i}^{n-m} = P\{Y_n = i | \tau = m\}. \tag{1.77}$$

Next, define $X_n^* = X_n$ for $n \le \tau$ and $X_n^* = Y_n$ otherwise. Clearly the chain X_n^* is equal in distribution to the chain X_n, since both chains have the same transition probabilities and initial distribution. Using this fact and (1.77),

$$\begin{aligned} |P\{X_n = i\} - P\{Y_n = i\}| &= |P\{X_n^* = i\} - P\{Y_n = i\}| \\ &= |P\{X_n^* = i, \tau > n\} - P\{Y_n = i, \tau > n\}| \\ &\le 2P\{\tau > n\} \to 0. \end{aligned}$$

This proves (1.76).

Recall from Theorem 59 that the stationary distribution for an ergodic Markov chain is also its limiting distribution. This result follows from statement (1.78) below that the probability measure of an ergodic Markov chain converges in *total variation distance*[20] to its stationary probability measure. This mode of convergence, of course, implies convergence in distribution.

Theorem 111. (Limiting Distributions) *If X_n is an ergodic Markov chain with stationary distribution π, then*

$$\sup_i |P\{X_n = i\} - \pi_i| \to 0, \quad \text{as } n \to \infty. \tag{1.78}$$

Hence π is the limiting distribution of X_n.

Proof. Let Y_n be a Markov chain defined on the same probability space as X_n with the same transition probabilities as X_n. Suppose the two chains are independent and that Y_n is stationary. Then $P\{Y_n = i\} = \pi_i$. Thus (1.78) follows by Theorem 110.

Next, we use Theorem 110 to prove the following result, which is a restatement of Theorem 38. This characterization of null-recurrence was used for classifying states of a Markov chain.

Theorem 112. *For an irreducible Markov chain on S with transition probabilities p_{ij}, a recurrent state i is null-recurrent if and only if*

[20] The total variation distance between two probability measures P and P' on a space S is $d(P, P') = \sup_B |P(B) - P'(B)|$, where the supremum is over all sets B in the σ-field of S. When S is countable, $d(P, P') = 1/2 \sum_{i \in S} |P(i) - P'(i)|$. Probability measures P_n on S converge in total variation distance to a probability measure P if $d(P_n, P) \to 0$ as $n \to \infty$.

$$\lim_{n\to\infty} p_{ii}^n = 0. \tag{1.79}$$

In this case, $\lim_{n\to\infty} p_{ji}^n = 0$, *for* $j \in S$.

Proof. We will prove the equivalent statement that i is positive recurrent if and only if

$$\limsup_{n\to\infty} p_{ii}^n > 0. \tag{1.80}$$

First assume (1.80) holds. Then the diagonal selection principle[21] applied to the probabilities p_{ij}^n yields the existence of a strictly increasing subsequence n_m and constants γ_j such that $\gamma_i > 0$ and $\lim_{m\to\infty} p_{ij}^{n_m} = \gamma_j$, $j \in S$. Furthermore,

$$\lim_{m\to\infty} p_{kj}^{n_m} = \gamma_j, \qquad k \in S.$$

Indeed, applying Theorem 110 to the middle term in the following, we have

$$|p_{kj}^{n_m} - \gamma_j| \le |p_{kj}^{n_m} - p_{ij}^{n_m}| + |p_{ij}^{n_m} - \gamma_j| \to 0.$$

We will now show that γ is a finite invariant measure. The γ is finite, since Fatou's lemma for sums[22] yields

$$\sum_j \gamma_j = \sum_j \lim_{m\to\infty} p_{ij}^{n_m} \le \liminf_{m\to\infty} \sum_j p_{ij}^{n_m} = 1.$$

Next, by the Chapman-Kolmogorov equations,

$$p_{ij}^{n_m+1} = \sum_k p_{ik}^{n_m} p_{kj} = \sum_k p_{ik} p_{kj}^{n_m}.$$

Then applying Fatou's lemma to the first sum and the dominated convergence theorem to the second sum yields

$$\sum_k \gamma_k p_{kj} \le \sum_k p_{ik}\gamma_j = \gamma_j.$$

If this inequality is a strict inequality for some j's, then summing on j,

$$\sum_k \gamma_k = \sum_j \sum_k \gamma_k p_{kj} < \sum_j \gamma_j,$$

which is a contradiction. Thus, $\gamma_j = \sum_k \gamma_k p_{kj}$, and hence γ is an invariant measure.

---

[21] The *diagonal selection principle* for bounded real numbers $\{a_j(n) : j \in S, n \ge 1\}$ states that there exists a strictly increasing subsequence of integers n_m such that the limit $\lim_{m\to\infty} a_j(n_m)$ exists for each j.

[22] From Theorem 12 in the Appendix, $\sum_j \liminf_{n\to\infty} a_j(n) \le \liminf_{n\to\infty} \sum_j a_j(n)$, for $a_j(n)$ that are bounded from below.

Now γ is a multiple of the invariant measure η in Theorem 53. From (1.42), the mean time between entrances to state i is given by $\mu_i = \sum_k \eta_k$, and this, being a multiple of $\sum_k \gamma_k$, is finite. Thus i is positive recurrent.

The next step is to show that if i is positive recurrent then (1.80) holds. The positive recurrence of i implies (using the notation in the last paragraph) that $\mu_i = \sum_k \eta_k$ is finite. Then $\pi_j = \eta_j/\mu_i$ is a stationary distribution for the chain. Now if the chain is aperiodic, Theorem 110 with $P\{Y_0 = j\} = \pi_j$, $j \in S$, yields $p_{ii}^n \to \pi_i > 0$, which proves (1.80). Also, when the chain is periodic, (1.80) follows from a slight variation of Exercise 29.

Finally, assume (1.79) holds. Using the strong Markov property,

$$p_{ji}^n = \sum_{m=1}^{n} f_{ji}^m p_{ii}^{n-m}, \qquad i, j \in S.$$

Recall that f_{ji}^m is the probability that, starting at j, the first entrance of the chain to i is at time m. Then as $n \to \infty$, the dominated convergence theorem yields $p_{ji}^n \to 0$, since (1.79) ensures $f_{ji}^m p_{ii}^{n-m} \mathbf{1}(m \le n) \to 0$.

1.21 Criteria for Positive Recurrence

We know by Theorem 54 that an irreducible Markov chain is positive recurrent if and only if it has a stationary distribution. For a complicated Markov chain, it may not be feasible to obtain a closed-form expression for its invariant measure or stationary distribution. However, it is still of interest to establish that an irreducible Markov chain is positive recurrent. One criterion for this is that the time to return to a fixed state has a finite mean (recall Theorem 54). This section presents more general criteria for positive recurrence based on showing that the time to hit a finite subset of states (instead of a fixed state) has a finite mean.

As usual, let X_n denote a Markov chain on S with transition probabilities p_{ij}. Consider the hitting time $\tau_F = \min\{n \ge 1 : X_n \in F\}$ of $F \subset S$. We begin with a preliminary result.

Proposition 113. *If X_n is irreducible, and there is a finite set $F \subset S$ such that $b = \max_{j \in F} E_j[\tau_F] < \infty$, then X_n is positive recurrent.*

Proof. It suffices to show that a single state in F is positive recurrent. To this end, consider the restriction of the chain to F defined by $X_n' = X_{\tau_F(n)}$, where $\tau_F(n)$ is the nth time X_n visits F and X_n' is the state of the chain at that visit. Example 106 showed that X_n' is an irreducible Markov chain on F; and it is positive recurrent since F is finite.

Now, for a fixed $i \in F$, the hitting time $\tau_i = \min\{n \ge 1 : X_n = i\}$ has the form

$$\tau_i = \sum_{m=0}^{\infty} (\tau_F(m+1) - \tau_F(m)) \mathbf{1}(\tau_i' > m),$$

where $\tau_F(0) = 0$ and $\tau_i' = \min\{n \geq 1 : X_n' = i\}$. Using the strong Markov property of X_n at $\tau_F(m)$,

$$\begin{aligned} E_i[\tau_i] &= \sum_{m=0}^{\infty} E_i\Big[E\Big[(\tau_F(m+1) - \tau_F(m))\mathbf{1}(\tau_i' > m)\Big| X_0, X_1, \ldots, X_{\tau_F(m)}\Big]\Big] \\ &= \sum_{m=0}^{\infty} E_i\Big[E_{X_m'}[\tau_F]\mathbf{1}(\tau_i' > m)\Big]. \end{aligned}$$

The last line uses the pull-through formula (1.90) and the fact that the event $\{\tau_i' > m\} = \{X_{\tau_k} \neq i;\, k \leq \tau_F(m)\}$ is a function of $X_0, X_1, \ldots, X_{\tau_F(m)}$. The positive recurrence of X_n' ensures that $E_i[\tau_i'] < \infty$. Using this and the assumption $E_{X_m'}[\tau_F] \leq b$ in the preceding, we have

$$E_i[\tau_i] \leq b \sum_{m=0}^{\infty} P_i\{\tau_i' > m\} = bE_i[\tau_i'] < \infty.$$

Thus i is positive recurrent for X_n.

The next result shows that the finite mean-hitting-time condition in Proposition 113 is satisfied if there exist a function v and a set F that satisfy (1.82). The function v assigns a real number to each state, which provides an "artificial" v-ordering of the states. The assumption (1.82) is equivalent to the existence of an $\varepsilon > 0$ such that

$$E_i[v(X_1)] < v(i) - \varepsilon, \qquad i \in F^c. \tag{1.81}$$

That is, the mean v-order of the chain decreases at each of its jumps initiated outside the finite set F. This ensures (the main part of the proof) that

$$E_j[\tau_F] < \varepsilon^{-1} v(j) < \infty, \qquad j \in F^c.$$

Finding such a function v is often done by trial and error since there are no known procedures for constructing it. Sufficient conditions for $v(i) = i$ are in the follow-on example.

Theorem 114. (Foster's Criterion) *Suppose X_n is an irreducible Markov chain, and there are a function $v : S \to \mathbb{R}_+$ and a set F such that $E_i[v(X_1)] < \infty$, $i \in F$, and*

$$\sup_{i \in F^c} E_i[v(X_1) - v(X_0)] < 0. \tag{1.82}$$

Then $E_i[\tau_F] < \infty$, $i \in S$. Furthermore, X_n is positive recurrent if F is finite.

Proof. Let $\tau = \tau_F$. We first prove $E_j[\tau] < \infty$, $j \in F^c$. Fix $j \in F^c$ and let $\gamma_j(n) = E_j[v(X_n)1(\tau > n)]$. Conditioning on $H_n = (X_0, \ldots, X_n)$ and using $\{\tau > n+1\} \subseteq \{\tau > n\}$, we have

$$\gamma_j(n+1) \le E_j\Big[E_j[v(X_{n+1})1(\tau > n)|H_n]\Big].$$

Since $1(\tau > n)$ is a function of H_n, and $X_n \in F^c$ when $\tau > n$, it follows by (1.81) that there is an $\varepsilon > 0$ such that

$$\begin{aligned} E_j[v(X_{n+1})1(\tau > n)|H_n] &= 1(\tau > n)E_{X_n}[v(X_{n+1})] \\ &< 1(\tau > n)(v(X_n) - \varepsilon). \end{aligned}$$

Using this inequality in the preceding display yields

$$\gamma_j(n+1) < \gamma_j(n) - \varepsilon P_j\{\tau > n\}.$$

Iterating this inequality for $n, n-1, \ldots, 1$, we have

$$0 \le \gamma_j(n+1) < \gamma_j(0) - \varepsilon \sum_{k=0}^{n} P_j\{\tau > k\}.$$

Letting $n \to \infty$, the sum converges to $E_j[\tau]$, and therefore

$$\varepsilon E_j[\tau] < \gamma_j(0) = v(j) < \infty, \qquad j \in F^c. \tag{1.83}$$

The aim is to prove $E_i[\tau] < \infty$ for all $i \in S$, and so it remains to show $E_i[\tau] < \infty$, $i \in F$. But this follows since conditioning on X_1 and using (1.83),

$$E_i[\tau] = \sum_{j \in F} p_{ij} + \sum_{j \in F^c} p_{ij}(1 + E_j[\tau]) \le 1 + \varepsilon^{-1} \sum_{j \in F^c} p_{ij} v(j) < \infty.$$

Finally, if F is finite, then X_n is positive recurrent by Proposition 113.

Example 115. Pake's Criterion. An irreducible Markov chain X_n on the nonnegative integers S is positive recurrent if $E_i[X_1] < \infty$, $i \in S$, and

$$\limsup_{i \to \infty} E_i[X_1 - i] < 0. \tag{1.84}$$

This result is a special case of Foster's criterion with $v(i) = i$, since (1.84) implies that there is an i^* such that $\sup_{i \ge i^*} E_i[X_1 - i] < 0$.

Example 116. Fork-Join Processing System. Consider a fork-join network shown in Figure 1.4 that processes jobs as follows. Jobs arrive according to a Bernoulli process, where p is the probability of an arrival at any discrete time. Each job arriving to the system instantaneously splits into m tasks, which are simultaneously assigned to the m nodes for processing. The m nodes operate

like $M/M/1$ service systems in discrete time as in Example 21, but not more than one job can be completed in each time period. The service time at node i has a geometric distribution where q_i is the probability of a service completion in any discrete time. When all of its m tasks are finished, the job is completed and exits the system.

Fork-join networks are natural models for a variety of computer, telecommunications and manufacturing systems that involve parallel processing. For instance, a fork-join computer or telecommunications network typically represents the processing of computer programs, data packets, telephone calls, etc. that involve parallel multi-tasking and splitting and joining of information. A manufacturing fork-join network represents the assembly of a product or system that requires several parts processed simultaneously at separate work stations or plant locations. A supply chain fork-join network typically represents filling an order by obtaining several products simultaneously from vendors (or warehouses or manufacturing plants).

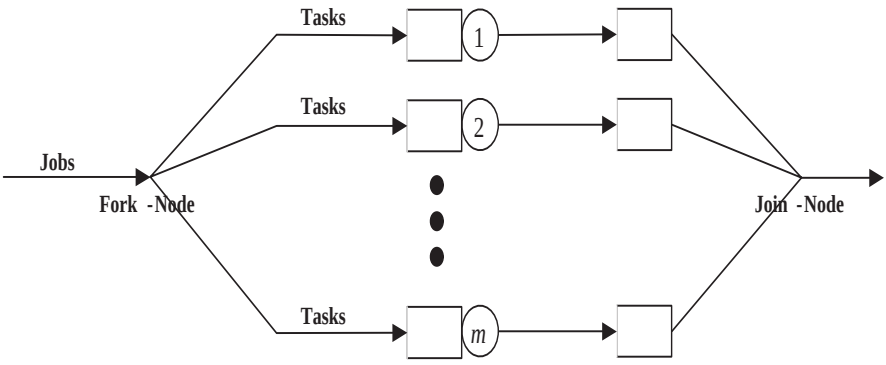

Fig. 1.4 Fork-Join Network.

The state of the fork-join network in Figure 1.4 at time n is represented by a vector-valued process X_n with states $x = (x_1, \ldots, x_m)$, where x_i denotes the quantity of tasks at node i. Under the assumptions above, X_n is a Markov chain with transition probabilities

$$p(x, x + (1, 1, \ldots, 1)) = p \prod_{\substack{j=1 \\ x_j > 0}}^{m} (1 - q_j)$$

$$p(x, x - e_i) = q_i \mathbf{1}(x_i > 0)(1 - p) \prod_{\substack{j=1, \\ j \neq i \\ x_j > 0}}^{m} (1 - q_j), \quad 1 \leq i \leq m.$$

Also, $p(x,x) = 1-$ the sum of the preceding probabilities, and $p(x,y) = 0$ elsewhere. Here e_i is the vector with 1 in position i and 0 elsewhere. This network chain X_n is one of those infamous queueing processes whose stationary distribution is intractable.

However, we can use Foster's criterion to prove that the chain is positive recurrent under the assumption

$$p < q_i, \qquad 1 \le i \le m. \tag{1.85}$$

That is, the arrival rate is less than the service rate at each node. Exercise 68 shows that (1.85) is a necessary as well as a sufficient condition for X_n to be positive recurrent.

A natural first guess is that Foster's criterion will work with the linear function $v(x) = \sum_{i=1}^m x_i$, but it does not as Exercise 68 verifies. A second choice is to consider the quadratic function $v(x) = \sum_{i=1}^m x_i^2$. Clearly,

$$E_x[v(X_1)] \le \sum_{i=1}^m (x_i+1)^2 < \infty, \qquad x \in S.$$

Next, for $x \ge (1,1,\ldots,1)$, consider

$$\begin{aligned} D(x) = E_x[v(X_1)] - v(x) &= \sum_{y \ne x} p(x,y)(v(y) - v(x)) \\ &= p(x, x+(1,1,\ldots,1)) \sum_{i=1}^m (2x_i+1) + \sum_{i=1}^m p(x, x-e_i)(-2x_i+1) \\ &= \prod_{j=1}^m (1-q_j) \sum_{i=1}^m \frac{g(x_i)}{1-q_i}, \end{aligned}$$

where $g(x_i) = 2x_i(p-q_i) + p(1-q_i) + q_i(1-p)$.

The aim is to find a finite set F for which $\sup_{x \in F^c} D(x) < 0$. Note that assumption (1.85) ensures that $D(x)$ is decreasing in x. Now, the smallest vector $b = (b_1, \ldots, b_m)$ for which $D(b) < 0$ is defined by

$$b_i = \min\{x_i \ge 1 : g(x_i) < 0\}.$$

Then setting $F = \{x : x_i < b_i, 1 \le i \le m\}$, it follows that

$$D(x) \le D(b) < 0, \qquad x \in F^c.$$

Thus the network chain X_n is positive recurrent by Foster's criterion.

1.22 Review of Conditional Probabilities

Applied probability involves extensive use of conditional probabilities and expectations. This section reviews these concepts for discrete random variables; analogous properties for continuous and general random variables are in the Appendix.

For this discussion, X and Y are discrete random variables that take values in countable sets S and S', respectively; and X and Y are defined on the same underlying probability space. For instance, X and Y could be discrete, real-valued random variables or vectors.

The *conditional probability measure* of Y given $X = x$, for $x \in S$, is

$$p(y|x) = P\{Y = y|X = x\} = \frac{P\{X = x, Y = y\}}{P\{X = x\}}, \quad y \in S',$$

provided $P\{X = x\} > 0$. This proviso will be assumed for all conditional probabilities without mention. For a real-valued Y ($S' \subset \mathbb{R}$), the *conditional expectation* (or mean) of Y given $X = x$ is

$$E[Y|X = x] = \sum_y yp(y|x), \qquad x \in S.$$

provided the sum is absolutely convergent. Unless specified otherwise, all the expectations in this section are assumed to be finite.

Consider a random variable of the form $g(Y)$, where $g : S' \to \mathbb{R}$. Then as above, the conditional expectation of $g(Y)$ given $X = x$ is

$$E[g(Y)|X = x] = \sum_y g(y)p(y|x), \qquad x \in S.$$

Conditional probabilities are often used to determine the distribution or mean of a random variable as follows. Suppose the issue is to find the distribution of Y or the mean of $g(Y)$, when the conditional probabilities $p(y|x)$ given $X = x$ are known and $p(x) = P\{X = x\}$ are also known. Using $P\{Y = y\} = \sum_x P\{X = x, Y = y\}$ and the definitions above,

$$P\{Y = y\} = \sum_x p(y|x)p(x), \quad E[g(Y)] = \sum_x \Big[\sum_y g(y)p(y|x)\Big]p(x). \quad (1.86)$$

Examples are given in Exercises 1 and 3.

We will often use the following standard shorthand notation for conditional probabilities and expectations. The conditional probability measure of Y given X (without specifying the value of X) is defined by

$$P\{Y = y|X\} = p(y|X) \qquad y \in S'.$$

Similarly, the conditional expectation of $g(Y)$ given X is

$$E[g(Y)|X] = h(X),$$

where $h(x) = E[g(Y)|X = x]$. That is, $E[Y|X] = \sum_y g(y)p(y|X)$. Note that these shorthand representations of conditional probabilities and expectations are random variables that are deterministic functions of X. With this notation, (1.86) becomes

$$P\{Y = y\} = E[P\{Y = y|X\}], \qquad E[g(Y)] = E[E[g(Y)|X].$$

The last formula is a special case of a result for general random variables as described in the Appendix. Namely, for a random variable X that takes values in a general space and a real-valued random variable Y,

$$E[Y] = E[E[Y|X]\,]. \tag{1.87}$$

This is a frequently-used formula for expectations.

Many properties of probabilities and expectations extend to conditional probabilities and expectations. For instance, suppose Z is an S''-valued random variable on the same probability space as X and Y. Then[23] for $h : S'' \to \mathbb{R}$,

$$\begin{aligned} E[g(Y) + h(Z)|X] &= E[g(Y)|X] + E[h(Z)|X], \\ E[g(Y)|X] &\le E[h(Z)|X] \quad \text{when } g(Y) \le h(Z). \end{aligned}$$

Another variation involves multiple conditioning, such as

$$E[h(Z)|X] = \sum_y E[h(Z)|X, Y = y]P\{Y = y|X\}.$$

An important point is that conditioning on $X = x$, allows one to replace X by x throughout in certain instances. In particular, from the definition of conditional expectation, for $G : S \times S' \to \mathbb{R}$,

$$E[G(X, Y)|X = x] = E[G(x, Y)|X = x], \qquad x \in S. \tag{1.88}$$

Moreover, $E[G(x, Y)|X = x] = E[G(x, Y)]$, when X and Y are independent. For instance, $E[X(Y - X)|X = x] = xE[(Y - x)|X = x]$. Similar expressions hold for conditional probabilities, such as

$$P\{X^2 + |Y - X| \le z|X = x\} = P\{x^2 + |Y - x| \le z|X = x\}.$$

In addition, for $H : S \times S' \to \mathbb{R}$,

[23] Statements like $g(Y) \le h(Z)$, $E[g(Y)|X] > V$ and $Y = 0$ hold a.s. but it is standard to suppress a.s.

$$E[G(X,Y)|H(X,Y)] \tag{1.89}$$
$$= \sum_x E[G(x,Y)|H(x,Y), X = x]P\{X = x|H(X,Y)\}.$$

Expression (1.88) also yields the *pull-through formula*: For $f : S \to \mathbb{R}$,

$$E[f(X)g(Y)|X] = f(X)E[g(Y)|X]. \tag{1.90}$$

That is, any function of the conditioning variable X can be pulled out of the expectation.

In some cases, we use conditioning statements in which only some of the variables are specified. For instance,

$$P\{Y = y|V, Z, X = x\} = h(V, Z, x),$$

where $h(v, z, x) = P\{Y = y|V = v, Z = z, X = x\}$.

Another important concept related to Markov chains is that of conditional independence. Recall that X and Y are *independent* if

$$P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}, \qquad x \in S, y \in S'.$$

Equivalently, for any $f : S \to \mathbb{R}$,

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$$

Analogously, Y and Z are *conditionally independent* given X if

$$P\{Y = y, Z = z|X\} = P\{Y = y|X\}P\{Z = z|X\}, \qquad y \in S',\ z \in S''.$$

Equivalently, $E[g(Y)h(Z)|X] = E[g(Y)|X]E[h(Z)|X]$ for any functions g and h for which the expectations exist. More generally, S'-valued random variables $Y_1, \ldots, Y_n$ are conditionally independent given X if

$$P\{Y_1 = y_1, \ldots, Y_n = y_n|X\} = \prod_{k=1}^{n} P\{Y_k = y_k|X\}.$$

Frequent use is made of the property that if Y is conditionally independent of $X_1, \ldots, X_{n-1}$ given X_n, then

$$P\{Y = y|X_1, \ldots, X_{n-1}, X_n = i\} = P\{Y = y|X_n = i\}.$$

1.23 Exercises

The first four exercises deal with conditional probabilities and expectations that are reviewed in Section 1.22.

Exercise 1. The length of time in microseconds for a computer to perform a task of type i has a geometric distribution $(1-p_i)p_i^{n-1}$, $n \geq 1$, with mean $1/(1-p_i)$, for $i = 1, \ldots, \ell$. Also, the type of task to be worked on is a random variable X, where $p(i) = P\{X = i\}$ is known for $i = 1, \ldots, m$. Under these assumptions, the time T to perform a task has the conditional probability measure $P\{T = n|X = i\} = (1-p_i)p_i^{n-1}$. Find expressions for $E[T|X = i]$, $P\{T = n\}$ and ET.

Exercise 2. Suppose $X_1, X_2, \ldots$ are real-valued i.i.d. random variables with mean μ and variance σ^2 that represent times to process jobs. The number of jobs to be processed in a week is a nonnegative integer-valued random variable N that is independent of the X_n. Prove that the mean and variance of the time to do the N jobs are

$$E[\sum_{n=1}^{N} X_n] = \mu E[N], \qquad \text{Var}[\textstyle\sum_{n=1}^{N} X_n] = E[N]\sigma^2 + \mu^2 \text{Var}[N].$$

Exercise 3. *Poisson Random Variable with a Randomized Mean.* The number of sales of a product in a period of length t has a Poisson probability measure $p_{\lambda t}(n) = (\lambda t)^n e^{-\lambda t}/n!$, $n \geq 0$, with mean λt. This is true when the sales over time occur according to a Poisson process with rate λ. Consider a variation of this setting in which the rate λ is a random variable Λ that may depend on the economic environment and other factors, but it is not affected by the sales. In this case, the number of sales N in the period of length t has the properties

$$P\{N = n|\Lambda = \lambda\} = p_{\lambda t}(n), \qquad E[N|\Lambda = \lambda] = \lambda t.$$

Letting $F_\Lambda(\lambda)$ denote the distribution of Λ, show that

$$E[N] = tE[\Lambda], \qquad P\{N = n\} = \int_{\mathbb{R}_+} p_{\lambda t}(n) F_\Lambda(d\lambda).$$

Exercise 4. *Continuation.* In the context of the preceding exercise, suppose the revenue from each sale is r, and the cost associated with making n sales is $\sum_{k=1}^{n} Y_k$, where Y_k is the cost to make the kth sale. Assume $Y_1, Y_2, \ldots$ are i.i.d. with mean μ and variance σ^2, and the Y_k are independent of Λ and the number of sales N. The net revenue from the N sales is

$$Z_N = \sum_{k=1}^{N} (r - Y_k).$$

Show that the mean and variance of this revenue are

$$E[Z_N] = tE[\Lambda](r-\mu), \qquad \text{Var}[Z_N] = tE[\Lambda][\sigma^2 + (r-\mu)^2].$$

You can use the results in Exercise 2. Recall that the mean and variance of the Poisson distribution $p_{\lambda t}$ are both λt.

Exercise 5. For a nonnegative integer-valued random variable N, prove

$$E[N] = \sum_{n=0}^{\infty} P\{N > n\}.$$

Exercise 6. *Geometric Sojourn Times in a State.* For a Markov chain X_n with $p_{ii} > 0$, the distribution of the sojourn time in state i is $P_i\{\tau_i = n\}$, where $\tau_i = \min\{n \geq 1 : X_n \neq i\}$. Show that

$$P_i\{\tau_i = n\} = (1-p_{ii})p_{ii}^{n-1}, \qquad n \geq 1,$$

which is a geometric distribution with parameter $1 - p_{ii}$. Use Exercise 5 to show $E[\tau_i] = 1/(1-p_{ii})$.

Exercise 7. *Geometric Memoryless Property.* Suppose X is a random variable with values in $\{1, 2, \ldots\}$. Show that X has a geometric distribution if and only if it satisfies the memoryless property

$$P\{X > n+1 | X > n\} = P\{X > 1\}, \qquad n \geq 0.$$

Hint: What is the unique solution of $f(n+1) = f(1)f(n)$, $n \geq 0$?
This memoryless property is analogous to the memoryless property of exponential random variables in Exercise 1 in Chapter 3.

Exercise 8. *Perishable Inventory or Perishable Service Model.* Quantities of a perishable resource arrive to a system in discrete time periods to satisfy demands. The resource is only available for a single time period, but unsatisfied demands are backlogged. For instance the resource might be food, blood or human organs that perish after a certain time period, or cargo space on an airline that disappears when the airline departs. This system could also be viewed as a queueing model in which potential services are occasionally available, but they are unused when there are no units waiting for services. Let U_n denote the quantity of the resource that is demanded in period n, and let V_n denote the quantity of the resource that arrives. Assume the pairs (U_n, V_n), for $n \geq 1$, are i.i.d. nonnegative integer-valued random variables. Let X_n denote the quantity of the resource that is backlogged at time n. Show that X_n is an input-output Markov chain process as in Example 23. Specify its transition probabilities when U_1 and V_1 are independent Poisson random variables with respective means α and β.

Exercise 9. *Moran Storage Model.* A reservoir with capacity c receives inputs $V_1, V_2, \ldots$ in discrete time periods, where V_n are i.i.d. nonnegative integer-valued random variables. In a period when the reservoir has room for y additional units (the reservoir level is $c - y$), and v units of input occur, then $(v - y)^+$ units of the input are discarded. In addition, the reservoir releases a non-random quantity of u units of water in each time period provided the reservoir level exceeds u, otherwise it releases all the water in the dam. Let X_n denote the amount of water in the reservoir at (the end of) time period n. Justify that X_n is a reflected random walk Markov chain and express its transition probabilities as a function of the distribution $q_n = P\{V_1 = n\}$.

Exercise 10. *Two-State Markov Chain.* Consider a machine (or a production system) that alternates, in discrete time, between being in operation (state 1) or being down for repair or reloading (state 2). The successive durations of time during which the machine is in operation are i.i.d. and the successive time durations the machine is down are i.i.d. and independent of the operation times. Let X_n denote the state of the machine at time n. Determine the type of distributions for the operation times and down times in order for X_n to be a Markov chain on $S = \{1, 2\}$ with transition matrix

$$\mathsf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix},$$

where all the entries are positive. Specify the meaning of the probability a, and specify the relation between a and the mean of a typical machine operation time. Show that

$$\mathsf{P}^n = \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b}\begin{bmatrix} a & -a \\ -b & b \end{bmatrix}.$$

Find $P\{X_n = 1 | X_0 = 1\}$, for $n \geq 1$. Establish that the chain is ergodic and its stationary distribution is $\pi_1 = b/(a+b)$ and $\pi_2 = a/(a+b)$. This two-state Markov chain can be used to model a variety of situations. Describe one.

Exercise 11. *Uniform Representation and Simulation of a Random Variable.* For a distribution function F, its left-continuous inverse is

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

Show that if U has a uniform distribution on the interval $[0, 1]$, then the random variable $X = F^{-1}(U)$ has the distribution F. (Use the fact that $F(x) \geq u$ if and only if $F^{-1}(u) \leq x$.)

Because of this property, one can generate a random sample of values from F by using the uniform distribution. First, one generates a random sample $U_1 = u_1, \ldots, U_n = u_n$ from the uniform distribution on $[0, 1]$. Then the values $x_1 = F^{-1}(u_1), \ldots, x_n = F^{-1}(u_n)$ form a random sample from F.

Exercise 12. *Non-homogeneous Markov Chains.* A stochastic process $\{X_n : n \geq 0\}$ on S is a non-homogeneous Markov chain if it satisfies the Markov property

$$P\{X_{n+1} = j | X_0, \ldots, X_{n-1}, X_n = i\} = P\{X_{n+1} = j | X_n = i\} = p_{ij}(n),$$

where the transition probabilities $p_{ij}(n)$ are functions of n. Show that the two-dimensional process $Y_n = (X_n, n)$ is a time-homogeneous Markov chain and specify its transition probabilities. Because of this formulation, some properties of non-homogeneous Markov chains (e.g., Markov decision problems) can be obtained from the theory of homogeneous Markov chains.

Exercise 13. *Continuation.* Suppose $\{X_n : n \geq 0\}$ is a stochastic process on S of the form $X_{n+1} = f_{n+1}(X_n, Y_{n+1})$, $n \geq 0$, where $Y_1, Y_2, \ldots$ are independent S'-valued random variables that are independent of X_0, and $f_n : S \times S' \to S$. Show that X_n is a non-time-homogeneous Markov chain with transition probabilities $p_{ij}(n) = P\{f_n(i, Y_n) = j\}$.

Exercise 14. *Unit-Demand Inventory System.* Consider an inventory or input-output system in discrete time, where X_n denotes the quantity of items in the system at the beginning of the nth period. At the beginning of each period, the inventory decreases by one unit provided the inventory level is positive, and otherwise the inventory remains at 0 until the end of the period. At the end of the nth period, the inventory is replenished by an amount V_n, where V_n are i.i.d. with distribution $p_i = P\{V_1 = i\}$, $i \geq 0$. Under these assumptions $X_{n+1} = (X_n - 1 + V_n)$ if $X_n > 0$ and $X_n = V_n$ if $X_n = 0$. Justify that X_n is a Markov chain and specify its matrix of transition probabilities. Is this Markov chain a special case of another one in this chapter?

Exercise 15. *Continuation: Unit-Supply Inventory System.* A dual of the model in the preceding exercise is an inventory system, where in each period, the inventory is replenished by one unit and it decreases by U_n (if possible). Then the inventory at the beginning of period $n+1$ is $X_{n+1} = (X_n - U_n + 1)^+$. Assume U_n are i.i.d. with distribution $p_i = P\{U_1 = i\}$, $i \geq 0$. Justify that X_n is a Markov chain and specify its matrix of transition probabilities.

Exercise 16. At a dock where trucks are unloaded one-at-a-time, it was observed that a small truck was followed by a large truck 10% of the time, and a large truck was followed by a small truck 60% of the time. Define a Markov chain model for representing the type of truck being unloaded and find the percentage of large trucks that are unloaded. Suppose the times to unload the small trucks are i.i.d. with mean μ and the times to unload the large trucks are i.i.d. with mean μ'. Find the percentage of unloading time devoted to large trucks.

Exercise 17. *Extreme-value Process.* Suppose Y_n are i.i.d. integer-valued random variables with distribution $p_k = P\{Y_1 = k\}$, where this probability is positive for some $k > 0$. Define $X_0 = 0$ and $X_n = \max\{Y_1, \ldots, Y_n\}$. Let $\tau_0 = 0$ and $\tau_{n+1} = \min\{m > \tau_n | X_m > X_{\tau_n}\}$, $n \geq 0$. Then $X'_n = X_{\tau_n}$ is the nth record value. Justify that X_n and X'_n are Markov chains and specify their transition probabilities. In addition, classify their states.

Exercise 18. The equality in distribution $X \stackrel{d}{=} Y$ of random variables is equivalent to $E[f(X)] = E[f(Y)]$ for any nonnegative function f on the set of values of the variables. This equivalence also holds for stochastic processes X and Y. Analogous equivalences hold for conditional expectations. In particular, for discrete random variables Y, Y^*, Z and Z^* in S,show that the following statements are equivalent:
(a) $P\{Y \in B|Z\} = P\{Y^* \in B|Z^*\}, \quad B \in S$
(b) $E[f(Y)|Z] = E[f(Y^*)|Z^*], \quad f : S^\infty \to \mathbb{R}$.
These statements are like (1.17) and (1.18).

Exercise 19. *Reflected Random Walks.* Show by induction or substitution that the solution to the recursion $X_n = a \vee [b \wedge (X_{n-1} + V_n - U_n)]$ in Example 23 is given by (1.10).

Exercise 20. Let $Y_1, \ldots, Y_{n+1}$ be random variables (possibly in a general space) such that $Y_1, \ldots, Y_n$ are i.i.d. Show that if Y_{n+1} is independent of $Y_1, \ldots, Y_n$ and $Y_{n+1} \stackrel{d}{=} Y_1$, then $Y_1, \ldots, Y_{n+1}$ are i.i.d.

Exercise 21. Suppose $\{X_n : n \geq 0\}$ is an S-valued stochastic process of the form $X_{n+1} = f(X_n, Y_{n+1})$, $n \geq 0$, where Y_n are S'-valued random variables such that, for each $n \geq 0$, Y_{n+1} is conditionally independent of $H_n \equiv \{X_{k-1}, Y_k, k \leq n\}$, and $G_i(A) = P\{Y_{n+1} \in A | X_n = i\}$ is independent of n. Show that X_n is a Markov chain and specify its transition probabilities.

Exercise 22. *Continuation.* Consider the processes X_n and Y_n in the preceding exercise with the history H_n changed to $H_n \equiv \{X_{k-1}, Y_{k+1}, k \leq n-1, X_{n-1}\}$ (it no longer contains Y_n). Is X_n is a Markov chain? If so specify its transition probabilities. Answer this under the additional assumption that $G_{i,i'}(A) = P\{Y_{n+1} \in A | X_n = i, Y_n = i'\}$ is independent of n.

Exercise 23. *Stopping Time Criteria.* Let τ be a random variable that takes values in $\{0, 1, \ldots, \infty\}$. Show that the following are equivalent statements:
(a) τ is a stopping time for a Markov chain X_n.
(b) $\{\tau > n\}$ is determined by $X_0, \ldots, X_n$ for any finite n.
(c) $\{\tau \leq n\}$ is determined by $X_0, \ldots, X_n$ for any finite n.
(d) $\mathbf{1}(\tau = n) = h_n(X_0, \ldots, X_n)$ for some $h_n : S^{n+1} \to \{0, 1\}$.

Exercise 24. *Communication is an Equivalence Relation.* Show that $i \to j$ in S if and only if there are states $i_1, \ldots, i_n \in S$ such that $p_{i,i_1} p_{i_1,i_2} \cdots p_{i_n,j} > 0$. Show that the communication relation $\leftrightarrow$ is an equivalence relation in that

it satisfies the following properties.
Reflexive: $i \leftrightarrow i$, $i \in S$.
Symmetric: $i \leftrightarrow j$ if and only if $j \leftrightarrow i$, $i, j \in S$.
Transitive: If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$, $i, j, k \in S$.

Exercise 25. Show that, for an irreducible class C, each of its states has the same period. Hint: Suppose $i, j \in C$ have periods d_i and d_j. Let m and n be the smallest integers such that $a = p_{ji}^m p_{ij}^n > 0$. Then use (1.27) along with $p_{ii}^{m+n} \geq p_{ij}^n p_{ji}^m = a$ to prove $d_i \leq d_j$. Then reverse the roles of i and j.

Exercise 26. Consider a Markov chain on $S = \{1, 2, \ldots, 10\}$ whose transition matrix has the following form, where $\star$ means a positive probability.

$$\mathsf{P} = \begin{bmatrix} \star & 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 \\ 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & \star & \star \\ \star & 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \star & \star & \star & 0 & \star & \star & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 & \star & 0 & 0 & 0 & 0 \\ 0 & \star & \star & 0 & \star & 0 & \star & \star & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \star & 0 & 0 \\ 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & 0 & \star \\ 0 & 0 & \star & 0 & 0 & 0 & 0 & 0 & \star & \star \end{bmatrix}$$

Draw the transition graph of this chain and identify its transient and closed irreducible sets. Display a more informative transition matrix for the chain by reordering the states according to Theorem 40.

Exercise 27. Suppose X_n and X'_n are independent ergodic Markov chains on the spaces S and S' with transition probabilities p_{ij} and p'_{ij} and stationary distributions π and π', respectively. Show that $Z_n = (X_n, X'_n)$ is an ergodic Markov chain on $S \times S'$ with transition probabilities $p_{ij,k\ell} = p_{ik} p'_{j\ell}$, and its stationary distribution is $\pi_{ij} = \pi_i \pi'_j$.

Exercise 28. Let X_n be an ergodic Markov chain on S with transition probabilities p_{ij} and stationary distribution π. Show that $\tilde{X}_n = (X_n, \ldots, X_{n+\ell})$ is an ergodic Markov chain on $S^{\ell+1}$ with stationary distribution

$$\pi(\mathbf{i}) = \pi_{i_0} p_{i_0, i_1} \cdots p_{i_{\ell-1}, i_\ell}.$$

Exercise 29. For an irreducible Markov chain X_n with period d, show that $X_n^* = X_{nd}$ is an irreducible aperiodic Markov chain and specify its transition probabilities. Show that $\tau_i = \min\{n \geq 1 : X_n = i\}$ is a multiple of $\tau_i^* = \min\{n \geq 1 : X_n^* = i\}$. Use Theorem 110 to prove that if i is positive recurrent, then $\lim_{n\to\infty} p_{ij}^{nd} = d/E_i[\tau_i]$. Use this result to prove (1.80).

Exercise 30. *Infinite Number of Stationary Distributions.* Consider a Markov chain, as in Theorem 40, whose state space S is the union of closed irreducible recurrent classes $C_1, \ldots, C_m$ with associated transition matrices $\mathsf{P}_1, \ldots, \mathsf{P}_m$, respectively. Suppose π_i^ℓ, $i \in C_\ell$, is a stationary distribution for P_ℓ ($\pi^\ell \mathsf{P}_\ell = \pi^\ell$), $\ell = 1, \ldots, m$. Show that, if α is an initial distribution for X_0, then $\pi_i = \sum_{\ell=1}^m \alpha_i \pi_i^\ell \mathbf{1}(i \in C_\ell)$ is a stationary distribution for the chain. In fact, all stationary distributions can be obtained this way.

Exercise 31. Using the dominated convergence for sums, show that a limiting distribution for a Markov chain (which need not be ergodic) is a stationary distribution. Give an example of a non-ergodic Markov chain with a stationary distribution that is not a limiting distribution.

Exercise 32. *Success Runs.* Let X_n denote the success runs Markov chain in Example 19. Find an expression for the probability f_{00} of ever reaching state 0 starting from 0. Show that the chain is irreducible, and it is recurrent if and only if $\sum_i (1 - p_i) = \infty$. In that case, show that an invariant measure for the chain is $\eta_i = \prod_{j=0}^{i-1} p_j$, $i \geq 0$.

Exercise 33. Suppose $Y_1, Y_2, \ldots$ are i.i.d. random variables with a finite mean μ. Show that $n^{-1} Y_1 \to 0$, and that $n^{-1} Y_n \to 0$, a.s. as $n \to \infty$. Hint: use $Y_n = \sum_{m=1}^n Y_m - \sum_{m=1}^{n-1} Y_m$. Next, suppose for some positive integer k that Y_m, $m > k$ are i.i.d. with mean μ, and $Y_1, \ldots, Y_k$ are general random variables with finite means (they need not be independent or independent of the other Y_m's). Show that $n^{-1} \sum_{m=1}^n Y_m \to \mu$ a.s. as $n \to \infty$. (This result also holds for a random k).

Exercise 34. *Limits of Expectations.* Let X_n be an ergodic Markov chain on S with stationary distribution π. Show that, for $B \subset S^{\ell+1}$,

$$\lim_{n \to \infty} P\{(X_n, \ldots, X_{n+\ell}) \in B\} = \sum_{\mathbf{i} \in B} p(\mathbf{i}),$$

where $\mathbf{i} = (i_0, \ldots, i_\ell)$ and $p(\mathbf{i}) = \pi_{i_0} p_{i_0, i_1} \cdots p_{i_{\ell-1}, i_\ell}$. Show that, for bounded $f : S^{\ell+1} \to \mathbb{R}$,

$$\lim_{n \to \infty} E[f(X_n, \ldots, X_{n+\ell})] = \sum_{\mathbf{i} \in S^{\ell+1}} f(\mathbf{i}) p(\mathbf{i}).$$

Exercise 35. *Brand Switching.* A Markov chain model for approximating the sales of several brands of a particular product that customers continually purchase (e.g., shampoo, soda, bread) is as follows. Suppose there are four brands of a product labeled $1, 2, 3, 4$, and the successive brands that a typical customer chooses to purchase over time is a Markov chain X_n with transition matrix

$$\mathsf{P} = \begin{bmatrix} .7 & .1 & .1 & .1 \\ .2 & .4 & .2 & .2 \\ 0 & .5 & .4 & .1 \\ 0 & 0 & .2 & .8 \end{bmatrix}$$

For instance, upon buying brand 3, a customer's next purchase is brand 2, 3 or 4 with respective probabilities .5, .4, .1, independent of past purchases. Show that the fraction of sales over time of brands 1 and 2 are 6/67 and 9/67, respectively. Suppose the profits from the four brands are \$10, \$12, \$15, \$16 per sale. Show that the average profit from the four brands is \$984/67 $\approx$ \$14.70 per sale, and the average profit from only brands 1 and 3 combined is \$300/67 $\approx$ \$4.48 per sale (of all brands). Is brand 4 more profitable per sale than brands 1 and 2 combined?

Exercise 36. *Setup Costs in Processing Systems.* As in Example 71, suppose the ergodic Markov chain X_n with stationary distribution π denotes a sequence of jobs a system processes (labeled by job type). Assume that whenever the system switches from processing a type i job to processing a type j job, the system incurs a setup cost $v(i,j)$ measured in time or money. Then the average setup cost per job processed is $\gamma = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} v(X_{n-1}, X_n)$. Find an expression for this average. Find an expression for the average setup cost $\gamma(i)$ per job processed for switching from a type i job to another type of job, where $\gamma = \sum_i \gamma(i)$.

Exercise 37. *Renewal Age Process.* Let X_n denote the renewal age process in Example 20, where the renewals times are i.i.d. machine lifetimes with distribution F and mean μ, and $F(1) > F(0) = 0$. Prove that X_n is ergodic and that its stationary distribution is $\pi_i = [1 - F(i)]/\mu$. Suppose there is a cost C each time the machine is replaced and a cost h_i for each period in which the age is i. Give an expression for the average cost for the chain.

Exercise 38. *Continuation.* Jobs are processed one at a time and the processing times are nonnegative integer-valued i.i.d. random variables with distribution F and mean μ. As in Example 20, let X_n denote the age of the job being processed at time n, and let X'_n denote the time needed to finish the job in process at time n. Show that X'_n is an ergodic Markov chain and that its stationary distribution is the same as that for X_n.

Exercise 39. Let X_n denote the number of visitors at a certain web site (e.g., a food recipe site) at time n. At each time period, each visitor at the site independently leaves with probability p, and the number of new visitors that enter the site has a Poisson distribution with mean λ (independent of everything else). Show that X_n is a Markov chain and determine its transition probabilities. Find a condition on p and λ under which the chain is ergodic. Show that its stationary distribution is a Poisson distribution and specify its mean.

Exercise 40. *Closed Network Process.* Show that the network chain X_n in Section 1.15 is irreducible or aperiodic or reversible if and only if the routing probabilities p_{ij} have these respective properties.

Exercise 41. *Open Network Process.* For the open network process described in Theorem 90, show that π given by (1.61) with $c = 1$ is an invariant measure for the network chain. Start with

$$\sum_y \pi(y)p(y,x) = \sum_{i=1}^m \sum_{j=1}^m \pi(x + e_i - e_j)p(x + e_i - e_j, x)\mathbf{1}(x_j > 0)$$
$$+ \sum_{i=1}^m \pi(x + e_i)p(x + e_i, x) + \sum_{j=1}^m \pi(x - e_j)p(x - e_j, x)\mathbf{1}(x_j > 0).$$

Exercise 42. *Doubly Stochastic Chains.* A Markov chain is doubly stochastic if its transition probabilities p_{ij} satisfy $\sum_i p_{ij} = 1$, for each j. Show that a doubly stochastic Markov chain on a state space with m states has a stationary distribution $\pi_i = 1/m$.

Exercise 43. *Ehrenfest Chain.* Consider a system in which ν particles move in discrete time between two locations 1 and 2. At each time period, a particle chosen at random moves from its location to the other one. Let X_n denote the number of particles in location 1 at time n. Justify that X_n is an irreducible Markov chain with period 2 and show that its stationary distribution is a binomial distribution.

Exercise 44. *Random Walk on a Graph.* Consider a connected graph with m nodes, and let G_i denote the set of nodes adjacent to node i. Whenever the particle is at node i, it moves to any $j \in G_i$ with equal probability (i.e., $1/|G_i|$). Let X_n denote the location of the particle at time n. Show that X_n is a Markov chain and specify its transition probabilities. Is the chain irreducible? aperiodic? recurrent? Assuming the graph is such that the chain is ergodic, show that its stationary distribution is $\pi_i = |G_i|/2|S|$.

Exercise 45. *Expected Length of Gambling Game.* In the Gambler's ruin model in Examples 4 and 44, let v_i denote the expected length of time until the gambler's fortune X_n reaches 0 or m, starting with the fortune i. Verify

$$v_i = \frac{m}{q-p}\Big[\frac{i}{m} - \frac{1-(q/p)^i}{1-(q/p)^m}\Big], \qquad \text{if } p \neq 1/2,$$

and $v_i = i(m-i)$, if $p = 1/2$. Find an expression for the mean number of times $\hat{v}_i$ the gambler's fortune equals i before it reaches 0 or m, when $X_0 = i$.

Exercise 46. *Continuation.* Consider the gambler's random walk in the preceding exercise in which there is no upper bound m at which the gambler stops playing. That is, the fortune X_n moves in $S = \{0, 1, \ldots\}$ until it is absorbed at 0 or remains positive forever. Let γ_i denote the probability of being absorbed at 0 when $X_0 = i$. Show that $\gamma_i = 1$ if $q \geq p$, and that the mean length of the game starting at i is $i/(q-p)$ if $q > p$. Thus a gambler is bound to loose in a series of gambles if the odds are not favorable.

In addition, show that $\gamma_i = (q/p)^i$ if $p > q$. Use the facts that γ_i must be bounded and $\sum_{i=1}^{\infty} \gamma_i = 1$.

Exercise 47. *Pattern Occurrences in Bernoulli Process.* Let Y_n denote a Bernoulli process with $P\{Y_n = 1\} = p$ and $P\{Y_n = 0\} = q = 1 - p$. Suppose one is interested in the occurrence of a three-letter pattern in the sequence Y_n. Let $X_n = Y_{n-2}Y_{n-1}Y_n$, for $n \geq 2$. Show that X_n is an ergodic Markov chain and find its stationary distribution.

In particular, consider the occurrence of the three-letter pattern 101 (various pattern occurrences like this are of interest in DNA sequences). Let $\tau = \inf\{n \geq 2 : X_n = 101\}$, and define $G_k(s) = E[s^\tau | X_0 \in A_k]$, where

$$A_0 = \{000, 100\}, \quad A_1 = \{001, 011, 111\}, \quad A_2 = \{110, 010\}, \quad A_3 = \{101\}.$$

Find an expression for $G_0(s)$. Begin by justifying that

$$G_0(s) = s[pG_1(s) + qG_0(s)], \quad G_1(s) = s[qG_2(s) + pG_0(s)]$$

and $G_2(s) = s[p + qG_0(s)]$, and then solve for $G_0(s)$. Find a tractable expression for $E[\tau | X_0 \in A_0]$.

Exercise 48. *Moving Averages.* Let X_n be an ergodic Markov chain with stationary distribution π. Consider the moving average process

$$Y_n = \sum_{m=-k}^{\ell} a_m f(X_{n+m}), \qquad \text{for } k, \ell \geq 0,\ a_m \in \mathbb{R}, \text{ and } f : S \to \mathbb{R}.$$

Specify the limiting distribution of Y_n for $k = 2 = m$, and give a formula for $\lim_{n\to\infty} E[Y_n]$.

Exercise 49. *Exhaustive Parallel Processing.* A set of m jobs are processed in parallel until they are all completed. At the completion time, another m jobs instantaneously enter the system and are processed similarly. This is repeated indefinitely. Assume the times to process jobs are independent with a geometric distribution with mean $1/p$. Let X_n denote the number of jobs being processed at time n. Show that X_n is a Markov chain on $S = \{1, \ldots, m\}$ and specify its transition probabilities (whenever all the jobs in the system are completed simultaneously, the next state is m). Show that the chain is ergodic and that an invariant distribution can be computed by the recursive formula

$$\pi_{m-k} = \frac{q^{m-k}}{1 - q^{m-k}} \sum_{j=1}^{k} \pi_{m-k-j} \binom{m-k-j}{j} p^j,$$

for $k = 1, \ldots, m$ starting with $\pi_m = 1$, where $q = 1 - p$. Use this formula (and successive substitutions) to derive an explicit expression for the stationary distribution when $m = 3$.

Exercise 50. *$M/M/1$ Queueing System.* Suppose X_n is the queue-length process in Example 21, where p is the probability of an arrival, and q is the probability of a service completion at any discrete time. Show that an invariant measure for this Markov chain is $\pi_i = \rho^i$, $i \geq 0$, where $\rho = p(1-q)/[q(1-p)]$. Prove that the chain is ergodic if and only if $p < q$, and in this case $\pi_i = (1-\rho)\rho^i$, $i \geq 0$ is its stationary distribution. Is this chain reversible? Suppose there is a cost s per unit time for serving a customer and a cost h per unit time of holding a customer in the system. Show that the average cost per unit time is $\rho[s + h/(1-\rho)]$. Now, assume there is a reward R for serving a customer. Show that the average net reward is $\rho q(1-p)R - \rho[s + h/(1-\rho)]$.

Exercise 51. *Continuation.* Consider the $M/M/1$ queueing model in the preceding exercise with $p < q$. Example 29 pointed out that the regenerative property of Markov chains in Proposition 67 implies that thetimes between empty epochs ξ_n are i.i.d. and the durations of the busy periods β_n are i.i.d. for $n \geq 2$. Find an expression for $E_0[\xi_1]$ and show that

$$E_0[\beta_1] = q(1-p)/(q-p) - 1/p.$$

Verify that $E_0[\beta_1]$ tends to ∞ as $p \uparrow q$ (the traffic is heavy), and it tends to 1 as $p \downarrow 0$ (the traffic is light).

Exercise 52. *Sharing a Buffer.* Let X_n and X'_n be independent $M/M/1$ queue-length processes as in the preceding exercise with parameters,

$$\rho = p(1-q)/[q(1-p)], \qquad \rho' = p'(1-q')/[q'(1-p')].$$

Now, consider the modification in which these queues must share a common buffer with capacity C so that a new arrival (in either system) is blocked from entering and disregarded when the buffer is full. Let Y_n and Y'_n denote the resulting queue-length processes (which are now dependent). Show that $Z_n = (Y_n, Y'_n)$ is a Markov chain on the state space $\tilde{S} = \{(i, i') : 0 \leq i + i' \leq C\}$ and specify its transition probabilities. Show that this chain is reversible and its stationary distribution has the form $\pi_{(i,i')} = c\rho^i(\rho')^{i'}$.
(These results follow from the definition of reversibility and a little algebra. Chapter 4 covers related models based on the following properties. (a) If X_n and X'_n are independent reversible Markov chains, then (X_n, X'_n) is also a reversible Markov chain. (b) If X_n is a reversible Markov chain on S with respect to π, then its restriction to a subset $\tilde{S} \subset S$ is also reversible with respect to π restricted to $\tilde{S}$.)

Exercise 53. Let $S_0 = 0$ and $S_n = Y_1 + \cdots Y_n$, for $n \geq 1$, where Y_k are i.i.d. nonnegative integer valued random variables. Let X_n denote the unit in the integer expansion of S_n (so X_n equals S_n modulo 10; if $S_n = 321$, then $X_n = 1$). Show that X_n is a Markov chain and specify its transition probabilities. Under what conditions is X_n irreducible. Assuming it is, find its stationary distribution. (Similar properties hold when X_n equals S_n modulo 10^k for some $k \geq 1$.)

Exercise 54. In the context of Example 71, suppose the cost of processing a type-i job in time v is $c(i, v)$, where $c : S \times \mathbb{R}_+ \to \mathbb{R}$. Then $c(X_n, V_n)$ is the cost of processing job X_n. Find an expression for the average cost of processing jobs, which is $\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} c(X_m, V_m)$.

Exercise 55. *Random Walk on an Edge-Weighted Graph.* Consider a Markov chain X_n on the set of vertices S of a finite graph. Associated with each pair of vertices i and j there is a nonnegative weight w_{ij}, which is 0 if (i, j) is not an edge; and the probability of a transition from i to j is directly proportional to this weight. Therefore, its transition probabilities are $p_{ij} = w_{ij} / \sum_k w_{ik}$. Assume the chain is irreducible. Show that the Markov chain is reversible and its stationary distribution is

$$\pi_i = \sum_j w_{ij} / \sum_{k,\ell} w_{i\ell}, \quad i \in S.$$

Find a formula for the average sojourn time of the chain in a subset A of vertices.

Exercise 56. Determine the stationary distribution of the random walk on a circle described in Example 100. Use the property that the balance equations for $i = 1, \ldots, \ell - 1$ are the same as those for a standard random walk, and so $\pi_i = \pi_0 \prod_{k=1}^{i} p_{k-1}/q_k$ for $i \le \ell - 1$. Then solve for π_ℓ and π_0. Consider the special case $p_i = p$ and $q_i = 1 - p$ for all i. For what value of p is the random walk reversible?

Exercise 57. Show that if a Markov chain is periodic with period greater than 2, then the chain is not reversible. For the random walk in Example 94, specify its periodicity for the following cases:

(a) $r_i = 0$ for each i. (b) $r_i > 0$ for some i.

Exercise 58. *McCabe Library.* Consider Example 101 for a library of three books. Compute the product (1.67) for the following path of states

$$(1, 2, 3) \to (2, 1, 3) \to (2, 3, 1) \to (3, 2, 1) \to (3, 1, 2).$$

An alternate way of modeling the state of the library is by the vector $\mathbf{z} = (z_1, \ldots, z_m)$, where z_b denotes the location of book b. This $\mathbf{z}$ is the inverse of the corresponding state $\mathbf{i}$ (which lists how the books are arranged on the shelf); indeed, $z_{i_k} = k$ and $i_{z_b} = b$. Use this one-to-one correspondence between $\mathbf{i}$'s and $\mathbf{z}$'s, to show that the successive values of this shelf variable Z_n is a reversible Markov chain. Describe the transition probabilities of Z_n in terms of the probabilities p_b of the book selections, and show that its stationary distribution is

$$\pi(\mathbf{z}) = c \prod_{b=1}^{m} p_b^{z_b}, \quad \mathbf{z} \in S.$$

One can make use of the results in Example 101, since $Z_n = \mathbf{z}$ if and only if the z_bth component of the process X_n equals b for each b.

Exercise 59. *Markov Chain Monte Carlo.* In the context of Section 1.18, consider the estimator of μ given by

$$\hat{\mu}_n = \frac{\sum_{m=1}^{n} g(X_m)\eta(X_m)/\gamma(X_m)}{\sum_{m=1}^{n} \eta(X_m)/\gamma(X_m)},$$

where γ is any fixed positive probability measure (it gives flexibility in implementing the estimation). Show that $\hat{\mu}_n$ is a consistent estimator of μ.

Suppose the target distribution has the form $\pi_i = c\eta(i)$, where the normalization constant c is unknown or is too difficult to compute. Show that a consistent estimator for c^{-1} is $\hat{c}_n^{-1} = n^{-1}\sum_{m=1}^{n} \eta(X_m)/\gamma(X_m)$.

Exercise 60. *Star-Like Collage.* Consider the Markov chain X_n in Example 107 with the variation that the set S_0 may consist of more that one state. In addition to the assumption that the chain restricted to each $S_0 \cup S_k$ is ergodic with stationary distribution p_i^k, assume the chain restricted to S_0 is ergodic with stationary distribution p_i^0. Show that X_n is ergodic and its stationary distribution is

$$\pi_i = \begin{cases} \pi_0 p_i^0 & \text{if } i \in S_0 \\ \pi_0 c_k p_i^k & \text{if } i \in S_k \text{ for some } k, \end{cases}$$

where $\pi_0 = \left[1 + \sum_{k\neq 0}(c_k - 1)\right]^{-1}$ and $c_k = \left[\sum_{i\in S_0} p_i^k\right]^{-1}$. Also, show that $\pi_0 = \sum_{i\in S_0} \pi_i$.

Exercise 61. *Branching Process Properties.* Consider the branching process X_n defined by (1.33), with the generalization that X_0 is a random variable that is independent of the ξ_{ni}. Show that $E[X_n] = \mu^n E[X_0]$. For the case $\mu > 1$, show that the extinction probability is

$$\lim_{n\to\infty} P\{X_n = 0\} = \sum_{i=1}^{\infty} z^i P\{X_0 = i\},$$

where z is the unique fixed point of $\phi(s)$ in $(0, 1)$. Assuming $X_0 = 1$ and $\sigma^2 = \mathrm{Var}[\xi_{n1}]$, show that

$$\begin{aligned} \mathrm{Var}[X_{n+1}] &= \mu^2 \mathrm{Var}[X_n] + \mu^n \sigma^2, \\ \mathrm{Var}[X_n] &= \sigma^2(\mu^{2n-2} + \mu^{2n-3} + \cdots \mu^{n-1}), \qquad n \geq 1. \end{aligned}$$

Exercise 62. *Geometric Branching Process.* Consider a branching process in which each item produces k items with probability $p_k = pq^k$, $k \geq 0$, where $q = 1 - p$. Determine the extinction probability z, and specify conditions under which $z = 1$ and $z < 1$. Show by induction that

$$E[s^{X_n}] = \frac{p[a_n - qsa_{n-1}]}{a_{n+1} - qsa_n}, \qquad \text{when } p \neq q,$$

where $a_n = q^n - p^n$. Let $p \to 1/2$ in this expression to obtain an expression for $E[s^{X_n}]$ when $p = q = 1/2$. Use this last generating function to obtain

$$P\{X_n = 0\} = n/(n+1), \qquad \text{when } p = q.$$

Exercise 63. Consider a branching process in which each item produces k items with probability $1/4$, where $0 \le k \le 3$. First show that the extinction probability z is in the interval [.30, .50]. Then compute the extinction probability of the process by the procedure in Remark 49 for $\varepsilon = .001$. Alternatively, the probability can be obtained – you need not do this – by solving the cubic equation $s = (1/4)(s^3 + s^2 + s + 1)$.

Exercise 64. *Total Progeny in Branching.* Let X_n denote the branching process defined by (1.33), where $X_0 = 1$. The total progeny in this branching process is $Y = \sum_{n=0}^{\infty} X_n$. Of course, $P\{Y < \infty\} = z$ (the extinction probability). Show that when $\mu < 1$,

$$E[Y] = 1/(1-\mu), \qquad \text{Var}[Y] = \text{Var}[X_1]/(1-\mu)^3.$$

Prove that the generating function $G(s) = E[s^Y]$ satisfies $G(s) = s\phi(G(s))$. It is also true (you need not prove it) that $G(s)$ is the unique solution of $\gamma = s\phi(\gamma)$, $\gamma \in (0, z]$. Using this result, find $G(s)$ for the geometric branching in Exercise 62.

Exercise 65. *Estimation of Transition Probabilities.* Suppose one wants to estimate the transition probabilities of an ergodic Markov chain X_n based on observing $X_0, \ldots, X_n$, where the number of states is known. An estimator of its transition probability p_{ij} is

$$\widehat{p}_{ij}(n) = \frac{\sum_{k=1}^{n} \mathbf{1}(X_{k-1} = i, X_k = j)}{\sum_{k=1}^{n} \mathbf{1}(X_{k-1} = i)}, \qquad i, j \in S.$$

This is the portion of the times the chain moves to j upon leaving i. Show that $\widehat{p}_{ij}(n)$ is a consistent estimator of p_{ij} in that $\widehat{p}_{ij}(n) \to p_{ij}$ a.s. as $n \to \infty$.

Exercise 66. Suppose X_n is an ergodic Markov chain that is stationary. Show that the following statements are equivalent.
(a) X_n is reversible.
(b) $(X_n, X_{n+1}) \stackrel{d}{=} (X_{n+1}, X_n), \quad n \ge 1.$
(c) $(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (X_n, \ldots, X_2, X_1), \quad n \ge 1.$

Exercise 67. *Time-Reversible Chains.* A Markov chain X_n is *reversible in time* if, for any $\nu \ge 1$,

$$(X_\nu, X_{\nu-1}, \ldots, X_{\nu-n}) \stackrel{d}{=} (X_0, X_1, \ldots, X_n), \qquad n \ge \nu.$$

That is, the chain viewed in reverse time beginning at ν (like viewing a video tape in reverse) is equal in distribution to viewing the chain in the forward

direction. Show that a Markov chain is time-reversible if and only if it is stationary and reversible (in the usual sense).

Exercise 68. *Fork-Join System.* Prove that (1.85) is also a necessary condition for the fork-join Markov chain X_n in Example 116 to be positive recurrent. Also, discuss why Foster's criterion does not work to prove positive recurrence with the linear function $v(x) = \sum_{i=1}^{m} x_i$.

Chapter 2
Renewal and Regenerative Processes

Renewal and regenerative processes are models of stochastic phenomena in which an event (or combination of events) occurs repeatedly over time, and the times between occurrences are i.i.d. Models of such phenomena typically focus on determining limiting averages for costs or other system parameters, or establishing whether certain probabilities or expected values for a system converge over time, and evaluating their limits.

The chapter begins with elementary properties of renewal processes, including several strong laws of large numbers for renewal and related stochastic processes. The next part of the chapter covers Blackwell's renewal theorem, and an equivalent key renewal theorem. These results are important tools for characterizing the limiting behavior of probabilities and expectations of stochastic processes. We present strong laws of large numbers and central limit theorems for Markov chains and regenerative processes in terms of a process with regenerative increments (which is essentially a random walk with auxiliary paths). The rest of the chapter is devoted to studying regenerative processes (including ergodic Markov chains), processes with regenerative increments, terminating renewal processes, and stationary renewal processes.

2.1 Renewal Processes

This section introduces renewal processes and presents several examples. The discussion covers Poisson processes and renewal processes that are "embedded" in stochastic processes.

We begin with notation and terminology for point processes that we use in later chapters as well. Suppose $0 \leq T_1 \leq T_2 \leq \ldots$ are finite random times at which a certain event occurs. The number of the times T_n in the interval $(0, t]$ is

$$N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t), \qquad t \geq 0.$$

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

We assume this counting process is finite valued for each t, which is equivalent to $T_n \to \infty$ a.s. as $n \to \infty$.

More generally, we will consider T_n as points (or locations) in $\mathbb{R}_+$ (e.g., in time, or a physical or virtual space) with a certain property, and $N(t)$ is the number of points in $[0, t]$. The process $\{N(t) : t \geq 0\}$, denoted by $N(t)$, is a *point process* on $\mathbb{R}_+$. The T_n are its *occurrence times* (or point locations). The point process $N(t)$ is *simple* if its occurrence times are distinct: $0 < T_1 < T_2 < \ldots$ a.s. (there is at most one occurrence at any instant).

Definition 1. A simple point process $N(t)$ is a *renewal process* if the inter-occurrence times $\xi_n = T_n - T_{n-1}$, for $n \geq 1$, are independent with a common distribution F, where $F(0) = 0$ and $T_0 = 0$. The T_n are called *renewal times*, referring to the independent or renewed stochastic information at these times. The ξ_n are the *inter-renewal times*, and $N(t)$ is the *number of renewals* in $(0, t]$.

Examples of renewal processes include the random times at which: customers enter a queue for service, insurance claims are filed, accidents or emergencies happen, or a stochastic process enters a special state of interest. In addition, T_n might be the location of the nth vehicle on a highway, or the location of the nth flaw along a pipeline or cable, or the cumulative quantity of a product processed in n production cycles. A discrete-time renewal process is one whose renewal times T_n are integer-valued. Such processes are used for modeling systems in discrete time, or for modeling sequential phenomena such as the occurrence of a certain character (or special data packet) in a string of characters (or packets), such as in DNA sequences.

To define a renewal process for any context, one only has to specify a distribution F with $F(0) = 0$ for the inter-renewal times. The F in turn defines the other random variables. More formally, there exists a probability space and independent random variables $\xi_1, \xi_2, \ldots$ defined on it that have the distribution F (see Corollary 6 in the Appendix). Then the other quantities are $T_n = \sum_{k=1}^{n} \xi_k$ and $N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t)$, where $T_n \to \infty$ a.s. by the strong law of large numbers (Theorem 72 in Chapter 1).

Here are two illustrations.

Example 2. Scheduled Maintenance. An automobile is lubricated when its owner has driven it L miles or every M days, whichever comes first. Let $N(t)$ denote the number of lubrications up to time t. Suppose the numbers of miles driven in disjoint time periods are independent, and the number of miles in any time interval has the same distribution, regardless of where the interval begins. Then it is reasonable that $N(t)$ is a renewal process. The inter-renewal distribution is $F(t) = P\{\tau \wedge M \leq t\}$, where τ denotes the time to accumulate L miles on the automobile.

This scheduled maintenance model applies to many types of systems where maintenance is performed when the system usage exceeds a certain level L or when a time M has elapsed. For instance, in reliability theory, the *Age*

Replacement model of components or systems, replaces a component with lifetime τ if it fails or reaches a certain age M (see Exercise 19).

Example 3. Service Times. An operator in a call center answers calls one at a time. The calls are independent and homogeneous in that the callers, the call durations, and the nature of the calls are independent and homogeneous. Also, the time needed to process a typical call (which may include post-call processing) has a distribution F. Then one would be justified in modeling the number of calls $N(t)$ that the operator can process in time t as a renewal process. The time scale here refers to the time that the operator is actually working; it is not the real time scale that includes intervals with no calls, operator work-breaks, etc.

Elementary properties of a renewal process $N(t)$ with inter-renewal distribution F are as follows. The times T_n are related to the counts $N(t)$ by

$$\{N(t) \geq n\} = \{T_n \leq t\},$$
$$T_{N(t)} \leq t < T_{N(t)+1}.$$

In addition, $N(T_n) = n$ and

$$N(t) = \max\{n : T_n \leq t\} = \min\{n : T_{n+1} > t\}.$$

These relations (which also hold for simple point processes) are used to derive properties of $N(t)$ in terms of T_n, and vice versa.

We have a good understanding of $T_n = \sum_{k=1}^{n} \xi_k$, since it is a sum of independent variables with distribution F. In particular, by properties of convolutions of distributions (see the Appendix), we know that

$$P\{T_n \leq t\} = F^{n\star}(t),$$

which is the n-fold convolution of F. Then $\{N(t) \geq n\} = \{T_n \leq t\}$ yields

$$P\{N(t) \leq n\} = 1 - F^{(n+1)\star}(t). \tag{2.1}$$

Also, using $E[N(t)] = \sum_{n=1}^{\infty} P\{N(t) \geq n\}$ (see Exercise 1), we have

$$E[N(t)] = \sum_{n=1}^{\infty} F^{n\star}(t). \tag{2.2}$$

The following result justifies that this mean and all moments of $N(t)$ are finite. Properties of moment generating functions are in the Appendix.

Proposition 4. *For each $t \geq 0$, the moment generating function $E[e^{\alpha N(t)}]$ exists for some α in a neighborhood of 0, and hence $E[N(t)^m] < \infty$, $m \geq 1$.*

Proof. It is clear that if $0 \leq X \leq Y$ and Y has a moment generating function on an interval $[0, \varepsilon]$, then so does X. Therefore, to prove the assertion it

suffices to find a random variable larger than $N(t)$ whose moment generating function exists.

To this end, choose $x > 0$ such that $p = P\{\xi_1 > x\} > 0$. Consider the sum $S_n = \sum_{k=1}^{n} 1(\xi_k > x)$, which is the number of successes in n independent Bernoulli trials with probability of success p. The number of trials until the mth success is $Z_m = \min\{n : S_n = m\}$.

Clearly $xS_n < T_n$, and so

$$N(t) = \max\{n : T_n \le t\} \le \max\{n : S_n = \lfloor t/x \rfloor\} \le Z_{\lfloor t/x \rfloor + 1}.$$

Now Z_m has a negative binomial distribution with parameters m and p, and its moment generating is given in Exercise 2. Thus, $Z_{\lfloor t/x \rfloor + 1}$ has a generating function, and hence $N(t)$ has one as well. Furthermore, this existence ensures that all moments of $N(t)$ exist (a basic property of moment generating functions for nonnegative random variables).

Keep in mind that the preceding properties of the renewal process $N(t)$ are true for any distribution F with $F(0) = 0$. When this distribution has a finite mean μ and finite variance σ^2, the distribution of $N(t)$, for large t, is approximately a normal distribution with mean t/μ and variance $t\sigma^2/\mu^3$ (this follows by the central limit theorem in Example 67 below). Refined asymptotic approximations for the mean of $N(t)$ are given in Proposition 84.

The rest of this section is devoted to examples of renewal processes. The most prominent renewal process is as follows.

Example 5. Poisson Process. Suppose the i.i.d. inter-renewal times of the renewal process $N(t)$ have the exponential distribution $F(t) = 1 - e^{-\lambda t}$ with rate λ (its mean is λ^{-1}). Then as we will see in the next chapter, $N(t)$ is *Poisson process* with rate λ.

In this case, by properties of convolutions

$$P\{T_n \le t\} = F^{n\star}(t) = \int_0^t \lambda^n x^{n-1} \frac{e^{-\lambda x}}{(n-1)!} dx.$$

This is a gamma distribution with parameters n and λ. Alternatively,

$$P\{T_n \le t\} = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This is justified by noting that the derivative of the summation equals the integrand (the gamma density) in the preceding integral. Then using the relation $\{N(t) \ge n\} = \{T_n \le t\}$, we arrive at

$$P\{N(t) \le n\} = \sum_{k=0}^{n} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This is the Poisson distribution with mean $E[N(t)] = \lambda t$.

Poisson processes are very important in the theory and applications of stochastic processes. We will discuss them further in Chapter 3. Note that the discrete-time analogue of a Poisson process is the Bernoulli process described in Exercise 2.

Example 6. Delayed Renewal Process. Many applications involve a renewal process $N(t)$ with the slight difference that the first renewal time ξ_1 does not have the same distribution as the other ξ_n, for $n \geq 2$. We call $N(t)$ a *delayed renewal process*. Elementary properties of delayed renewal processes are similar to those for renewal processes with the obvious changes (e.g., if ξ_1 has distribution G, then the time T_n of the nth renewal has the distribution $G \star F^{(n-1)\star}(t)$). More important, we will see that many limit theorems for renewal processes apply to delayed renewal processes.

In addition to being of interest by themselves, renewal processes play an important role in analyzing more complex stochastic processes. Specifically, as a stochastic process evolves over time, it is natural for some event associated with its realization to occur again and again. When the "embedded" occurrence times of the event are renewal times, they may be useful for gleaning properties about the parent process. Stochastic processes with embedded renewal times include discrete- and continuous-time Markov chains, Markov-Renewal processes and more general regenerative processes (which are introduced in later chapters).

The next example describes renewal processes embedded in ergodic Markov chains due the regenerative property of Markov chains.

Example 7. Ergodic Markov Chain. Let X_n denote a discrete-time Markov chain on a countable state space that is ergodic (aperiodic, irreducible and positive recurrent). Consider any state i and let $0 < \nu_1 < \nu_2 < \ldots$ denote the (discrete) times at which X_n enters state i. Theorem 67 in Chapter 1 showed that the times ν_n form a discrete-time renewal process when $X_0 = i$. These times form a delayed renewal process when $X_0 \neq i$. The Bernoulli process in Exercise 2 is a special case.

Example 8. Cyclic Renewal Process. Consider a continuous-time stochastic process $X(t)$ that cycles through states $0, 1, \ldots, K-1$ in that order, again and again. That is, it starts at $X(0) = 0$, and its nth state is j if $n = mK + j$ for some m. For instance, in modeling the status of a machine or system, $X(t)$ might be the amount of deterioration of a system, or the number of shocks (or services) it has had, and the system is renewed whenever it ends a sojourn in state $K-1$.

Assume the sojourn times in the states are independent, and let F_j denote the sojourn time distribution for state j, where $F_j(0) = 0$. The time for the process $X(t)$ to complete a cycle from state 0 back to 0 has the distribution $F = F_0 \star F_1 \star \cdots \star F_{K-1}$. Then it is clear that the times at which $X(t)$ enters state 0 form a renewal process with inter-renewal distribution F. We call $X(t)$ a *cyclic* renewal process.

There are many other renewal processes embedded in $X(t)$. For instance, the times at which the process enters any fixed state i form a delayed renewal process with the same distribution F. Another more subtle delayed renewal process is the sequence of times at which the processes $X(t)$ bypasses state 0 by jumping from state $K-1$ to state 1 (assuming $F_0(0) > 0$); see Exercise 7. It is quite natural for a single stochastic process to contain several such embedded renewal processes.

Example 9. Alternating Renewal Process. An *alternating* renewal process is a cyclic renewal process with only two states, say 0 and 1. This might be appropriate for indicating whether a system is working (state 1) or not working (state 0), or whether a library book is available or unavailable for use.

2.2 Strong Laws of Large Numbers

This section begins our study of the long run behavior of renewal and related stochastic processes. In particular, we present a framework for deriving strong laws of large numbers for a variety of processes. We have already seen SLLNs in Chapter 1 for sums of i.i.d. random variables and for functions of Markov chains.[1]

Throughout this section, assume that $N(t)$ is a point process on $\mathbb{R}_+$ with occurrence times T_n. With no loss in generality, assume that $N(t) \uparrow \infty$ a.s. as $t \to \infty$. The first result says that T_n satisfies a SLLN if and only if $N(t)$ does. Here $1/\mu$ is 0 when $\mu = \infty$.

Theorem 10. *For a constant $\mu \leq \infty$ (or random variable), the following statements are equivalent:*

$$\lim_{n\to\infty} n^{-1}T_n = \mu \quad a.s. \tag{2.3}$$

$$\lim_{t\to\infty} t^{-1}N(t) = 1/\mu \quad a.s. \tag{2.4}$$

Proof. Suppose (2.3) holds. We know $T_{N(t)} \leq t < T_{N(t)+1}$. Dividing these terms by $N(t)$ (for large enough t so $N(t) > 0$), we have

$$\frac{T_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{T_{N(t)+1}}{N(t)+1}\frac{N(t)+1}{N(t)}.$$

Supposition (2.3) along with $N(t) \uparrow \infty$ and $(N(t)+1)/N(t) \to 1$ ensure that the first and last terms in this display converge to μ. Since $t/N(t)$ is sandwiched between these terms, it must also converge to their limit μ. This proves (2.4).

[1] The limit statements here and below are for the a.s. mode of convergence, but we sometimes suppress the term a.s., especially in the proofs.

Conversely, suppose (2.4) holds. When $N(t)$ is simple, $N(T_n) = n$, and so $T_n/n = T_n/N(T_n) \to \mu$, which proves (2.3). When $N(t)$ is not simple, $N(T_n) \geq n$ and (2.3) follows by Exercise 18.

Corollary 11. (SLLN for Renewal Processes) *If $N(t)$ is a renewal process whose inter-renewal times have a mean $\mu \leq \infty$, then*

$$t^{-1}N(t) \to 1/\mu \quad \textit{a.s. as } t \to \infty.$$

Proof. This follows by Theorem 10, since the classical SLLN (Theorem 72 in Chapter 1) ensures that $n^{-1}T_n \to \mu$.

Example 12. Statistical Estimation. Suppose $N(t)$ is a Poisson process with rate λ, but this rate is not known, and one wants to estimate it. One approach is to observe the process for a fixed time interval of length t and record $N(t)$. Then an estimator for λ is

$$\hat{\lambda}_t = t^{-1}N(t).$$

This estimator is unbiased in that $E[\hat{\lambda}_t] = \lambda$. It is also a *consistent estimator* since $\hat{\lambda}_t \to \lambda$ by Corollary 11. Similarly, if $N(t)$ is a renewal process whose inter-renewal distribution has a finite mean μ, then $\hat{\mu}_t = t/N(t)$ is a consistent estimator for μ (but it is not unbiased).

Of course, if it is practical to observe a fixed number n of renewals (rather than observing over a "fixed" time), then $n^{-1}T_n$ is an unbiased and consistent estimator of μ.

We now present a framework for obtaining limiting averages (or SLLNs) for a variety of stochastic processes. Consider a real-valued stochastic process $\{Z(t) : t \geq 0\}$ on the same probability space as the point process $N(t)$. Our interest is in natural conditions under which the limit of its average value $t^{-1}Z(t)$ exists. For instance, $Z(t)$ might denote a cumulative utility (e.g., cost or reward) associated with a system, and one is interested in the utility per unit time $t^{-1}Z(t)$ for large t.

The following theorem relates the limit of the *time average* $t^{-1}Z(t)$ to the limit of the embedded *interval average* $n^{-1}Z(T_n)$. An important quantity is

$$M_n = \sup_{T_{n-1} < t \leq T_n} |Z(t) - Z(T_{n-1})|,$$

which is the maximum fluctuation of $Z(t)$ in the interval $(T_{n-1}, T_n]$. We impose the rather weak assumption that this maximum does not increase faster than n does.

Theorem 13. *Suppose that $n^{-1}T_n \to \mu$ a.s. as $n \to \infty$, where $\mu \leq \infty$ is a constant or random variable. Let a be a constant or random variable that may be infinite when μ is finite, and consider the limit statements*

$$\lim_{t\to\infty} t^{-1}Z(t) = a/\mu \quad a.s. \tag{2.5}$$

$$\lim_{n\to\infty} n^{-1}Z(T_n) = a \quad a.s. \tag{2.6}$$

Statement (2.5) implies (2.6). Conversely, (2.6) implies (2.5) if the process $Z(t)$ is increasing, or if $\lim_{n\to\infty} n^{-1}M_n = 0$ a.s.

Proof. Clearly (2.5) implies (2.6) since

$$n^{-1}Z(T_n) = T_n^{-1}Z(T_n)(T_n/n) \to a.$$

Next, suppose (2.6) holds, and consider

$$t^{-1}Z(t) = t^{-1}Z(T_{N(t)}) + r(t).$$

where $r(t) = t^{-1}[Z(t) - Z(T_{N(t)})]$. By Theorem 10, $n^{-1}T_n \to \mu$ implies $N(t)/t \to 1/\mu$. Using the latter and (2.6), we have

$$t^{-1}Z(T_{N(t)}) = [Z(T_{N(t)})/N(t)][N(t)/t] \to a/\mu.$$

Then to prove $t^{-1}Z(t) \to a/\mu$, it remains to show $r(t) \to 0$.

In case $Z(t)$ is increasing, (2.6) and $N(t)/t \to 1/\mu$ ensure that

$$|r(t)| \le \frac{[Z(T_{N(t)+1}) - Z(T_{N(t)})]}{N(t)} \frac{N(t)}{t} \to 0.$$

Also, in the other case in which $n^{-1}M_n \to 0$,

$$|r(t)| \le [M_{N(t)+1}/(N(t)+1)][(N(t)+1)/t] \to 0.$$

Thus $r(t) \to 0$, which completes the proof that (2.6) implies (2.5).

Here is a consequence of Theorem 13 that applies to processes with regenerative increments, which are discussed in Section 2.10.

Corollary 14. *If $N(t)$ is a renewal process, and $(Z(T_n) - Z(T_{n-1}), M_n)$, $n \ge 1$, are i.i.d. with finite means, then*

$$t^{-1}Z(t) \to E[Z(T_1) - Z(0)]/E[T_1] \quad a.s. \text{ as } t \to \infty. \tag{2.7}$$

Proof. By the classical SLLN, $n^{-1}Z(T_n) \to E[Z(T_1) - Z(0)]$. Also, since M_n are i.i.d., it follows by Exercise 33 in the preceding chapter that $n^{-1}M_n \to 0$. Then Theorem 13 yields (2.7).

We will see a number of applications of Theorem 13 throughout this chapter. Here are two elementary examples.

Example 15. Renewal Reward Process. Suppose $N(t)$ is a renewal process associated with a system in which a reward Y_n (or cost or utility value) is

received at time T_n, for $n \geq 1$. Then the total reward in $(0, t]$ is[2]

$$Z(t) = \sum_{n=1}^{\infty} Y_n \mathbf{1}(T_n \leq t) = \sum_{n=1}^{N(t)} Y_n, \qquad t \geq 0.$$

For instance, Y_n might be claims received by an insurance company at times T_n, and $Z(t)$ would represent the cumulative claims.

The process $Z(t)$ is a *renewal reward process* if the pairs (ξ_n, Y_n), $n \geq 1$, are i.i.d. (ξ_n and Y_n may be dependent). Under this assumption, it follows by Theorem 13 that the average reward per unit time is

$$\lim_{t\to\infty} t^{-1} Z(t) = E[Y_1]/E[\xi_1] \quad \text{a.s.},$$

provided the expectations are finite. This result is very useful in many diverse contexts. One only has to justify the renewal conditions and evaluate the expectations. In complicated systems with many activities, a little thought may be needed to identify the renewal times as well as the associated rewards.

Example 16. Cyclic Renewal Process. Let $X(t)$ be a cyclic renewal process on $0, \ldots, K-1$ as in Example 8. Recall that the entrance times to state 0 form a renewal process, and the mean inter-renewal time is $\mu = \mu_0 + \cdots + \mu_{K-1}$, where μ_i is the mean sojourn time in state i. Suppose a cost or value $f(i)$ per unit time is incurred whenever $X(t)$ is in state i. Then the average cost per unit time is

$$\lim_{t\to\infty} t^{-1} \int_0^t f(X(s))ds = \frac{1}{\mu} \sum_{i=0}^{K-1} f(i)\mu_i \quad \text{a.s..} \tag{2.8}$$

This follows by applying Corollary 13 to $Z(t) = \int_0^t f(X(s))ds$ and noting that $E[Z(T_1)] = \sum_{i=0}^{K-1} f(i)\mu_i$.

A particular case of (2.8) says that the portion of time $X(t)$ spends in a subset of states J is

$$\lim_{t\to\infty} t^{-1} \int_0^t \mathbf{1}(X(s) \in J)ds = \frac{1}{\mu} \sum_{j\in J} \mu_j \quad \text{a.s..}$$

2.3 The Renewal Function

This section describes several fundamental properties of renewal processes in terms of the their mean value functions.

[2] Recall the convention that $\sum_{n=1}^{0}(\cdot) = 0$.

For this discussion, suppose that $N(t)$ is a renewal process with inter-renewal distribution F with a finite mean μ. We begin by showing that the mean value function $E[N(t)]$ contains all the probabilistic information about the process. It is more convenient to use the slight variation of the mean value function defined as follows.

Definition 17. The *renewal function* associated with the distribution F (or the process $N(t)$) is

$$U(t) = \sum_{n=0}^{\infty} F^{n\star}(t), \qquad t \in \mathbb{R}, \tag{2.9}$$

where $F^{0\star}(t) = \mathbf{1}(t \geq 0)$. Clearly $U(t) = E[N(t)] + 1$, for $t \geq 0$, is the expected number of renewals up to time t, including a "fictitious renewal" at time 0.

Note that $U(t)$ is similar to a distribution function in that it is nondecreasing and right-continuous on $\mathbb{R}$, but $U(t) \uparrow \infty$ as $t \to \infty$. Keep in mind that $U(t)$ is 0 for $t < 0$ and it has a unit jump at $t = 0$. Although a renewal function is ostensibly very simple, it has some remarkable uses as we will soon see.

Our first observation is that if the inter-renewal times are continuous random variables, then the renewal function has a density.

Proposition 18. *Suppose the inter-renewal distribution F has a density f. Then $U(t)$ also has a density for $t > 0$, and it is $U'(t) = \sum_{n=1}^{\infty} f^{n\star}(t)$. In addition,*

$$P\{N(t) > N(t-)\} = 0, \qquad t \geq 0. \tag{2.10}$$

Proof. The first assertion follows since $U(t) = \sum_{n=0}^{\infty} F^{n\star}(t)$, and the derivative of $F^{n\star}(t)$ is $f^{n\star}(t)$. The second assertion, which is equivalent to $N(t) - N(t-) = 0$ a.s., will follow if $E[N(t) - N(t-)] = 0$. But the last equality is true since, by the monotone convergence theorem (Theorem 13 in the Appendix) and the continuity of U,

$$E[N(t-)] = E[\lim_{s \uparrow t} N(s)] = \lim_{s \uparrow t} U(s) - 1 = U(t) - 1 = E[N(t].$$

Expression (2.10) tells us that the probability of a renewal at any time is 0, when the inter-renewal times are continuous. Here is an important case.

Remark 19. If $N(t)$ is a Poisson process with rate λ, then the probability of a jump at any time t is 0.

Some of the results below have slight differences depending on whether the inter-renewal distribution is or is not arithmetic. The distribution F is *arithmetic* (or periodic) if it is piecewise constant and its points of increase are contained in a set $\{0, d, 2d, \ldots\}$; the largest $d > 0$ with this property is

the *span*. In this case, it is clear that the distributions $F^{n\star}$ and the renewal function $U(t)$ also have this arithmetic property. If F is not arithmetic, we call it *non-arithmetic*. A distribution with a continuous part is necessarily non-arithmetic.

The rest of this chapter makes extensive use of Riemann-Stieltjes integrals; see the review in the Appendix. In particular, the expectation of a function $g : \mathbb{R} \to \mathbb{R}$ on a finite or infinite interval I with respect to F will be expressed as the *Riemann-Stieltjes integral*[3]

$$\int_I g(t)dF(t).$$

All the functions in this book like g are assumed to be measurable (see the Appendix); we will not repeat this assumption unless emphasis is needed. Riemann-Stieltjes integrals with respect to U are defined similarly, since U is like a distribution function. A typical integral is

$$\int_{[0,b]} g(t)dU(t) = g(0) + \int_{(0,b]} g(t)dU(t).$$

The right-hand side highlights that $g(0)U(0) = g(0)$ is the contribution from the unit jump of U at 0. Since $U(t) = 0$ for $t < 0$, we will only consider integrals with respect to U on intervals in $\mathbb{R}_+$.

An important property of the renewal function $U(t)$ is that it uniquely determines the distribution F. To see this, we will use Laplace transforms. The Laplace-Stieltjes or simply the *Laplace transform* of F is defined by

$$\hat{F}(\alpha) = \int_{\mathbb{R}_+} e^{-\alpha t} dF(t), \qquad \alpha \geq 0.$$

A basic property is that the transform $\hat{F}$ uniquely determines F and vice versa. The Laplace transform $\hat{U}(\alpha)$ of $U(t)$ is defined similarly. Now, taking the Laplace transform of both sides in (2.9), we have

$$\hat{U}(\alpha) = \sum_{n=0}^{\infty} \widehat{F^{n\star}}(\alpha) = \sum_{n=0}^{\infty} \hat{F}(\alpha)^n = 1/(1 - \hat{F}(\alpha)).$$

This yields the following result.

Proposition 20. The Laplace transforms $\hat{U}(\alpha)$ and $\hat{F}(\alpha)$ determine each other uniquely by the relation $\hat{U}(\alpha) = 1/(1-\hat{F}(\alpha))$. Hence U and F uniquely determine each other.

One can sometimes use this result for identifying that a renewal process is of a certain type. For instance, a Poisson process has a renewal function

[3] This integral is the usual Riemann integral $\int_I g(t)f(t)dt$ when F has a density f. Also, $\int_I h(t)dt$ is written as $\int_a^b h(t)dt$ when I is $(a, b]$ or $[a, b]$ etc.

$U(t) = \lambda t + 1$, and so any renewal process with this type of renewal function is a Poisson process.

Remark 21. A renewal process $N(t)$, whose inter-renewal times have a finite mean, is a Poisson process with rate λ if and only if $E[N(t)] = \lambda t$, for $t \geq 0$.

Other examples of renewal processes with tractable renewal functions are those whose inter-renewal distribution is a convolution or mixture of exponential distributions; see Exercises 6 and 12. Sometimes the Laplace transform $\hat{U}(\alpha) = 1/(1 - \hat{F}(\alpha))$ can be inverted to determine $U(t)$. Unfortunately, nice expressions for renewal processes are the exception rather than the rule.

In addition to characterizing renewal processes as discussed above, renewal functions arise naturally in expressions for probabilities and expectations of functions associated with renewal processes. Such expressions are the focus of much of this chapter.

The next result describes an important family of functions of point processes as well as renewal processes. Expression (2.11) is a special case of Campbell's formula in the theory of point processes (see Theorem 106 in Chapter 4).

Theorem 22. *Let $N(t)$ be a simple point process with point locations T_n such that $\eta(t) = E[N(t)]$ is finite for each t. Then for any function $f : \mathbb{R}_+ \to \mathbb{R}$,*

$$E\Big[\sum_{n=1}^{N(t)} f(T_n)\Big] = \int_{(0,t]} f(s)d\eta(s), \qquad t \geq 0, \tag{2.11}$$

provided the integral exists. Moreover, if $X_1, X_2, \ldots$ are random variables defined on the same probability space as the process $N(t)$ such that $E[X_n|T_n = s] = f(s)$, independent of n. Then

$$E\Big[\sum_{n=1}^{N(t)} X_n\Big] = \int_{(0,t]} f(s)d\eta(s), \qquad t \geq 0, \tag{2.12}$$

provided the integral exists.

Proof. We will prove (2.11) by a standard approach for proving formulas for integrals. For convenience, denote the equality (2.11) by $\Sigma(f) = I(f)$. First, consider the simple piecewise-constant function

$$f(s) = \sum_{k=1}^{m} a_k \mathbf{1}(s \in (s_k, t_k]),$$

for fixed $0 \leq s_1 < t_1 < \cdots \leq s_m < t_m \leq t$. In this case,

$$\Sigma(f) = E\Big[\sum_{k=1}^{m} a_k[N(t_k) - N(s_k)]\Big]$$
$$= \sum_{k=1}^{m} a_k[\eta(t_k) - \eta(s_k)] \;=\; I(f).$$

Next, for any nonnegative function f one can define simple functions f_m as above such that $f_m(s) \uparrow f(s)$ as $m \to \infty$ for each s. For instance,

$$f_m(s) = m \wedge (\lfloor 2^m f(s) \rfloor / 2^m)\mathbf{1}(s \in [-2^m, 2^m]).$$

Then by the monotone convergence theorem (see the Appendix, Theorem 13) and the first part of this proof,

$$\Sigma(f) = \lim_{m\to\infty} \Sigma(f_m) = \lim_{m\to\infty} I(f_m) = I(f).$$

Thus, (2.11) is true for nonnegative f.

Finally, (2.11) is true for a general function f, since $f(s) = f(s)^+ - f(s)^-$ and the preceding part of the proof for nonnegative functions yield

$$\Sigma(f) = \Sigma(f^+) - \Sigma(f^-) = I(f^+) - I(f^-) = I(f).$$

It suffices to prove (2.12) for nonnegative X_n. Conditioning on T_n, we have

$$E\Big[\sum_{n=1}^{N(t)} X_n\Big] = \sum_{n=1}^{\infty} E[X_n\mathbf{1}(T_n \le t)] \;=\; \sum_{n=1}^{\infty} \int_{[0,t]} f(s)dF^{n\star}(s)$$
$$= \sum_{n=1}^{\infty} E[f(T_n)\mathbf{1}(T_n \le t)] \;=\; E\Big[\sum_{n=1}^{N(t)} f(T_n)\Big].$$

Then applying (2.11) to the last term yields (2.12).

Remark 23. Theorem 22 applies to a renewal process $N(t)$ with its renewal function U being equal to η. For instance, (2.12) would be

$$E\Big[\sum_{n=1}^{N(t)} X_n\Big] = \int_{(0,t]} f(s)dU(s).$$

Note that this integral does not include the unit jump of U at 0. An extension that includes a value X_0 with $f(0) = E[X_0]$ would be

$$E\Big[\sum_{n=0}^{N(t)} X_n\Big] = \int_{[0,t]} f(s)dU(s). \tag{2.13}$$

This remark yields the following special case of a general Wald identity for stopping times in Corollary 25 in Chapter 5.

Corollary 24. (Wald Identity for Renewals) *For the renewal process $N(t)$,*

$$E[T_{N(t)+1}] = \mu E[N(t)+1], \qquad t \geq 0.$$

Proof. Using Remark 23 with $f(s) = E[\xi_{n+1}|T_n = s] = \mu$, it follows that

$$\begin{aligned} E[T_{N(t)+1}] &= E\Big[\sum_{n=0}^{N(t)} \xi_{n+1}\Big] \\ &= \mu U(t) = \mu E[N(t)+1]. \end{aligned}$$

In light of this result, one might suspect that $E[T_{N(t)}] = \mu E[N(t)]$. However, this is not the case. In fact, $E[T_{N(t)}] \leq \mu E[N(t)]$; and this is a strict inequality for a Poisson process; see Exercise 22.

Example 25. Discounted Rewards. Suppose a renewal process $N(t)$ has rewards associated with it such that a reward (or cost) Y_n is obtained at the nth renewal time T_n. The rewards are discounted continuously over time and if a reward y occurs a time t, it has a discounted value of $ye^{-\alpha t}$. Then the total discounted reward up to time t is

$$Z(t) = \sum_{n=1}^{N(t)} Y_n e^{-\alpha T_n}.$$

As in Theorem 22, assume there is a function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $E[Y_n|T_n = s] = f(s)$, independent of n. Then applying Remark 23 to $X_n = Y_n e^{-\alpha T_n}$ yields

$$E[Z(t)] = \int_{(0,t]} e^{-\alpha s} f(s) dU(s).$$

The next examples describe several systems modeled by renewal processes with the same type of inter-renewal distribution shown in (2.14) below (also see Exercise 16).

Example 26. Single-Server System. Pallets are scheduled to arrive at an automatically guided vehicle (AGV) station according to a renewal process $N(t)$ with inter-arrival distribution F. The station is attended by a single AGV, which can transport only one pallet at a time. Pallets scheduled to arrive when the AGV is already busy transporting a pallet are diverted to another station. Assume the transportation times are independent with common distribution G.

Let us consider the times $\widetilde{T}_n$ at which the AGV begins to transport a pallet (the times at which pallets arrive and the AGV is idle). For simplicity, assume a transport starts at time 0. To describe $\widetilde{T}_1$, let τ denote a transport time for the first pallet. Then $\widetilde{T}_1$ equals τ plus the waiting time $T_{N(\tau)+1} - \tau$ for the next pallet to arrive after transporting the first pallet. That is, $\widetilde{T}_1 = T_{N(\tau)+1}$. When the next pallet arrives at time $T_{N(\tau)+1}$, the system is renewed and these cycles are repeated indefinitely. Thus $\widetilde{T}_n$ are renewal times.

The inter-renewal distribution of $\widetilde{T}_1$ and its mean have reasonable expressions in terms of the arrival process. Indeed, conditioning on τ, which is independent of $N(t)$, yields

$$P\{\widetilde{T}_1 \le t\} = \int_{\mathbb{R}_+} P\{T_{N(x)+1} \le t\} dG(x). \tag{2.14}$$

Also, if F has a finite mean μ, then by Wald's identity,

$$E[\widetilde{T}_1] = \int_{\mathbb{R}_+} E[T_{N(x)+1}] dG(x) = \mu \int_{\mathbb{R}_+} U(x) dG(x).$$

Example 27. $G/G/1/1$ System. Consider a system in which customers arrive at a processing station according to a renewal process with inter-arrival distribution F and are processed by a single server. The processing or service times are independent with the common distribution G, and are independent of the arrival process. Also, customer arrivals during a service time are blocked from being served — they either go elsewhere or go without service. In this context the times $\widetilde{T}_n$ at which customers begin services are renewal times as in the preceding example with inter-renewal distribution (2.14). This system is called a $G/G/1/1$ system: G/G means the inter-arrival and service times are i.i.d. (with general distributions) and $1/1$ means there is one server and at most one customer in the system.

Example 28. Geiger Counters. A classical model of a Geiger counter assumes that electronic particles arrive at the counter according to a Poisson or renewal process. Upon recording an arrival of a particle, the counter is locked for a random time during which arrivals of new particles are not recorded. The times of being locked are i.i.d. and independent of the arrivals. Under these assumptions, it follows that the times $\widetilde{T}_n$ at which particles are recorded are renewal times, and have the same structure as those for the $G/G/1/1$ system described above. This so-called Type I model assumes that particles arriving while the counter is locked do not affect the counter.

A slightly different Type II Geiger counter model assumes that whenever the counter is locked and a particle arrives, that particle is not recorded, but it extends the locked period by another independent locking time. The times at which particles are registered are renewal times, but the inter-renewal distribution is more intricate than that for the Type I counter.

2.4 Future Expectations

We have just seen the usefulness of the renewal function for characterizing a renewal process and for describing some expected values of the process. In the following sections, we will discuss the major role a renewal function plays in describing the limiting behavior of probabilities and expectations associated with renewal and regenerative phenomena. This section outlines what to expect in the next three sections, which cover the heart of renewal theory.

The analysis to follow will use convolutions of functions with respect to the renewal function $U(t)$, such as

$$U \star h(t) = \int_{[0,t]} h(t-s)dU(s) = h(0) + \int_{(0,t]} h(t-s)dU(s),$$

where h is bounded on finite intervals and equals 0 for $t < 0$.

We will see that many probabilities and expectations associated with a renewal process $N(t)$ can be expressed as a function $H(t)$ that satisfies a recursive equation of the form

$$H(t) = h(t) + \int_{[0,t]} H(t-s)dF(s), \qquad t \geq 0.$$

This "renewal equation", under minor technical conditions given in the next section, has a unique solution of the form $H(t) = U \star h(t)$.

The next topic we address is the limiting behavior of such functions as $t \to \infty$. We will present Blackwell's theorem, and an equivalent key renewal theorem, which establishes

$$\lim_{t\to\infty} U \star h(t) = \frac{1}{\mu} \int_{\mathbb{R}_+} h(s)ds.$$

This is for non-arithmetic F; an analogous result holds for arithmetic F. Also, the integral is slightly different from the standard Riemann integral.

We cover the topics outlined above — Renewal Equations, Blackwell's Theorem and the Key Renewal Theorem — in the next three sections. Thereafter, we discuss applications of these theorems that describe the limiting behavior of probabilities and expectations associated with renewal, regenerative and Markov chains.

2.5 Renewal Equations

We begin our discussion of renewal equations with a concrete example.

Example 29. Let $X(t)$ be a cyclic renewal process on $0, 1, \ldots, K-1$, and consider the probability $H(t) = P\{X(t) = i\}$ as a function of time, for a fixed state i. To show $H(t)$ satisfies a renewal equation, the standard approach is to condition on the time T_1 of the first renewal (the first entrance to state 0). The result is

$$H(t) = P\{X(t) = i, T_1 > t\} + P\{X(t) = i, T_1 \le t\}, \tag{2.15}$$

where the last probability, conditioning on the renewal at T_1, is

$$\int_{[0,t]} P\{X(t) = i | T_1 = s\} dF(s) = \int_{[0,t]} H(t-s) dF(s).$$

Therefore, the recursive equation (2.15) that $H(t)$ satisfies is

$$H(t) = h(t) + F \star H(t),$$

where $h(t) = P\{X(t) = i, T_1 > t\}$. This type of equation is a renewal equation, which is defined as follows.

Definition 30. Let $h(t)$ be a real-valued function on $\mathbb{R}$ that is bounded on finite intervals and equals 0 for $t < 0$. The *renewal equation* for $h(t)$ and the distribution F is

$$H(t) = h(t) + \int_{[0,t]} H(t-s) dF(s), \qquad t \ge 0, \tag{2.16}$$

where $H(t)$ is a real-valued function. That is $H = h + F \star H$. We say $H(t)$ is a *solution of this equation* if it satisfies the equation, and is bounded on finite intervals and equals 0 for $t < 0$.

We first observe that a renewal equation has a unique solution.

Proposition 31. *The function $U \star h(t)$ is the unique solution to the renewal equation (2.16).*

Proof. Clearly $U \star h(t) = 0$ for $t < 0$, and it is bounded on finite intervals since

$$\sup_{s \le t} |U \star h(s)| \le \sup_{s \le t} |h(s)| U(t) < \infty, \qquad t \ge 0.$$

Also, $U \star h$ is a solution to the renewal equation, since by the definition of U and $F^{0\star} \star h = h$,

$$U \star h = \left(F^{0\star} + F \star \sum_{n=1}^{\infty} F^{(n-1)\star} \right) \star h = h + F \star (U \star h).$$

To prove $U \star h$ is the unique solution, let $H(t)$ be any solution to the renewal equation, and consider the difference $D(t) = H(t) - U \star h(t)$. From the renewal

equation, we have $D = F \star D$, and so iterating this yields $D = F^{n\star} \star D$. Now, the finiteness of $U(t)$ implies $F^{n\star}(t) \to 0$, as $n \to \infty$, and hence $D(t) = 0$ for each t. This proves that $U \star h(t)$ is the unique solution of the renewal equation.

The standard approach for deriving a renewal equation is by conditioning on the first renewal time to obtain the function $h(t)$ (recall Example 29). Upon establishing that a function $H(t)$ satisfies a renewal equation, one automatically knows that $H(t) = U \star h(t)$ by Proposition 31. For instance, Example 29 showed that the probability $P\{X(t) = i\}$ for a cyclic renewal process satisfies a renewal equation for $h(t) = P\{X(t) = i, T_1 > t\}$, and hence

$$P\{X(t) = i\} = U \star h(t). \tag{2.17}$$

Although $H(t) = U \star h(t)$ is a solution of the renewal equation, it is not an explicit expression for the function $H(t)$ in that $h(t)$ generally depends on $H(t)$. For instance, $h(t) = P\{X(t) = i, T_1 > t\}$ in (2.17) is part of the probability $H(t) = P\{X(t) = i\}$.

Only in very special settings is the formula $H(t) = U \star h(t)$ tractable enough for computations. On the other hand, we will see in Section 2.7 that the function $U \star h(t)$ is the framework of the Key Renewal Theorem that yields limit theorems for a variety of stochastic processes.

2.6 Blackwell's Theorem

The next issue is to characterize the limiting behavior of functions of the form $U \star h(t)$ as $t \to \infty$. This is based on the limiting behavior of $U(t)$, which we now consider.

Throughout this section, assume that $N(t)$ is a renewal process with renewal function $U(t)$ and mean inter-renewal time μ, which may be finite or infinite. In Section 2.2, we saw that $N(t)/t \to 1/\mu$ a.s., and so $N(t)$ behaves asymptotically like t/μ as $t \to \infty$ (recall that $1/\mu = 0$ when $\mu = \infty$). This suggests $U(t) = E[N(t)] + 1$ should also behave asymptotically like t/μ. Here is a confirmation.

Theorem 32. (Elementary Renewal Theorem)

$$t^{-1}U(t) \to 1/\mu, \qquad \textit{as } t \to \infty.$$

Proof. For finite μ, using $t < T_{N(t)+1}$ and Wald's identity (Corollary 24),

$$t < E[T_{N(t)+1}] = \mu U(t).$$

This yields the lower bound $1/\mu < t^{-1}U(t)$. Also, this inequality holds trivially when $\mu = \infty$. With this bound in hand, to finish proving the assertion

it suffices to show

$$\limsup_{t\to\infty} t^{-1}U(t) \le 1/\mu. \tag{2.18}$$

To this end, for a constant b, define a renewal process $\overline{N}(t)$ with inter-renewal times $\overline{\xi}_n = \xi_n \wedge b$. Define $\overline{T}_n$ and $\overline{U}(t)$ accordingly. Clearly, $U(t) \le \overline{U}(t)$. Also, by Wald's identity and $\overline{T}_{\overline{N}(t)+1} \le t+b$ (since the $\overline{\xi}_n$ are bounded by b),

$$E[\xi_1 \wedge b]\overline{U}(t) = E[\overline{T}_{\overline{N}(t)+1}] \le t+b.$$

Consequently,

$$t^{-1}U(t) \le t^{-1}\overline{U}(t) \le \frac{1+b/t}{E[\xi_1 \wedge b]}.$$

Letting $t \to \infty$ and then letting $b \to \infty$ (whereupon the last fraction tends to $1/\mu$, even when $\mu = \infty$), we obtain (2.18), which finishes the proof.

A more definitive description of the asymptotic behavior of $U(t)$ is given in the following major result.

Theorem 33. (Blackwell) *For non-arithmetic F and $a > 0$,*

$$U(t+a) - U(t) \to a/\mu, \qquad \text{as } t \to \infty.$$

If F is arithmetic with span d, the preceding limit holds with $a = md$ for any integer m.

Proof. A proof for non-arithmetic F using a coupling argument is in Section 2.15 below. A simpler proof for the arithmetic case is as follows.

Suppose F is arithmetic and, for simplicity, assume the span is $d = 1$. Then renewals occur only at integer times, and $p_i = F(i) - F(i-1)$ is the probability that an inter-renewal time is of length i, where $p_0 = 0$.

We will represent the renewal times by the backward recurrence time process $\{A(t) : t = 0, 1, 2, \ldots\}$, which we know is a Markov chain with transition probabilities

$$p_{i0} = \frac{p_i}{\sum_{j=i}^{\infty} p_j} = 1 - p_{i,i+1}, \qquad i \ge 0.$$

(recall Example 20 and Exercises 37 and 38 in Chapter 1). This chain is irreducible, and hits state 0 at and only at the renewal times. Then the chain is ergodic since the time between renewals has a finite mean μ. Theorem 54 in Chapter 1 yields $P\{A(t) = 0\} \to 1/\mu$ (the representation for π_0).

Then because $U(t+m) - U(t)$ is the expected number of renewals exactly at the times $t+1, \ldots, t+m$, it follows that

$$U(t+m) - U(t) = \sum_{k=1}^{m} P\{A(t+k) = 0\} \to m/\mu, \qquad \text{as } t \to \infty.$$

Blackwell's theorem says that the renewal function $U(t)$ is asymptotically linear. This raises the question: "Does the asymptotic linearity of $U(t)$ lead

to a nice limit for functions of the form $U \star h(t)$?" The answer is yes, as we will see shortly.

As a preliminary, let us investigate the limit of $U \star h(t)$ for a simple piecewise-constant function

$$h(s) = \sum_{k=1}^{m} a_k \mathbf{1}(s \in [s_k, t_k)),$$

where $0 \le s_1 < t_1 \le s_2 < t_2 < \cdots \le s_m < t_m < \infty$. In this case,

$$\begin{aligned} U \star h(t) = \int_{[0,t]} h(t-s) dU(s) &= \sum_{k=1}^{m} a_k \int_0^t \mathbf{1}(t-s \in [s_k, t_k)) dU(s) \\ &= \sum_{k=1}^{m} a_k [U(t-s_k) - U(t-t_k)]. \end{aligned} \tag{2.19}$$

The last equality follows since the integral is over $s \in [t-t_k, t-s_k)$, and $U(t) = 0$ when $t < 0$. By Theorem 33, we know

$$U(t-s_k) - U(t-t_k) \to (t_k - s_k)/\mu.$$

Applying this to (2.19) yields

$$\lim_{t\to\infty} U \star h(t) = \frac{1}{\mu} \sum_{k=1}^{m} a_k (t_k - s_k) = \frac{1}{\mu} \int_{\mathbb{R}_+} h(s) ds. \tag{2.20}$$

This result suggests that a limit of this form would also be true for general functions $h(t)$. That is what we will establish next.

2.7 Key Renewal Theorem

This section will complete our development of renewal functions and solutions of renewal equations. The issue here is to determine limits of functions of the form $U \star h(t)$ as $t \to \infty$.

We begin with preliminaries on integrals of functions on the infinite axis $\mathbb{R}_+$. Recall that the Riemann integral $\int_0^t h(s) ds$ is constructed by Riemann sums on grids that become finer and finer (see Definition 86 below). The integral exists when h is continuous on $[0, t]$, or is bounded and has a countable number of discontinuities. Furthermore, the Riemann integral of h on $\mathbb{R}_+$ is defined by

$$\int_{\mathbb{R}_+} h(s) ds = \lim_{t\to\infty} \int_0^t h(s) ds, \tag{2.21}$$

provided the limit exits. In that case, h is *Riemann integrable* on $\mathbb{R}_+$.

The Key Renewal Theorem requires a slightly different notion of a function being *directly Riemann integrable* on $\mathbb{R}_+$. A DRI function is defined in Section 2.17, where an integral is constructed "directly" on the entire axis $\mathbb{R}_+$ by Riemann sums, analogously to the construction of a Riemann integral on a finite interval. A DRI function is Riemann integrable in the usual sense, but the converse is not true; see Exercise 32.

For our purposes, we only need the following properties from Proposition 88 below (also see Exercise 33).

Remark 34. A function $h : \mathbb{R}_+ \to \mathbb{R}$ is DRI in the following cases.
(a) $h(t) \geq 0$ is decreasing and Riemann integrable.
(b) h is continuous except possibly on a set of Lebesgue measure 0, and $|h(t)| \leq b(t)$, where b is DRI.

Here is the main result. Its proof is in Section 2.17.

Theorem 35. (Key Renewal Theorem) *If F is non-arithmetic and $h(t)$ is DRI, then*

$$\lim_{t\to\infty} U \star h(t) = \frac{1}{\mu} \int_{\mathbb{R}_+} h(s)ds. \tag{2.22}$$

Remark 36. This theorem is equivalent to Blackwell's Theorem 33, which asserts that $U(t+a)-U(t) \to a/\mu$. Indeed, Section 2.17 shows that Blackwell's theorem implies the Key Renewal Theorem. Conversely, (2.22) applied to $h(s) = 1(a < s < t + a)$ (so $h(t - s) = 1(t - a < s \leq t)$) yields Blackwell's theorem.

An analogous key renewal theorem for arithmetic F is as follows. It can also be proved by Blackwell's renewal theorem — with fewer technicalities — as suggested in Exercise 31.

Theorem 37. (Arithmetic Key Renewal Theorem) *If F is arithmetic with span d, then for any $u < d$,*

$$\lim_{n\to\infty} U \star h(u + nd) = \frac{d}{\mu} \sum_{k=0}^{\infty} h(u + kd),$$

provided the sum is absolutely convergent.

The next order of business is to show how the limit statement in the key renewal theorem applies to limits of time-dependent probabilities and expected values of stochastic processes. We know that any function $H(t)$ that satisfies a renewal equation has the form $H(t) = U \star h(t)$. It turns out that this functional form is "universal" in the following sense.

Proposition 38. *Any function $H(t)$ that is bounded on finite intervals and is 0 for $t < 0$ can be expressed as*

$$H(t) = U \star h(t), \qquad \text{where} \quad h(t) = H(t) - F \star H(t).$$

Proof. This follows since $U = F^{0\star} + U \star F$, and so

$$H = F^{0\star} \star H = (U - U \star F) \star H = U \star (H - F \star H).$$

Knowing that $U \star h(t)$ is a universal form for any function that is bounded on finite intervals, the remaining issue is, "How to relate $U \star h(t)$ to probabilities and expectations of stochastic processes?" A natural vehicle is the following type of stochastic process.

Definition 39. A real-valued stochastic process $X(t)$ is *crudely regenerative* at a positive random time T if

$$E[X(T+t)|T] = E[X(t)], \qquad t \geq 0, \tag{2.23}$$

and these expectations are finite.

An important connection between crudely regenerative processes and functions $U \star h(t)$ is as follows.

Proposition 40. *Suppose that $X(t)$ is a crudely regenerative process at T, which has the distribution F. If $E[X(t)]$ is bounded on finite intervals, then*

$$E[X(t)] = U \star h(t), \qquad \textit{where} \quad h(t) = E[X(t)\mathbf{1}(T > t)].$$

Proof. Applying Proposition 38 to $H(t) = E[X(t)]$, it follows that $E[X(t)] = U \star h(t)$, where

$$h(t) = H(t) - F \star H(t) = E[X(t)] - \int_{[0,t]} E[X(t-s)]dF(s).$$

By the crude regeneration property, $E[X(t)|T = s] = E[X(t-s)]$, $s \leq t$, and

$$h(t) = E[X(t)] - \int_{[0,t]} E[X(t)|T = s]dF(s) = E[X(t)\mathbf{1}(T > t)].$$

This completes the proof.

The family of crudely regenerative processes is very large; it includes ergodic Markov chains in discrete and continuous time, regenerative processes, and many functions of these processes as well. More details on these processes are in the next sections. Typically, $X(t)$ is a real-valued function of one or more stochastic processes. An important example is a probability $P\{Y(t) \in A\} = E[X(t)]$, when $X(t) = \mathbf{1}(Y(t) \in A)$.

The following major result is a version of the key renewal theorem that characterizes limiting distributions and expectations. Many applications in the next sections are based on this formulation.

Theorem 41. (Crude Regenerations) *Suppose that $X(t)$ is a real-valued process that is crudely regenerative at T, and define $M = \sup\{|X(t)| : t \leq T\}$. If T is non-arithmetic and M and MT have finite means, then*

$$\lim_{t\to\infty} E[X(t)] = \frac{1}{\mu}\int_{\mathbb{R}_+} h(s)ds, \tag{2.24}$$

where $h(t) = E[X(t)\mathbf{1}(T > t)]$.

Proof. Since $E[X(t)] = U \star h(t)$ by Proposition 40, where T has the nonarithmetic distribution F, the assertion (2.24) will follow by the key renewal theorem provided $h(t)$ is DRI.

To prove this, note that $|h(t)| \le b(t) = E[M\mathbf{1}(T > t)]$. Now, by the dominated convergence theorem in the Appendix and $E[M] < \infty$, we have $b(t) \downarrow 0$. Also,

$$\int_{\mathbb{R}_+} b(s)ds = E\Big[\int_0^T M\, ds\Big] = E[MT] < \infty. \tag{2.25}$$

Then $b(t)$ is DRI by Remark 34 (a), and so $h(t)$ is DRI by Remark 34 (b).

2.8 Regenerative Processes

The primary use of the key renewal theorem is in characterizing the limiting behavior of regenerative processes and their relatives via Theorem 41. This section covers limit theorems for regenerative processes, and the next three sections cover similar results for Markov chains, and processes with regenerative increments.

We begin by defining regenerative processes. Loosely speaking, a discrete- or continuous-time stochastic process is regenerative if there is a renewal process such that the segments of the process between successive renewal times are i.i.d. More precisely, let $\{X(t) : t \ge 0\}$ denote a continuous-time stochastic process with a state space S that is a metric space (e.g., the Euclidean space $\mathbb{R}^d$ or a Polish space; see the Appendix). This process need not be a jump process like the continuous-time Markov chains we discuss later. However, we assume that the sample paths of $X(t)$ are right-continuous with left-hand limits a.s. This ensures that the sample paths are continuous except possibly on a set of Lebesgue measure 0.

Let $N(t)$ denote a renewal process on $\mathbb{R}_+$, defined on the same probability space as $X(t)$, with renewal times T_n and inter-renewal times $\xi_n = T_n - T_{n-1}$, which have a distribution F with a finite mean μ.

Definition 42. For the process $\{(N(t), X(t)) : t \ge 0\}$, its sample path in the time interval $[T_{n-1}, T_n)$ is described by

$$\zeta_n = (\xi_n, \{X(T_{n-1} + t) : 0 \le t < \xi_n\}). \tag{2.26}$$

This ζ_n is the nth *segment* of the process. The process $X(t)$ is *regenerative over the times* T_n if its segments ζ_n are i.i.d.

Classic examples of regenerative processes are ergodic Markov chains in discrete and continuous time. An important fact that follows directly from the definition is that functions of regenerative processes inherit the regenerative property.

Remark 43. Inheritance of Regenerations. If $\widetilde{X}(t)$ with state space $\widetilde{S}$ is regenerative over T_n, then $X(t) = f(\widetilde{X}(t))$ is also regenerative over T_n, for any $f : \widetilde{S} \to S$.

For instance, we can express the distribution of a regenerative process $\widetilde{X}(t)$ as the expectation $P\{\widetilde{X}(t) \in B\} = E[X(t)]$, where $X(t) = \mathbf{1}(\widetilde{X}(t) \in B)$ (a function of $\widetilde{X}$) is a real-valued regenerative process.

To include the possibility that the first segment of the process $X(t)$ in the preceding definition may differ from the others, we say $X(t)$ is a *delayed* regenerative process if ζ_n are independent, and $\zeta_2, \zeta_3, \ldots$ have the same distribution, which may be different from the distribution of ζ_1. We discuss more general regenerative-like processes with stationary segments in Section 2.19.

Remark 44. Regenerative processes are crudely regenerative, but not vice versa.

Indeed, if $X(t)$ is regenerative over the times T_n, then $X(t)$ is crudely regenerative at T_1. Next, consider the process $X(t) = X_n(t)$, if $t \in [n-1, n]$ for some n, where $\{X_n(t) : t \in [0,1]\}$ for $n \geq 1$, are independent stochastic processes with identical mean functions ($E[X_n(t)] = E[X_1(t)]$ for each n), but non-identical variance functions. Clearly X is crudely regenerative at $T = 1$, but it is not regenerative.

To proceed, a few comments are in order concerning convergence in distribution. For a process $X(t)$ on a countable state space S, a probability measure P on S is the limiting distribution of $X(t)$ if

$$\lim_{t\to\infty} P\{X(t) \in B\} = \mathrm{P}(B), \qquad B \subset S. \tag{2.27}$$

This definition, however, is too restrictive for uncountable S, where (2.27) is not needed for all subsets B. In particular, when the state space S is the Euclidean space $\mathbb{R}^d$, then P on $S = \mathbb{R}^d$ is defined to be the limiting distribution of $X(t)$ if (2.27) holds for $B \in \mathcal{S}$ (the Borel sets of S) such that $\mathrm{P}(\partial B) = 0$, where ∂B is the boundary of B.

Equivalently, P on S is the *limiting distribution* of $X(t)$ if

$$\lim_{t\to\infty} E[f(X(t))] = \int_S f(x)\mathrm{P}(dx), \tag{2.28}$$

for any continuous function $f : S \to [0,1]$. This means that the distribution of $X(t)$ converges weakly to P (see Section 6.9 in the Appendix for more details on weak convergence).

We are now ready to apply Theorem 41 to characterize the limiting distribution of regenerative processes. For simplicity, assume throughout this section that the inter-renewal distribution F (for the times between regenerations) is non-arithmetic.

Theorem 45. (Regenerative Processes) *Suppose the process $X(t)$ on a metric state space S (e.g. $\mathbb{R}^d$) with Borel σ-field $\mathcal{S}$ is regenerative over T_n. For $f : S \to \mathbb{R}$ define $M = \sup\{|f(X(t))| : t \le T_1\}$. If M and MT_1 have finite means, then*

$$\lim_{t\to\infty} E[f(X(t))] = \frac{1}{\mu} E\Big[\int_0^{T_1} f(X(s))ds\Big]. \tag{2.29}$$

In particular, the limiting distribution of $X(t)$ is

$$\mathrm{P}(B) = \lim_{t\to\infty} P\{X(t) \in B\} = \frac{1}{\mu} E\Big[\int_0^{T_1} \mathbf{1}(X(s) \in B)ds\Big], \quad B \in \mathcal{S}. \tag{2.30}$$

Proof. Assertion (2.29) follows by Theorem 41, since $f(X(t))$ is regenerative over T_n and therefore it satisfies the crude-regeneration property. Clearly, (2.30) is a special case of (2.29).

Theorems 41 and 45 provide a framework for characterizing limits of expectations and probabilities of regenerative processes. For expectations, one must check that the maximum M of the process during an inter-renewal interval has a finite mean. The main step in applying these theorems, however, is to evaluate the integrals $\int_{\mathbb{R}_+} h(s)ds$ or $\int_S f(x)\mathrm{P}(dx)$. Keep in mind that one need not set up a renewal equation or check the DRI property for each application — these properties have already been verified in the proof of Theorem 41.

Theorem 45 and most of those to follow are true, with slight modifications, for delayed regenerative processes. This is due to the property in Exercise 42 that the limiting behavior of a delayed regenerative process is the same as the limiting behavior of the process after its first regeneration time T_1. Here is an immediate consequence of Theorem 45 and Exercise 42.

Corollary 46. (Delayed Regenerations) *Suppose the process $X(t)$ with a metric state space S is a delayed regenerative process over T_n. If $f : S \to \mathbb{R}$ is such that the expectations of $M = \sup\{|f(X(t))| : T_1 \le t \le T_2\}$ and $M\xi_2$ are finite, then*

$$\lim_{t\to\infty} E[f(X(t))] = \frac{1}{\mu} E\Big[\int_{T_1}^{T_2} f(X(s))\, ds\Big].$$

In particular, the limiting distribution of $X(t)$ is

$$\mathrm{P}(B) = \frac{1}{\mu} E\Big[\int_{T_1}^{T_2} \mathbf{1}(X(s) \in B)ds\Big], \quad B \in \mathcal{S}.$$

We end this section with applications of Theorem 45 to three regenerative processes associated with a renewal process.

Definition 47. *Renewal Process Trinity.* For a renewal process $N(t)$, the following three processes provide more information about renewal times:

$A(t) = t - T_{N(t)}$, the *backward recurrence time* at t (or the *age*), which is the time since the last renewal prior to t.

$B(t) = T_{N(t)+1} - t$, the *forward recurrence time* at t (or the *residual renewal time*), which is the time to the next renewal after t.

$L(t) = \xi_{N(t)+1} = A(t) + B(t)$, *length of the renewal interval* covering t.

For instance, a person arriving at a bus stop at time t would have to wait $B(t)$ minutes for the next bus to arrive, or a call-center operator returning to answer calls at time t would have to wait for a time $B(t)$ before the next call. Also, if a person begins analyzing an information string at a location t looking for a certain character (or pattern), then $A(t)$ and $B(t)$ would be the distances to the left and right of t where the next character occurs.

Note that the three-dimensional process $(A(t), B(t), L(t))$ is regenerative over T_n, and so is each process by itself. Each of the processes $A(t)$ and $B(t)$ is a continuous-time Markov process with piece-wise deterministic paths on the state space $\mathbb{R}_+$; see Exercises 34 and 35. A convenient expression for their joint distribution is, for $0 \le x < t$, $y \ge 0$,

$$P\{A(t) > x, B(t) > y\} = P\{N(t+y) - N((t-x)-) = 0\}. \tag{2.31}$$

This is simply the probability of no renewals in $[t-x, t+y]$. Although this probability is generally intractable, one can show that it is the solution of a renewal equation, and so it has the form $U \star h(t)$; see Exercises 36 and 37.

Example 48. Trinity in Equilibrium. One can obtain the limiting distributions of $A(t)$ and $B(t)$ separately from Theorem 45. Instead, we will derive their joint limiting distribution. Since $(A(t), B(t))$ is regenerative over T_n, Theorem 41 yields

$$\lim_{t\to\infty} P\{A(t) > x, B(t) > y\} = 1 - \frac{1}{\mu}\int_0^{x+y} [1 - F(s)]ds, \tag{2.32}$$

since, by the definitions of the variables,

$$h(t) = P\{A(t) > x, B(t) > y, T_1 > t\} = P\{T_1 > t + y\}\mathbf{1}(t > x).$$

From (2.32), it immediately follows that

$$\lim_{t\to\infty} P\{A(t) \le x\} = \lim_{t\to\infty} P\{B(t) \le x\} = \frac{1}{\mu}\int_0^{x} [1 - F(s)]ds. \tag{2.33}$$

This limiting distribution, which is called the *equilibrium distribution* associated with F, is important in other contexts. We will see its significance in Section 2.15 for stationary renewal processes.

One can also obtain the limiting distribution of $L(t) = A(t) + B(t)$ by Theorem 41. Namely,

$$\lim_{t\to\infty} P\{L(t) \le x\} = \frac{1}{\mu} \int_{[0,x]} s dF(s), \tag{2.34}$$

since

$$h(t) = P\{L(t) \le x, T_1 > t\} = P\{T_1 \le x, T_1 > t\} = (F(x) - F(t))\mathbf{1}(x > t).$$

Alternatively, one can derive (2.34) directly from (2.32).

Additional properties of the three regenerative processes $A(t)$, $B(t)$ and $L(t)$ are in Exercises 34–41. These processes are especially nice for a Poisson process.

Example 49. Poisson Recurrence Times. If $N(t)$ is a Poisson process with rate λ, then from (2.31)

$$P\{A(t) > x, B(t) > y\} = e^{-\lambda(x+y)}, \qquad 0 \le x < t, y \ge 0, \tag{2.35}$$

which is the Poisson probability of no renewals in an interval of length $x+y$. In particular, setting $x = 0$, and then $y = 0$, yields

$$P\{B(t) > y\} = e^{-\lambda y}, \qquad P\{A(t) > x\} = e^{-\lambda x}\mathbf{1}(x < t).$$

Thus $B(t)$ is exponentially distributed with rate λ; this also follows by the memoryless property of the exponential distribution (Exercise 1 in Chapter 3). Note that $A(t)$ has the same exponential distribution, but it is truncated at $x = t$. The limiting distribution of each of these processes, however, is exponential with rate λ. Since $L(t) = A(t) + B(t)$, its distribution can be obtained from (2.35); its mean is shown in Exercise 39.

Even though recurrence time processes $A(t)$ and $B(t)$ are typically not tractable for a fixed t, their equilibrium distribution F_e in (2.33) may be.

Example 50. Uniformly Distributed Renewals. Suppose $N(t)$ is a renewal process with uniform inter-renewal distribution $F(x) = x$, for $x \in [0, 1]$. Its associated equilibrium distribution (2.33) is simply $F_e(x) = 2x - x^2$.

Interestingly, $F_e(x) \ge F(x)$ for each x. That is, the distribution F_e for the forward recurrence time $B(t)$ in equilibrium is greater than the distribution F of the forward recurrence time $B(0) = \xi_1$ at time 0. This means that $B(t)$ in equilibrium is *stochastically smaller* than $B(0)$. This is due to the fact that the failure rate $F'(x)/(1 - F(x)) = 1/(1 - x)$ of F is increasing. Compare this property with the inspection paradox in Exercise 39.

2.9 Limiting Distributions for Markov Chains

This section covers the classical renewal argument for determining the limiting distributions of ergodic Markov chains. The argument uses limit theorems in the preceding section, which are manifestations of the key renewal theorem for regenerative processes. We present a similar characterization of limiting distributions for continuous-time Markov chains in Chapter 4.

Assume that X_n is an ergodic Markov chain on a countable state space S, with limiting distribution

$$\pi_j = \lim_{n\to\infty} P\{X_n = j\}, \qquad j \in S,$$

which does not depend on X_0. Recall that Theorems 59 and 54 in Chapter 1 established that the limiting distribution is also the stationary distribution and it is the unique distribution π that satisfies the balance equation $\pi = \pi \mathrm{P}$. They also showed (via a coupling proof) that the stationary distribution has the following form, which we will now prove by a classical renewal argument.

Theorem 51. (Markov Chains) *The ergodic Markov chain X_n has a unique limiting distribution given as follows: for a fixed $i \in S$,*

$$\pi_j = \frac{1}{\mu_i} E\Big[\sum_{n=0}^{\tau_1(i)-1} \mathbf{1}(X_n = j) \Big| X_0 = i \Big], \qquad j \in S, \tag{2.36}$$

where $\mu_i = E[\tau_1(i)|X_0 = i]$. Another expression for this probability is

$$\pi_j = \frac{1}{\mu_j}, \qquad j \in S. \tag{2.37}$$

Proof. We will prove this by applying the key renewal theorem. The main idea is that the strong Markov property ensures that X_n is a (discrete-time) delayed regenerative process over the times $0 < \tau_1(i) < \tau_2(i) < \dots$ at which X_n enters a fixed state i (Theorem 67 in Chapter 1). In light of this fact, the assertion (2.36) follows by Corollary 46. Also, setting $i = j$ in (2.36) yields (2.37), since the sum in (2.36) is the sojourn time in state j, which is 1.

2.10 Processes with Regenerative Increments

Many cumulative cost or utility processes associated with ergodic Markov chains and regenerative processes can be formulated as processes with regenerative increments. These processes are basically random walks with auxiliary paths or information. We will show that the classical SLLN and central limit theorem for random walks extend to processes with regenerative increments.

This section presents a SLLN based on material in Section 2.2, and Section 2.13 presents a central limit theorem. Functional central limit theorems for random walks and processes with regenerative increments are the topic of Section 5.9 in Chapter 5.

For this discussion, $N(t)$ will denote a renewal process whose inter-renewal times $\xi_n = T_n - T_{n-1}$ have a distribution F and finite mean μ.

Our focus will be on the following processes that are typically associated with cumulative information of regenerative processes.

Definition 52. Let $Z(t)$ be a real-valued process with $Z(0) = 0$ defined on the same probability space as a renewal process $N(t)$. For the two-dimensional process $\{(N(t), Z(t)) : t \geq 0\}$, its increments in the time interval $[T_{n-1}, T_n)$ are described by

$$\zeta_n = (\xi_n, \{Z(t + T_{n-1}) - Z(T_{n-1}) : 0 \leq t < \xi_n\}).$$

The process $Z(t)$ has *regenerative increments over the times* T_n if ζ_n are i.i.d. A process with "delayed" regenerative increments is defined in the obvious way, where the distribution ζ_1 is different from the others.

Under this definition, $(T_n - T_{n-1}, Z(T_n) - Z(T_{n-1}))$ are i.i.d. These i.i.d. increments of N and Z leads to many nice limit theorems based on properties of random walks.

A primary example of a process with regenerative increments is a cumulative functional $Z(t) = \int_0^t f(X(s))ds$, where $X(t)$ is regenerative over T_n and $f(i)$ is a cost rate (or utility rate) when the process $X(t)$ is in state i.

Hereafter, assume that $Z(t)$ is a process with regenerative increments over T_n. Keep in mind that $Z(0) = 0$. Although the distribution and mean of $Z(t)$ are generally not tractable for computations, we do have a Wald identity for some expectations.

Proposition 53. (Wald Identity for Regenerations) *For the process $Z(t)$ with regenerative increments and finite $a = E[Z(T_1)]$,*

$$E[Z(T_{N(t)+1})] = aE[N(t) + 1], \qquad t \geq 0. \tag{2.38}$$

Proof. By Theorem 22,

$$E[Z(T_{N(t)+1})] = E\left[\sum_{n=0}^{N(t)} [Z(T_{n+1}) - Z(T_n)]\right] = a \cup (t).$$

By the classical SLLN, we know that

$$n^{-1}Z(T_n) = n^{-1}\sum_{k=1}^{n} \big[Z(T_k) - Z(T_{k-1})\big] \to E[Z(T_1)], \qquad \text{a.s. as } n \to \infty.$$

This extends to $Z(t)$ as follows, which is a special case of Corollary 14. Here

$$M_n = \sup_{T_{n-1} < t \leq T_n} |Z(t) - Z(T_{n-1})|, \qquad n \geq 1.$$

Theorem 54. *For the process $Z(t)$ with regenerative increments, suppose the mean of M_n is finite, and $E[T_1]$ and $a = E[Z(T_1)]$ exist, but are not both infinite. Then $t^{-1}Z(t) \to a/\mu$, a.s. as $t \to \infty$.*

The next result is a special case of Theorem 54 for a functional of a regenerative process, where the limiting average is expressible in terms of the limiting distribution of the regenerative process. The convergence of the expected value per unit time is also shown in (2.40); a refinement of this is given in Theorem 85 below.

Theorem 55. *Let $X(t)$ be a regenerative process over T_n with a metric state space S (e.g. $\mathbb{R}^d$), and let* P *denote the limiting distribution of $X(t)$ given by (2.30), where $\mu = E[T_1]$ is finite. Suppose $f : S \to \mathbb{R}$ is such that $\int_0^{T_1} |f(X(s))|ds$ and $|f(\overline{X})|$ have finite means, where $\overline{X}$ has the distribution* P*. Then*

$$\lim_{t\to\infty} t^{-1} \int_0^t f(X(s))ds = E[f(\overline{X})], \qquad a.s. \tag{2.39}$$

If, in addition, $E[T_1 \int_0^{T_1} |f(X(s))|ds]$ is finite, and T_1 has a non-arithmetic distribution, then

$$\lim_{t\to\infty} t^{-1} E\Big[\int_0^t f(X(s))ds \Big] = E[f(\overline{X})]. \tag{2.40}$$

Proof. Applying Theorem 54 to $Z(t) = \int_0^t f(X(s))ds$ and noting that

$$E[M_n] \le E[\int_0^{T_1} |f(X(s))|ds] < \infty,$$

we obtain $t^{-1}Z(t) \to E[Z(T_1)]/\mu$. Then (2.39) follows since by expression (2.29) for P,

$$\begin{aligned} E[Z(T_1)]/\mu &= \frac{1}{\mu} E\Big[\int_0^{T_1} f(X(s))ds \Big] \\ &= \int_S f(x)\mathrm{P}(dx) \;=\; E[f(\overline{X})]. \end{aligned}$$

To prove (2.40), note that $E[f(X(t))] \to E[f(\overline{X})]$ by Theorem 45. Then (2.40) follows by the fact that $t^{-1} \int_0^t g(s)ds \to c$ if $g(t) \to c$.

Remark 56. Limiting Averages as Expected Values. The limit (2.39) as an expected value is a common feature of many strong laws of large numbers when $f(X(t)) \xrightarrow{d} f(\overline{X})$. However, there are non-regenerative processes that satisfy the strong law (2.39), but not (2.40).

2.11 Average Sojourn Times in Regenerative Processes

We now show how SLLNs yield fundamental formulas, called Little laws, that relate the average sojourn times in queues to the average input rate and average queue length. We also present similar formulas for average sojourn times of a regenerative process in a region of its state space.

Consider a general service system or input-output system where discrete items (e.g., customers, jobs, data packets) are processed, or simply visit a location for a while. The items arrive to the system at times τ_n that form a point process $N(t)$ on $\mathbb{R}_+$ (it need not be a renewal process). Let W_n denote the total time the nth item spends in the system. Here the waiting or sojourn time W_n includes the item's service time plus any delay waiting in queue for service. The item exits the system at time $\tau_n + W_n$. Then the quantity of items in the system at time t is

$$Q(t) = \sum_{n=1}^{\infty} \mathbf{1}(\tau_n \le t < \tau_n + W_n), \qquad t \ge 0.$$

There are no assumptions concerning the processing or visits of the items or the stochastic nature of the variables W_n and τ_n, other than their existence. For instance, items may arrive and depart in batches, an item may reenter for multiple services, or the items may be part of a larger network that affects their sojourns.

We will consider the following three standard system performance parameters:

$$\begin{aligned} L &= \lim_{t\to\infty} t^{-1} \int_0^t Q(s)ds \qquad \text{average quantity in the system,} \\ \lambda &= \lim_{t\to\infty} t^{-1} N(t) \qquad \text{arrival rate,} \\ W &= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} W_k \qquad \text{average waiting time.} \end{aligned}$$

There are many diverse systems in which two of the averages L, λ, and W exist, and the issue is whether the third one exists. We will consider this issue under the following assumption, which is very natural for most queueing systems.

Empty-System Assumption. Let T_n denote the nth time at which an item arrives to an empty system, i.e., $Q(T_n-) = 0$ and $Q(T_n) > 0$. Assume the times T_n form a point process on $\mathbb{R}_+$ such that the limit $\mu = \lim_{n\to\infty} n^{-1} T_n$ exists and is positive.[4] This simply says that the system empties out infinitely often, and it does so at times that have a limiting average.

[4] Keep in mind that the arrival process $N(t)$ is "not" the counting process associated with these empty times T_n.

Theorem 57. (Little Law) *Suppose the system described above satisfies the empty-system assumption. If any two of the averages L, λ or W exists, then the other one also exists, and $L = \lambda W$.*

Proof. With no loss in generality, we may assume the system is empty at time 0 and an item arrives. We begin with the key observation that in the time interval $[0, T_n)$, all of the $\nu_n = N(T_n-)$ items that arrive in the interval also depart by the empty-system time T_n, and their total waiting time is

$$\sum_{k=1}^{\nu_n} W_k = \sum_{k=1}^{\infty} \int_0^{T_n} \mathbf{1}(\tau_k \le s < \tau_k + W_k) ds = \int_0^{T_n} Q(s) ds. \tag{2.41}$$

The first equality follows since the system is empty just prior to T_n, and the second equality follows from the definition of $Q(t)$.

Also, observe that under the assumptions $t^{-1}N(t) \to \lambda$ and $T_n/n \to \mu$,

$$n^{-1}\nu_n = T_n^{-1} N(T_n-)(n^{-1}T_n) \to \lambda\mu. \tag{2.42}$$

First assume that λ and W exist. Then by (2.41), we have

$$n^{-1} \int_0^{T_n} Q(s)\, ds = (\nu_n/n)\nu_n^{-1} \sum_{k=1}^{\nu_n} W_k \to \lambda\mu W.$$

Therefore, an application of Theorem 13 to the nondecreasing process $Z(t) = \int_0^t Q(s)\, ds$ and the times T_n yields

$$L = \lim_{t\to\infty} t^{-1} Z(t) = \lambda W.$$

Next, assume that λ and L exist. Then by (2.41),

$$n^{-1} \sum_{k=1}^{\nu_n} W_k = (n^{-1}T_n)\Big(T_n^{-1} \int_0^{T_n} Q(s) ds\Big) \to \mu L, \qquad \text{a.s. as } t \to \infty.$$

Now, by a discrete-time version of Theorem 13 for the nondecreasing process $Z_n' = \sum_{k=1}^n W_k$ and integer-valued indices ν_n, which satisfy (2.42), it follows that

$$W = \lim_{n\to\infty} n^{-1} Z_n' = L/\lambda.$$

Thus, W exists and $L = \lambda W$.

Exercise 23 shows that if L and W exist then λ exists and $L = \lambda W$.

The preceding Little law applies to a wide variety of queueing systems as long as two of the averages λ, L or W exist. Here are a few examples.

Example 58. Regenerative Processing System. Suppose the system described above satisfies the empty-system assumption, the arrival process is a renewal

process with a finite mean $1/\lambda$, and $Q(t)$ is a regenerative process over the empty-system times T_n. Assume that T_1 and $\int_0^{T_1} Q(s)ds$ have finite means.

By the SLLN for the renewal input process, the arrival rate is λ. Also, applying Theorem 54 to $Z(t) = \int_0^t Q(s)ds$, we have $L = E[\int_0^{T_1} Q(s)ds]/E[T_1]$. Therefore, by Theorem 57, the average waiting time W exists and $L = \lambda W$; that is, $W = E[\int_0^{T_1} Q(s)ds]/(\lambda E[T_1])$.

In some queueing systems, the Little law $L = \lambda W$ we have been discussing for averages has an analogue in which the averages are means.

Example 59. Little Laws for Means. Consider the system in the preceding example with the additional assumption that the sequence of sojourn times W_n is regenerative over the discrete times $\nu_n = N(T_n-)$. Since $Q(t)$ is regenerative over T_n, and W_n is regenerative over ν_n,

$$Q(t) \xrightarrow{d} \overline{Q} \text{ as } t \to \infty, \quad \text{and} \quad W_n \xrightarrow{d} \overline{W} \text{ as } n \to \infty,$$

where the distributions of $\overline{Q}$ and $\overline{W}$ are described in Theorem 45. Furthermore, by Theorem 55,

$$L = \lim_{t\to\infty} t^{-1} \int_0^t Q(s)ds = E[\overline{Q}] \quad \text{a.s.},$$

$$W = \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} W_k = E[\overline{W}] \quad \text{a.s.}$$

Also, the renewal arrival rate λ can be represented as $\lambda = E[\tilde{N}(1)]$, where $\tilde{N}(t)$ is a stationary version of $N(t)$ as described in Theorem 76 below. Then the Little law $L = \lambda W$ that holds for averages has the following analogue for expected values:

$$E[\overline{Q}] = E[\tilde{N}(1)]E[\overline{W}].$$

Example 60. $G/G/1$ Queueing System. A general example of a regenerative queueing system is a $G/G/1$ system, where arrivals form a renewal process with mean inter-arrival time $1/\lambda$, the services times are i.i.d., independent of the arrivals, and customers are served by a single server under a first-in-first-out (FIFO) discipline. Assume that the mean service time is less than the mean inter-arrival time, and that T_1 and $\int_0^{T_1} Q(s)ds$ have finite means. In this case, the sojourn times W_n are regenerative over $\nu_n = N(T_n-)$, and W exists by Theorem 118 in Chapter 4. Then it follows by Theorem 57 that the average queue length L exists and $L = \lambda W$.

Special cases of the $G/G/1$ system are an $M/G/1$ system when the arrival process is a Poisson process, a $G/M/1$ system when the service times are exponentially distributed, and an $M/M/1$ system when the arrivals are Poisson and the service times are exponential.

Theorem 57 also yields expected waiting times in Jackson networks, which we discuss in Chapter 5.

There are several Little laws for input-output systems and general utility processes not related to queueing [101]. The next result is an elementary but very useful example.

Let $X(t)$ be a regenerative process over T_n with state space S. Assume $X(t)$ is a pure jump process (piecewise constant paths, etc.) with a limiting distribution $p(B) = \lim_{n\to\infty} P\{X(t) \in B\}$, which is known. Let B denote a fixed subset of the state space whose complement B^c is not empty. The expected number of times that $X(t)$ enters B between regenerations is

$$\gamma(B) = E\Big[\sum_n \mathbf{1}\Big(X(\tau_{n-1}) \in B^c, X(\tau_n) \in B, \tau_n \in (T_1, T_2]\Big)\Big],$$

where τ_n is the time of the nth jump of $X(t)$. The expected number of transitions of $X(t)$ between regenerations is $\gamma(S)$, which we assume is finite.

Consider the average sojourn time of $X(t)$ in B defined by

$$W(B) = \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} W_k(B),$$

where $W_n(B)$ is its sojourn time in B at its nth visit to the set.

Proposition 61. (Sojourns in Regenerative Processes) *For the regenerative process $X(t)$ defined above, its average sojourn time in B exists and is $W(B) = p(B)\gamma(S)/\gamma(B)$.*

Proof. Consider $Q(t) = \mathbf{1}(X(t) \in B)$ as an artificial queueing process that only takes values 0 or 1. Clearly $Q(t)$ is regenerative over T_n, since $X(t)$ is regenerative over T_n; and $Q(t)$ satisfies the empty-system assumption. Now, the limiting average of $Q(t)$ is

$$L = \lim_{t\to\infty} t^{-1} \int_0^t \mathbf{1}(X(s) \in B)\, ds = p(B).$$

The arrival rate λ is the rate $\gamma(B)/\gamma(S)$ at which $X(t)$ enters B. Thus, Theorem 57 yields $p(B) = \lambda W(B) = (\gamma(B)/\gamma(S))W(B)$, which proves the assertion.

2.12 Batch-Service Queueing System

For service systems that process items in batches, a basic problem is to determine when to serve batches and how many items should be in the batches. This is a dynamic control problem or a Markov decision problem. We will

address this problem for a particular setting and show how to obtain certain control parameters by using a SLLN for regenerative processes.

Consider a single-server station that serves items or customers in batches as follows. Items arrive to the station according to a Poisson process with rate λ and they enter a queue where they wait to be served. The server can serve items in batches, and the number of items in a batch can be any number less than or equal to a fixed number $K \leq \infty$ (the service capacity). The service times of the batches are independent, identically distributed and do not depend on the arrival process or the batch size (think of a computer, bus, or truck). Only one batch can be served at a time and, during a service, additional arrivals join the queue.

The server observes the queue length at the times at which an arrival occurs and the server is idle, or whenever a service is completed. At each of these observation times, the server takes one of the following actions:

- No items are served.
- A batch consisting of all or a portion of the items waiting is served (the batch size cannot exceed $i \wedge K$, where i is the queue length).

These actions control the batch sizes and the timing of the services. If the server takes the first action, the next control action is taken when the next item arrives, and if the server takes the second action to serve a batch, the next control action is taken when the service is completed. A *control policy* is a rule for selecting one of these actions at each observation time. The general problem is to find a control policy that minimizes the average cost (or discounted cost) of serving items over an infinite time horizon.

This Markov decision problem was solved in [33] for natural holding and service cost functions for both the average-cost and discounted-cost criteria. In either case, the main result is that there is an optimal M-policy of the following form: At each observation time when the queue length is i, do not serve any items if $i < M$, and serve a batch of $i \wedge K$ items if $i \geq M$. Here M is an "optimal" level that is a function of the costs.

We will now describe an optimal level M for a special case. Suppose the system is to operate under the preceding M-policy, where the capacity K is infinite, and the service times are exponentially distributed with rate γ. Assume there is cost C for serving a batch and a cost hi per unit time for holding i items in the queue.

Theorem 62. *Under the preceding assumptions, the average cost per unit time is minimized by setting the level M to be*

$$M = \min\{m \geq 0 : m(m+1) \geq 2[(\lambda/\gamma)^2 p^m + C\lambda/h]\}, \tag{2.43}$$

where $p = \lambda/(\lambda + \gamma)$.

Proof. Let $X_m(t)$ denote the number of items in the queue at time t, when the system is operated under an m-policy. Let T_n denote the time at which the server initiates the nth service, and let $N(t)$ denote the associated counting

process. For simplicity, assume that a service has just been completed at time 0, and let $T_0 = 0$.

We will show that, under the m-policy with exponential service times, the T_n are renewal times; and the service plus holding cost in $[0, t]$ is

$$Z_m(t) = CN(t) + h \int_0^t X_m(s)\, ds.$$

Next, we will establish the existence of the average cost

$$f(m) = \lim_{t\to\infty} t^{-1} Z_m(t),$$

and then show that $f(m)$ is minimized at the M specified in (2.43).

Let Q_n denote the queue length at the nth service initiation time T_n, for $n \geq 0$. Note that Q_n is just the number of arrivals that occur during the nth service period, since all the waiting items are served in the batch. Because of the exponential service times, Q_n, for $n \geq 1$, are i.i.d. with

$$P\{Q_n = i\} = \int_{\mathbb{R}_+} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \gamma e^{-\gamma t}\, dt = p^i(1-p), \qquad i \geq 0. \tag{2.44}$$

For notational convenience, assume the initial queue length Q_0 has this distribution and is independent of everything else.

Next, observe that the quantity Q_n determines the time $\xi_{n+1} = T_{n+1} - T_n$ until the next service initiation. Specifically, if $Q_n \geq m$, then ξ_{n+1} is simply a service time; and if $Q_n = i < m$, then ξ_{n+1} is the time it takes for $m - i$ more items to arrive plus a service time. Since the Q_n are i.i.d., it follows that T_n are renewal times. Furthermore, conditioning on Q_0, the inter-renewal distribution is

$$P\{\xi_1 \leq t\} = P\{Q_0 \geq m\} G_\gamma(t) + \sum_{i=0}^{m-1} P\{Q_0 = i\} G_\lambda^{(m-i)\star} \star G_\gamma(t),$$

where G_λ is an exponential distribution with rate λ.

Then using the distribution (2.44), the inter-renewal distribution and its mean (indexed by m) are:[5]

$$P\{\xi_1 \leq t\} = p^m G_\gamma(t) + (1-p) \sum_{i=0}^{m-1} p^i G_\lambda^{(m-i)\star} \star G_\gamma(t),$$
$$\mu_m = \gamma^{-1} + m\lambda^{-1} - (1-p^m)\gamma^{-1}.$$

Now, the increasing process $Z_m(t)$ is such that $Z_m(T_n) - Z_m(T_{n-1})$, for $n \geq 1$, are i.i.d. with mean

[5] The identity $\sum_{i=1}^k i p^{i-1} = \frac{d}{dp}(\sum_{i=0}^k p^i)$ is used to derive the formula for μ_m.

$$E[Z_m(T_1)] = C + hE\Big[\int_0^{T_1} X_m(s)\,ds\Big].$$

Then by Theorem 13, the average cost, as a function of m, is

$$f(m) = \lim_{t\to\infty} t^{-1} Z_m(t) = \mu_m^{-1} E[Z_m(T_1)].$$

To evaluate this limit, let $\tilde{N}(t)$ denote the Poisson arrival process with exponential inter-arrival times $\tilde{\xi}_n$, and let τ denote an exponential service time with rate γ. Then we can write

$$\int_0^{T_1} X_m(s)\,ds = Q_0\tau + \int_0^{\tau} \tilde{N}(s)\,ds + \sum_{i=0}^{m-1} \mathbf{1}(Q_0 = i) \sum_{k=1}^{m-i} (i+k-1)\tilde{\xi}_k. \quad (2.45)$$

The first two terms on the right-hand side represent the holding time of items during the service period, and the last term represents the holding time of items (which is 0 if $Q_0 \geq m$) prior to the service period. Then from the independence of Q_0 and τ and Exercise 14,

$$E\Big[\int_0^{T_1} X_m(s)\,ds\Big] = \Big[1/(1-p)\gamma + \lambda/\gamma^2 + (1-p)\lambda^{-1} \sum_{i=0}^{m-1} p^i \sum_{k=1}^{m-i} (i+k-1)\Big].$$

Substituting this in the expression above for $f(m)$, it follows from lengthy algebraic manipulations that

$$f(m+1) - f(m) = h(1-p^{m+1})D_m/(\lambda^2\mu_m\mu_{m+1}),$$

where $D_m = m(m+1) - 2[(\lambda/\gamma)^2 p^m + C\lambda/h]$. Now, D_m is increasing in m and the other terms in the preceding display are positive. Therefore $f(m)$ is monotone decreasing and then increasing and has a unique minimum at $M = \min\{m : D_m \geq 0\}$, which is equivalent to (2.43).

Analysis similar to that above yields a formula for the optimal level M when the service capacity K is finite; see Exercise 52 in Chapter 4.

2.13 Central Limit Theorems

For a real-valued process $Z(t)$ with regenerative increments over T_n, we know that under the conditions in Theorem 54,

$$Z(t)/t \to a = E[Z(T_1)]/E[T_1] \quad \text{a.s.} \quad \text{as } t \to \infty.$$

In other words, $Z(t)$ behaves asymptotically like at. Further information about this behavior can be obtained by characterizing the limiting

distribution of the difference $Z(t) - at$ as $t \to \infty$. We will now present a central limit theorem that gives conditions under which this limiting distribution is a normal distribution. Special cases of this result are CLT's for renewal and Markovian processes.

We will obtain the CLT for regenerative processes by applying the following classical CLT for sums of independent random variables (which is proved in standard probability texts). The analysis will involve the notion of convergence in distribution of random variables; see Section 6.9 in the Appendix.

Theorem 63. (Classical CLT) *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with mean μ and variance $\sigma^2 > 0$, and define $S_n = \sum_{m=1}^{n}(X_m - \mu)$. Then*

$$P\{S_n/n^{1/2} \leq x\} \to \int_{-\infty}^{x} \frac{e^{-y^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\, dy, \qquad x \in \mathbb{R}.$$

This convergence in distribution is denoted by

$$S_n/n^{1/2} \xrightarrow{d} N(0, \sigma^2), \qquad \text{as } n \to \infty,$$

where $N(0, \sigma^2)$ is a normal random variable with mean 0 and variance σ^2.

We will also use the following result for randomized sums; see for instance p.216 in [26]. This result and the ones below are contained in the functional central limit theorems in Chapter 5, which focus on the convergence of entire stochastic processes instead of random variables.

Theorem 64. (Anscombe) *In the context of Theorem 63, let $N(t)$ be an integer-valued process defined on the same probability space as the X_n, where $N(t)$ may depend on the X_n. If $t^{-1}N(t) \xrightarrow{d} c$, where c is a positive constant, then*

$$S_{N(t)}/t^{1/2} \xrightarrow{d} N(0, c\sigma^2), \text{ as } t \to \infty.$$

The following is a regenerative analogue of the classical CLT.

Theorem 65. (Regenerative CLT) *Suppose $Z(t)$ is a real-valued process with regenerative increments over T_n such that $\mu = E[T_1]$ and $a = E[Z(T_1)]/\mu$ are finite. In addition, let*

$$M_n = \sup_{T_{n-1} < t \leq T_n} |Z(t) - Z(T_{n-1})|, \quad n \geq 1,$$

and assume $E[M_1]$ and $\sigma^2 = Var[Z(T_1) - aT_1]$ are finite, and $\sigma > 0$. Then

$$(Z(t) - at)/t^{1/2} \xrightarrow{d} N(0, \sigma^2/\mu), \qquad \text{as } t \to \infty. \tag{2.46}$$

Proof. The process $Z(t)$ is "asymptotically close" to $Z(T_{N(t)})$, when dividing them by $t^{1/2}$, because their difference is bounded by $M_{N(t)+1}$, which is a regenerative process that is 0 at regeneration times. Consequently, the normalized process

$$\tilde{Z}(t) = (Z(t) - at)/t^{1/2}$$

should have the same limit as the process

$$Z'(t) = (Z(T_{N(t)}) - aT_{N(t)})/t^{1/2}.$$

Based on this conjecture, we will prove

$$Z'(t) \xrightarrow{d} N(0, \sigma^2/\mu), \qquad \text{as } t \to \infty, \tag{2.47}$$

$$|\tilde{Z}(t) - Z'(t)| \xrightarrow{d} 0, \qquad \text{as } t \to \infty. \tag{2.48}$$

Then it will follow by a standard property of convergence in distribution (see Exercise 53 of Chapter 5), that

$$\tilde{Z}(t) = Z'(t) + (\tilde{Z}(t) - Z'(t)) \xrightarrow{d} N(0, \sigma^2/\mu).$$

To prove (2.47), note that

$$Z'(t) = t^{-1/2} \sum_{n=1}^{N(t)} X_n,$$

where $X_n = Z(T_n) - Z(T_{n-1}) - a(T_n - T_{n-1})$. Since $Z(t)$ has regenerative increments over T_n, the X_n are i.i.d. with mean 0 and variance σ^2. Also, $t^{-1}N(t) \to 1/\mu$ by the SLLN for renewal processes. In light of these observations, Anscombe's theorem above yields (2.47).

To prove (2.48), note that

$$\tilde{Z}(t) - Z'(t) = t^{-1/2}[Z(t) - Z(T_{N(t)}) - a(t - T_{N(t)})].$$

Then letting $Y_n = M_n + a(T_{n+1} - T_n)$, it follows that

$$|\tilde{Z}(t) - Z'(t)| \le t^{-1/2} Y_{N(t)} = \sqrt{N(t)/t}\Big(N(t)^{-1/2} Y_{N(t)}\Big).$$

Since $Z(t)$ has regenerative increments, the Y_n are i.i.d., and so

$$n^{-1/2} Y_n \stackrel{d}{=} n^{-1/2} Y_1 \to 0 \quad \text{a.s.}$$

Using this and $N(t)/t \to 1/\mu$ a.s. in the preceding proves (2.48).

An important use of a CLT is to find confidence intervals for certain parameters. Here is an example.

Example 66. Confidence Interval for the Mean. Under the assumptions of Theorem 65, let us construct a confidence interval for the mean a of the regenerative-increment process $Z(t)$ based on observing the process up to a fixed time t. Assume (which is reasonable) that we do not know the variance parameter $\sigma^2 = \text{Var}[Z(T_1) - aT_1]$.

We first note that by the SLLN for $Z(t)$,

$$t^{-1}(Z(t) - at) \to 0, \qquad t^{-1/2}\sqrt{Z(t) - at} \to \sigma/\sqrt{\mu} \quad \text{a.s. as } t \to \infty.$$

Then the latter combined with Theorem 65 yields

$$(Z(t) - at)/\sqrt{Z(t) - at} \xrightarrow{d} N(0, 1).$$

Therefore, a confidence interval with approximate confidence coefficient $1-\alpha$ (e.g., see [99]) is

$$\Big[\frac{Z(t) - at}{\sqrt{Z(t) - at}} - z_{\alpha/2},\ \frac{Z(t) - at}{\sqrt{Z(t) - at}} + z_{\alpha/2}\Big],$$

where

$$P\{-z_{\alpha/2} \le N(0, 1) \le z_{\alpha/2}\} = 1 - \alpha.$$

Insights on simulation procedures for this and related models are in [43].

What would be an analogous confidence interval when $Z(t)$ is observed only at regeneration times? See Exercise 52.

Applying Theorem 65 to a regenerative-increment process involves determining conditions on the process under which the main assumptions are satisfied and then finding expressions for the normalization constants a and σ. Here are some examples.

Example 67. CLT for Renewal Processes. Suppose that $N(t)$ is a renewal process whose inter-renewal distribution has a finite mean μ and variance σ^2. Then $Z(t) = N(t)$ satisfies the assumptions in Theorem 65, and so

$$(N(t) - t/\mu)/t^{1/2} \xrightarrow{d} N(0, \sigma^2/\mu^3), \qquad \text{as } t \to \infty,$$

where

$$a = 1/\mu, \qquad \mathrm{Var}[Z(T_1) - aT_1] = \sigma^2\mu^{-2}.$$

Example 68. CLT for Markov Chains. Let X_n be an ergodic Markov chain on S with limiting distribution π. Consider the sum

$$Z_n = \sum_{m=1}^{n} f(X_m), \qquad n \ge 0,$$

where $f(j)$ is a real-valued cost or utility for the process being in state j. For simplicity, fix an $i \in S$ and assume $X_0 = i$ a.s. Then Z_n has regenerative increments over the discrete times ν_n at which X_n enters state i. We will apply a discrete-time version of Theorem 65 to Z_n.

Accordingly, assume $\mu_i = E[\nu_1]$ and $E\Big[\max_{1\le n\le \nu_1} |Z_n|\Big]$ are finite. The latter is true when $E\Big[\sum_{n=1}^{\nu_1} |f(X_n)|\Big]$ is finite. In addition, assume

$$a = \frac{1}{\mu_i} E_i[Z_{\nu_1}] = \sum_{j \in S} \pi_j f(j), \quad \text{and} \quad \sigma^2 = \frac{1}{\mu_i} \text{Var}[Z_{\nu_1} - a\nu_1]$$

are finite and $\sigma > 0$. Letting $\tilde{f}(j) = f(j) - a$, Exercise 54 shows that

$$\sigma^2 = E[\tilde{f}(X_0)^2] + 2 \sum_{n=1}^{\infty} E[\tilde{f}(X_0)\tilde{f}(X_n)], \tag{2.49}$$

where $P\{X_0 = i\} = \pi_i$. Then Theorem 65 (in discrete time) yields

$$(Z_n - an)/n^{1/2} \xrightarrow{d} N(0, \sigma^2), \qquad \text{as } n \to \infty. \tag{2.50}$$

This result also applies to random functions of Markov chains as follows (see Exercise 33 in Chapter 4 for a related continuous-time version). Suppose

$$Z_n = \sum_{m=1}^{n} f(X_m, Y_m), \qquad n \geq 0,$$

where $f : S \times S' \to \mathbb{R}$, and Y_m are conditionally independent given X_n $(n \geq 0)$, and $P\{Y_m \in B | X_n, n \geq 0\}$ only depends on X_m and $B \in \mathcal{S}'$. Here S' need not be discrete. In this setting, the cost or utility $f(X_m, Y_m)$ at time m is partially determined by the auxiliary or environmental variable Y_m. Then the argument above yields the CLT (2.50). In this case, $a = \sum_{j \in S} \pi_j \alpha(j)$, $\alpha(j) = E[f(j, Y_1)]$, and

$$\sigma^2 = E[(f(X_0, Y_1) - \alpha(X_0))^2] + 2 \sum_{n=1}^{\infty} E[m(X_0, X_n)],$$

where $m(j, k) = E[(f(j, Y_1) - \alpha(j))(f(k, Y_2) - \alpha(k))]$.

2.14 Terminating Renewal Processes

In this section, we discuss renewal processes that terminate after a random number of renewals. Analysis of these terminating (or transient) renewal processes uses renewal equations and the key renewal theorem applied a little differently than above.

Consider a sequence of renewal times T_n with inter-renewal distribution F. Suppose that at each time T_n (including $T_0 = 0$), the renewals terminate with probability $1 - p$, or continue until the next renewal epoch with probability p. These events are independent of the preceding renewal times, but may depend on the future renewal times.

Under these assumptions, the total number of renewals ν over the entire time horizon $\mathbb{R}_+$ has the distribution

$$P\{\nu \geq n\} = p^n, \qquad n \geq 0,$$

and $E[\nu] = p/(1-p)$. The number of renewals in $[0,t]$ is

$$N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t, \nu \geq n), \qquad t \geq 0.$$

Of course $N(t) \to \nu$ a.s. Another quantity of interest is the time T_ν at which the renewals terminate.

We will also use the following equivalent formulation of this terminating renewal process. Assume that $N(t)$ counts renewals in which the independent inter-renewal times have an *improper* distribution $G(t)$, with $p = G(\infty) < 1$. Then p is the probability of another renewal and $1 - p = 1 - G(\infty)$ is the probability that an inter-renewal time is "infinite", which terminates the renewals. This interpretation is consistent with that above since necessarily $G(t) = pF(t)$, where F as described above is the conditional distribution of an inter-renewal time given that it is allowed (or is finite).

Similarly to renewal processes, we will address issues about the process $N(t)$ with the use of its *renewal function*

$$V(t) = \sum_{n=0}^{\infty} G^{n\star}(t) = \sum_{n=0}^{\infty} p^n F^{n\star}(t).$$

We first observe that the counting process $N(t)$ and the termination time T_ν are finite a.s., and their distributions and means are

$$\begin{array}{ll} P\{N(t) \geq n\} = G^{n\star}(t), & E[N(t)] = V(t) - 1, \\ P\{T_\nu \leq t\} = (1-p)V(t), & E[T_\nu] = p\mu/(1-p). \end{array} \tag{2.51}$$

To establish these formulas, recall that the events $\nu = n$ (to terminate at n) and $\nu > n$ (to continue to the $n+1$st renewal) are assumed to be independent of $T_1, \ldots, T_n$. Then

$$\begin{aligned} P\{N(t) \geq n\} = P\{\nu \geq n, T_n \leq t\} &= p^n F^{n\star}(t) = G^{n\star}(t), \\ E[N(t)] = \sum_{n=1}^{\infty} P\{N(t) \geq n\} &= V(t) - 1. \end{aligned}$$

Similarly, using the independence and $T_\nu = \sum_{n=0}^{\infty} \mathbf{1}(\nu = n) T_n$,

$$P\{T_\nu \le t\} = \sum_{n=0}^{\infty} P\{\nu = n, T_n \le t\} \;=\; (1-p)\sum_{n=0}^{\infty} p^n F^{n\star}(t),$$
$$E[T_\nu] = \sum_{n=1}^{\infty} P\{\nu = n\} E[T_n] \;=\; \mu p/(1-p).$$

Although a regular renewal function tends to infinity, the renewal function for a terminating process has a finite limit.

Remark 69. As $t \to \infty$

$$V(t) = \sum_{n=0}^{\infty} p^n F^{n*}(t) \to 1/(1-p).$$

Corollary 71 below describes the convergence rate.

We will now discuss limits of certain functions associated with the terminating renewal process. As in Proposition 31, it follows that $H(t) = V \star h(t)$ is the unique solution to the renewal equation

$$H(t) = h(t) + G \star H(t).$$

We will consider the limiting behavior of $H(t)$ for the case in which the limit

$$h(\infty) = \lim_{t\to\infty} h(t)$$

exists, which is common in applications. Since $V(t) \to 1/(1-p)$ and $h(t)$ is bounded on compact sets and converges to $h(\infty)$, it follows by dominated convergence that

$$\begin{aligned} H(t) = h \star V(t) \;&=\; h(\infty)V(t) + \int_{[0,t]} [h(t-s) - h(\infty)]dV(s) \\ &\to h(\infty)/(1-p) \qquad \text{as } t \to \infty. \end{aligned} \tag{2.52}$$

The next result describes the rate of this convergence under a few more technical conditions. Assume there is a positive β such that

$$\int_{\mathbb{R}_+} e^{\beta t} dG(t) = 1.$$

The existence of a unique β is guaranteed under the weak condition that $\int_{\mathbb{R}_+} e^{\beta t} dG(t)$ is finite for some $\beta > 0$. Indeed, this function of β is continuous and increasing and, being finite at one point, its range contains the set $[p, \infty)$; thus, it must equal 1 for some β. We also assume the distribution

$$F^{\#}(t) = \int_{[0,t]} e^{\beta s} dG(s)$$

is non-arithmetic and has a mean $\mu^{\#}$.

Theorem 70. *In addition to the preceding assumptions, assume the function* $e^{\beta t}[h(t) - h(\infty)]$ *is DRI. Then*

$$H(t) = h(\infty)/(1-p) + ce^{-\beta t}/\mu^{\#} + o(e^{-\beta t}), \quad \text{as } t \to \infty, \tag{2.53}$$

where $c = \int_{\mathbb{R}_+} e^{\beta s}[h(s) - h(\infty)]\, ds - h(\infty)/\beta$.

Proof. Multiplying the renewal equation $H = h + G \star H$ by $e^{\beta t}$ yields the renewal equation $H^{\#} = h^{\#} + F^{\#} \star H^{\#}$ where $H^{\#}(t) = e^{\beta t}H(t)$ and $h^{\#}(t) = e^{\beta t}h(t)$.

We can now describe the limit of $H(t) - h(\infty)/(1-p)$ by the limit of $H^{\#}(t) - v(t)$, where $v(t) = e^{\beta t}h(\infty)/(1-p)$. From Lemma 83 below,

$$H^{\#}(t) = v(t) + \frac{1}{\mu^{\#}} \int_{\mathbb{R}_+} \overline{h}(s)\, ds + o(1), \quad \text{as } t \to \infty, \tag{2.54}$$

provided $\overline{h}(t) = h^{\#}(t) - v(t) + F^{\#} \star v(t)$ is DRI. In this case,

$$\overline{h}(t) = e^{\beta t}[h(t) - h(\infty)] - \Big[\frac{h(\infty)e^{\beta t}}{1-p}(p - G(t))\Big]. \tag{2.55}$$

Now, the first term on the right-hand side is DRI by assumption. Also,

$$e^{\beta t}(p - G(t)) \le \int_{(t,\infty)} e^{\beta s} dG(s) = 1 - F^{\#}(t).$$

This bound is decreasing to 0 and its integral is $\mu^{\#}$, and so the last term in brackets in (2.55) is DRI. Thus $\overline{h}(t)$ is DRI. Finally, an easy check shows that $\int_{\mathbb{R}_+} \overline{h}(s)\, ds = c$, the constant in (2.53). Substituting this in (2.54) and dividing by $e^{\beta t}$ yields (2.53).

Corollary 71. *Under the assumptions preceding Theorem 70,*

$$V(t) = 1/(1-p) - e^{-\beta t}/(\beta\mu^{\#}) + o(e^{-\beta t}),$$
$$P\{T_\nu > t\} = (1-p)e^{-\beta t}/(\beta\mu^{\#}) + o(e^{-\beta t}), \quad \text{as } t \to \infty.$$

Proof. The first line follows by Theorem 70 with $h(t) = 1$, since by its definition, $V(t) = 1 + G \star V(t)$. The second follows from the first line and (2.51).

Example 72. Waiting Time for a Gap in a Poisson Process. Consider a Poisson process with rate λ that terminates at the first time a gap of size $\ge c$ occurs. That is, the termination time is T_ν, where $\nu = \min\{n : \xi_{n+1} \ge c\}$, where $\xi_n = T_n - T_{n-1}$ and T_n are the occurrence times of the Poisson process. Now, at each time T_n, the process either terminates if $\xi_{n+1} \ge c$, or it continues until the next renewal epoch if $\xi_{n+1} < c$. These events are clearly independent of $T_1, \ldots, T_n$.

Under these assumptions, the probability of terminating is

$$1 - p = P\{\xi_{n+1} \geq c\} = e^{-\lambda c}.$$

The conditional distribution of the next renewal period beginning at T_n is

$$F(t) = P\{\xi_{n+1} \leq t | \xi_{n+1} < c\} = p^{-1}(1 - e^{-\lambda t}), \qquad 0 \leq t \leq c.$$

Then from (2.51), the distribution and mean of the waiting time for a gap of size c are

$$P\{T_\nu \leq t\} = e^{-\lambda c} V(t), \qquad E[T_\nu] = (e^{\lambda c} - 1)/\lambda.$$

Now, assume $\lambda c > 1$. Then the condition $\int_{\mathbb{R}_+} e^{\beta t} p dF(t) = 1$ above for defining β reduces to $\lambda e^{(\beta - \lambda)c} = \beta$, for $\beta < \lambda$. Such a β exists as in Figure 1.3 in Chapter 1 for the branching model. Using this formula and integration by parts, we have

$$\mu^{\#} = \int_{[0,c]} t e^{\beta t} p dF(t) = p(c\beta - 1)/(\beta - \lambda).$$

Then by Corollary 71,

$$P\{T_\nu > t\} = \Big(\frac{1 - \beta/\lambda}{1 - \beta c}\Big) e^{-\beta t} + o(e^{-\beta t}), \quad \text{as } t \to \infty.$$

Example 73. Cramér-Lundberg Risk Model. Consider an insurance company that receives capital at a constant rate c from insurance premiums, investments, interest etc. The company uses the capital to pay claims that arrive according to a Poisson process $N(t)$ with rate λ. The claim amounts $X_1, X_2, \ldots$ are i.i.d. positive random variables with mean μ, and are independent of the arrival times. Then the company's capital at time t is

$$Z_x(t) = x + ct - \sum_{n=1}^{N(t)} X_n, \qquad t \geq 0,$$

where x is the capital at time 0.

An important performance parameter of the company is the probability

$$R(x) = P\{Z_x(t) \geq 0, t \geq 0\},$$

that the capital does not go negative (the company is not ruined). We are interested in approximating this survival probability when the initial capital x is large. Exercise 25 shows that $R(x) = 0$, regardless of the initial capital x, when $c < \lambda\mu$ (the capital input rate is less than the payout rate).

We will now consider the opposite case $c > \lambda\mu$. Conditioning on the time and size of the first claim, one can show (e.g., see [37, 92, 94]) that $R(x)$ satisfies a certain differential equation whose corresponding integral equation

is the renewal equation

$$R(x) = R(0) + R \star G(x), \tag{2.56}$$

where $R(0) = 1 - \lambda\mu/c$ and

$$G(y) = \lambda c^{-1} \int_0^y P\{X_1 > u\}\, du.$$

The G is a defective distribution with $G(\infty) = \lambda\mu/c < 1$. Then applying (2.52) to $R(x) = h \star V(x) = R(0)V(x)$, we have

$$R(x) \to R(0)/(1 - \lambda\mu/c) = 1, \qquad \text{as } x \to \infty.$$

We now consider the rate at which the "ruin" probability $1 - R(x)$ converges to 0 as $x \to \infty$. Assume there is a positive β such that

$$\lambda c^{-1} \int_{\mathbb{R}_+} e^{\beta x} P\{X_1 > x\}\, dx = 1,$$

and that

$$\mu^{\#} = \lambda c^{-1} \int_{\mathbb{R}_+} x e^{\beta x} P\{X_1 > x\}\, dx < \infty.$$

Then by Theorem 70 (with $R(x)$, $R(0)$ in place of $H(t)$, $h(t)$), the probability of ruin has the asymptotic form

$$1 - R(x) = \frac{1}{\beta\mu^{\#}}(1 - \lambda\mu/c)e^{-\beta x} + o(e^{-\beta x}), \quad \text{as } x \to \infty.$$

2.15 Stationary Renewal Processes

Recall that a basic property of an ergodic Markov chain is that it is stationary if the distribution of its state at time 0 is its stationary distribution (which is also its limiting distribution). This section addresses the analogous issue of determining an appropriate starting condition for a delayed renewal process so that its increments are stationary in time.

We begin by defining the notion of stationarity for stochastic processes and point processes. A continuous-time stochastic process $\{X(t) : t \geq 0\}$ on a general space is *stationary* if its finite-dimensional distributions are invariant under any shift in time: for each $0 \leq s_1 < \ldots < s_k$ and $t \geq 0$,

$$(X(s_1 + t), \ldots, X(s_k + t)) \stackrel{d}{=} (X(s_1), \ldots, X(s_k)). \tag{2.57}$$

Remark 74. A Markov process $X(t)$ is stationary if $X(t) \stackrel{d}{=} X(0)$, $t \geq 0$. This simpler criterion follows as in the proofs of Proposition 52 in Chapter 1 and Exercise 55.

Now, consider a point process $N(t) = \sum_n \mathbf{1}(\tau_n \leq t)$ on $\mathbb{R}_+$, with points at $0 < \tau_1 < \tau_2 < \ldots$ Another way of representing this process is by the family $N = \{N(B) : B \in \mathbb{B}_+\}$, where $N(B) = \sum_n \mathbf{1}(\tau_n \in B)$ is the number of points τ_n in the Borel set B. We also define $B + t = \{s + t : s \in B\}$. The process N is *stationary* (i.e., it has *stationary increments*) if, for any $B_1, \ldots, B_k \in \mathbb{B}_+$,

$$(N(B_1 + t), \ldots, N(B_k + t)) \stackrel{d}{=} (N(B_1), \ldots, N(B_k)), \qquad t \geq 0. \tag{2.58}$$

A basic property of a stationary point process is that its mean value function is linear.

Proposition 75. If N is a stationary point process and $E[N(1)]$ is finite, then $E[N(t)] = tE[N(1)]$, $t \geq 0$.

Proof. To see this, consider

$$E[N(s + t)] = E[N(s)] + E[N(s + t) - N(s)] = E[N(s)] + E[N(t)].$$

This is a linear equation $f(s + t) = f(s) + f(t)$, $s, t \geq 0$. The only nondecreasing function that satisfies this linear equation is $f(t) = ct$ for some c. In our case, $c = f(1) = E[N(1)]$, and hence $E[N(t)] = tE[N(1)]$.

We are now ready to characterize stationary renewal processes. Assume that $N(t)$ is a delayed renewal process, where the distribution of ξ_1 is G, and the distribution of ξ_n, $n \geq 2$, is F, which has a finite mean μ. The issue is how to select the initial distribution G such that N is stationary. The answer, according to (iv) below, is to select G to be F_e, which is the limiting distribution of the forward and backward recurrence times for a renewal process with inter-renewal distribution F. The following result also shows that the stationarity of N is equivalent to the stationarity of its forward recurrence time process.

Theorem 76. *The following statements are equivalent.*
(i) *The delayed renewal process N is stationary.*
(ii) *The forward recurrence time process $B(t) = T_{N(t)+1} - t$ is stationary.*
(iii) $E[N(t)] = t/\mu$, *for* $t \geq 0$.
(iv) $G(t) = F_e(t) = \frac{1}{\mu} \int_0^t [1 - F(s)]ds$.
When these statements are true, $P\{B(t) \leq x\} = F_e(x)$, *for* $t, x \geq 0$.

Proof. (i) $\Leftrightarrow$ (ii): Using $T_n = \inf\{u : N(u) = n\}$, we have

$$\begin{aligned} B(t) = T_{N(t)+1} - t &= \inf\{u - t : N(u) = N(t) + 1\} \\ &\stackrel{d}{=} \inf\{t' : N((0, t'] + t) = 1\}. \end{aligned} \tag{2.59}$$

Consequently, the stationarity property (2.58) of N implies $B(t) \stackrel{d}{=} B(0)$, $t \geq 0$. Then B is stationary by Remark 74, because it is a Markov process (Exercise 55).

Conversely, since N counts the number of times $B(t)$ jumps upward,

$$N(A+t) = \sum_{u \in A} \mathbf{1}(B(u+t) > B((u+t)-)). \tag{2.60}$$

Therefore, the stationarity of B implies N is stationary.

(i) $\Rightarrow$ (iii): If N is stationary, Proposition 75 ensures $E[N(t)] = tE[N(1)]$. Also, $E[N(1)] = 1/\mu$ since $t^{-1}E[N(t)] \to 1/\mu$ by Proposition 32. Therefore, $E[N(t)] = t/\mu$.

(iii) $\Rightarrow$ (iv): Assume $E[N(t)] = t/\mu$. Exercise 53 shows $U \star F_e(t) = t/\mu$, and so $E[N(t)] = U \star F_e(t)$. Another expression for this expectation is

$$E[N(t)] = \sum_{n=1}^{\infty} G \star F^{(n-1)\star}(t) = G \star U(t).$$

Equating these expressions, we have $U \star F_e(t) = G \star U(t)$. Taking the Laplace transform of this equality yields

$$\hat{U}(\alpha)\hat{F}_e(\alpha) = \hat{G}(\alpha)\hat{U}(\alpha), \tag{2.61}$$

where the hat symbol denotes Laplace transform; e.g., $\hat{G}(\alpha) = \int_{\mathbb{R}_+} e^{-\alpha t} dG(t)$. By Proposition 20, we know $\hat{U}(\alpha) = 1/(1 - \hat{F}(\alpha))$ is positive. Using this in (2.61) yields $\hat{F}_e(\alpha) = \hat{G}(\alpha)$. Since these Laplace transforms uniquely determine the distributions, we obtain $G = F_e$.

(iv) $\Rightarrow$ (ii): By direct computation as in Exercise 37, it follows that

$$P\{B(t) > x\} = 1 - G(t+x) + \int_{[0,t]} [1 - F(t+x-s)] dV(s), \tag{2.62}$$

where $V(t) = E[N(t)] = G \star U(t)$. Now, the assumption $G = F_e$, along with $U \star F_e(t) = t/\mu$ from Exercise 53, yield

$$V(t) = G \star U(t) = U \star G(t) = U \star F_e(t) = t/\mu.$$

Using this in (2.62), along with a change of variable in the integral, we have

$$P\{B(t) > x\} = 1 - G(t+x) + F_e(x+t) - F_e(x). \tag{2.63}$$

Since $G = F_e$, this expression is simply $P\{B(t) > x\} = 1 - F_e(x)$, $t \geq 0$. Thus, the distribution of $B(t)$ is independent of t. This condition is sufficient for $B(t)$ to be stationary since it is a Markov process (see Exercise 55).

Example 77. Suppose the inter-renewal distribution for the delayed renewal process N is the beta distribution

$$F(t) = 30 \int_0^t s^2(1-s)^2\, ds, \qquad t \in [0,1].$$

The equilibrium distribution associated with F is clearly

$$F_e(t) = 2t - 5t^4 + 6t^5 - 2t^6, \qquad t \in [0,1].$$

Then by Theorem 76, N is stationary if and only if $G = F_e$.

One consequence of Theorem 76 is that Poisson processes are the only non-delayed renewal processes (whose inter-renewal times have a finite mean) that are stationary.

Corollary 78. *The renewal process $N(t)$ with no delay, and whose inter-renewal times have a finite mean, is stationary if and only if it is a Poisson process.*

Proof. By Theorem 76 (vi), $N(t)$ is stationary if and only if $E[N(t)] = t/\mu$, $t \geq 0$, which is equivalent to $N(t)$ being a Poisson process by Remark 21.

An alternate proof is to apply Theorem 76 (iii) and use the fact (Exercise 4 in Chapter 3) that $F = F_e$ if and only if F is an exponential distribution.

Here is another useful stationarity property.

Remark 79. If $N(t)$ is a stationary renewal process, then

$$E\Big[\sum_{n=1}^{N(t)} f(T_n)\Big] = \frac{1}{\mu}\int_0^t f(s)ds.$$

This follows by Theorem 22 and $E[N(t)] = t/\mu$.

Many stationary processes arise naturally as functions of stationary processes (two examples are in the proof of Theorem 76). A general statement to this effect is as follows; it is a consequence of the definition of stationarity.

Remark 80. Hereditary Property of Stationarity. Suppose $X(t)$ is a stationary process. Then the process $Y(t) = f(X(t))$ is also stationary, where f is a function on the state space of X to another space. More generally, $Y(t) = g(\{X(s+t) : s \geq 0\})$ is stationary, where g is a function on the space of sample paths of X to some space. Analogously, N is a stationary point process if, for any bounded set B and $t > 0$,

$$N(B+t) = g(\{X(s+t) : s \geq 0\}, B) \tag{2.64}$$

(see for instance (2.59) and (2.60)).

Example 81. Let $X(t)$ be a delayed regenerative process (e.g., a continuous-time Markov chain as in Chapter 4) over the times $0 < T_1 < T_2 < \ldots$ at which $X(t)$ enters a special state x^*. Let N denote the point process of these

times. If $X(t)$ is stationary, then N is a stationary renewal process. This follows since, like (2.64),

$$N(B+t) = \sum_{s \in B} \mathbf{1}(X((s+t)-) \neq x^*, X(s+t) = x^*).$$

Although the bounded set B may be uncountable, only a finite number of its values will contribute to the sum.

Because a stationary renewal process $N(t)$ has a stationary forward recurrence time process, it seems reasonable that the backward recurrence time process $A(t) = t - T_{N(t)}$ would also be stationary. This is not true, since the distribution of $A(t)$ is not independent of t; in particular, $A(t) = t$, for $t < T_1$. However, there is stationarity in the following sense.

Remark 82. Stationary Backward Recurrence Time Process. Suppose the stationary renewal process is extended to the negative time axis with (artificial or virtual) renewals at times $\ldots < T_{-1} < T_0 < 0$. One can think of the renewals occurring since the beginning of time at $-\infty$. Consistent with the definition above, the backward recurrence process is

$$A(t) = t - T_n, \qquad \text{if } t \in [T_n, T_{n+1}), \text{ for some } n \in \mathbb{R}.$$

Assuming N is stationary on $\mathbb{R}_+$, the time $A(0) = T_1$ to the first renewal has the distribution F_e. Then one can show, as we proved (i) $\Leftrightarrow$ (ii) in Theorem 76, that the process $\{A(t) : t \in \mathbb{R}\}$ is stationary with distribution F_e.

2.16 Refined Limit Laws

We will now describe applications of the key renewal theorem for functions that are not asymptotically constant.

The applications of the renewal theorem we have been discussing are for limits of functions $H(t) = U \star h(t)$ that converge to a constant (i.e., $H(t) = c + o(1)$). However, there are many situations in which $H(t)$ tends to infinity, but the key renewal theorem can still be used to describe limits of the form $H(t) = v(t) + o(1)$ as $t \to \infty$, where the function $v(t)$ is the asymptotic value of $H(t)$.

For instance, a SLLN $Z(t)/t \to b$ suggests $E[Z(t)] = bt + c + o(1)$ might be true, where the constant c gives added information on the convergence. In this section, we discuss such limit theorems.

We first note that an approach for considering limits $H(t) = v(t) + o(1)$ is simply to consider a renewal equation for the function $H(t) - v(t)$ as follows.

Lemma 83. *Suppose $H(t) = U \star h(t)$ is a solution of a renewal equation for a non-arithmetic distribution F, and $v(t)$ is a real-valued function on $\mathbb{R}$ that*

is bounded on finite intervals and is 0 *for negative* t*. Then*

$$H(t) = v(t) + \frac{1}{\mu}\int_{\mathbb{R}_+} \overline{h}(s)\,ds + o(1), \quad \text{as } t \to \infty, \tag{2.65}$$

provided $\overline{h}(t) = h(t) - v(t) + F \star v(t)$ *is DRI. In particular, for a linear function* $v(t) = bt$*,*

$$H(t) = bt + \frac{b(\sigma^2+\mu^2)}{2\mu} + \frac{1}{\mu}\int_{\mathbb{R}_+} (h(s) - b\mu)\,ds + o(1), \quad \text{as } t \to \infty, \tag{2.66}$$

where σ^2 *is the variance of* F*, provided* $h(t) - b\mu$ *is DRI.*

Proof. Clearly $H - v$ satisfies the renewal equation

$$H - v = (h - v + F \star v) + F \star (H - v).$$

Then $H - v = U \star \overline{h}$ by Proposition 31, and its limit (2.65) is given by the key renewal theorem.

Next, suppose $v(t) = bt$ and $h(t) - b\mu$ is DRI. Then using $\mu = \int_{\mathbb{R}_+}[1 - F(x)]\,dx$ and the change of variable $x = t - s$ in the integral below, we have

$$\begin{aligned}\overline{h}(t) &= h(t) - bt + b\int_0^t F(t-s)\,ds \\ &= h(t) - b\mu + bg(t),\end{aligned}$$

where $g(t) = \int_t^\infty [1 - F(x)]\,dx$. Now $g(t)$ is continuous and decreasing and

$$\int_0^\infty g(t)\,dt = \frac{1}{2}\int_{\mathbb{R}_+} t^2 dF(t) = \frac{\sigma^2+\mu^2}{2}. \tag{2.67}$$

Then $g(t)$ is DRI by Proposition 88 (a), and hence $\overline{h}(t) = h(t) - b\mu + bg(t)$ is DRI. Thus, by what we already proved, (2.65) is true but it reduces to (2.66) in light of (2.67).

Our first use of the preceding result is a refinement of $t^{-1}U(t) \to 1/\mu$ from Proposition 32.

Proposition 84. *If* $N(t)$ *is a renewal process whose inter-renewal times have a non-arithmetic distribution with mean* μ *and variance* σ^2*, then*

$$U(t) = t/\mu + (\sigma^2+\mu^2)/2\mu^2 + o(1), \qquad \text{as } t \to \infty.$$

Proof. This follows by Lemma 83 with $H(t) = U(t)$, $h(t) = 1$, and $v(t) = t/\mu$ (that $h(t) - b\mu$ is DRI need not be verified since it equals 0).

We will now apply Lemma 83 to a real-valued stochastic process $Z(t)$ whose sample paths are right-continuous with left-hand limits. Assume

that $Z(t)$, has *crude regenerative increments at* T in the sense that

$$E[Z(T+t) - Z(T)|T] = E[Z(t)], \qquad t \geq 0. \tag{2.68}$$

If $Z(t)$ has regenerative increments over T_n, then $Z(t)$ has crude regenerative increments at T_1.

Theorem 85. *For the process $Z(t)$ defined above, let*

$$M = \sup\{|Z(T) - Z(t)| : t \leq T\}.$$

If the expectations of M, MT, T^2, $|Z(T)|$, and $\int_0^T |Z(s)|ds$ are finite, then

$$E[Z(t)] = at/\mu + a(\sigma^2 + \mu^2)/2\mu^2 + c + o(1), \qquad \text{as } t \to \infty, \tag{2.69}$$

where $a = E[Z(T)]$ and $c = \frac{1}{\mu}E\Big[\int_0^T Z(s)ds - TZ(T)\Big]$.

Proof. Because $Z(t)$ has crude regenerative increments, it would be natural that that $t^{-1}E[Z(t)] \to a/\mu$. So to prove (2.69), we will apply Lemma 83 with $v(t) = at/\mu$.

We first derive a renewal equation for $E[Z(t)]$. Conditioning on T,

$$E[Z(t)] = E[Z(t)\mathbf{1}(T > t)] + \int_{[0,t]} E[Z(t)|T = s]dF(s).$$

Using $E[Z(t)|T = s] = E[Z(t-s)] + E[Z(s)|T = s]$ from assumption (2.68) and some algebra, it follows that the preceding is a renewal equation $H = h + F \star H$, where $H(t) = E[Z(t)]$ and

$$h(t) = a + E\Big[(Z(t) - Z(T))\mathbf{1}(T > t)\Big].$$

Now, by Lemma 83 for $v(t) = at/\mu$, we have

$$E[Z(t)] = at/\mu + \frac{\sigma^2 + \mu^2}{2\mu^2} + \frac{1}{\mu}\int_{\mathbb{R}_+} g(s)\, ds + o(1), \quad \text{as } t \to \infty, \tag{2.70}$$

provided $g(t) = h(t) - a = E\Big[(Z(t) - Z(T))\mathbf{1}(T > t)\Big]$ is DRI. Clearly

$$|g(t)| \leq b(t) = E[M\mathbf{1}(T > t)].$$

Now, $b(t) \downarrow 0$; and as in (2.25), $\int_{\mathbb{R}_+} b(s)ds = E[MT]$ is finite. Then $b(t)$ is DRI by Proposition 88 (a). Hence $g(t)$ is also DRI by Proposition 88 (c). Finally, observe that

$$\int_{\mathbb{R}_+} g(t)\, dt = E\Big[\int_0^T Z(s)ds - TZ(T)]\Big].$$

Substituting this formula in (2.70) proves (2.69).

2.17 Proof of the Key Renewal Theorem*

This section proves the key renewal theorem by applying Blackwell's theorem, which is proved in the next section.

The key renewal theorem involves real-valued functions that are integrable on the entire axis $\mathbb{R}_+$ as follows.

Definition 86. Similarly to the definition of a Riemann integral on a finite interval, it is natural to approximate the integral of a real-valued function $h(t)$ on the entire domain $\mathbb{R}_+$ over a grid $0, \delta, 2\delta, \ldots$ by the upper and lower Riemann sums

$$I^\delta(h) = \delta \sum_{k=0}^{\infty} \sup\{h(s) : k\delta \le s < (k+1)\delta\},$$
$$I_\delta(h) = \delta \sum_{k=0}^{\infty} \inf\{h(s) : k\delta \le s < (k+1)\delta\}.$$

The function $h(t)$ is *directly Riemann integrable* (DRI) if $I^\delta(h)$ and $I_\delta(h)$ are finite for each δ, and they both converge to the same limit as $\delta \to 0$. The limit is necessarily the usual Riemann integral

$$\int_{\mathbb{R}_+} h(s)ds = \lim_{t\to\infty} \int_0^t h(s)ds,$$

where the last integral is the limit of the Riemann sums on $[0, t]$.

A DRI function is clearly Riemann integrable in the usual sense, but the converse is not true; see Exercise 28. From the definition, it is clear that $h(t)$ is DRI if it is Riemann integrable and it is 0 outside a finite interval. Also, $h(t)$ is DRI if and only if its positive and negative parts $h^+(t)$ and $h^-(t)$ are both DRI. Further criteria for DRI are given in Proposition 88 and Exercise 33.

We are now ready for the main result.

Theorem 87. (Key Renewal Theorem) *If $h(t)$ is DRI and F is non-arithmetic, then*

$$\lim_{t\to\infty} U \star h(t) = \frac{1}{\mu} \int_{\mathbb{R}_+} h(s)ds.$$

Proof. Fix $\delta > 0$ and define $\overline{h}_k = \sup\{h(s) : k\delta \le s < (k+1)\delta\}$ and

$$\overline{h}(t) = \sum_{k=0}^{\infty} \overline{h}_k \mathbf{1}(k\delta \le t < (k+1)\delta).$$

* The star at the end of a section title means the section contains advanced material that need not be covered in a first reading.

Define $\underline{h}(t)$ and $\underline{h}_k$ similarly, with sup replaced by inf. Obviously,

$$U \star \underline{h}(t) \le U \star h(t) \le U \star \overline{h}(t). \tag{2.71}$$

Letting $d_k(t) = U(t - k\delta) - U(t - (k+1)\delta)$, we can write (like (2.19))

$$U \star \overline{h}(t) = \sum_{k=0}^{\infty} \overline{h}_k d_k(t).$$

Now $\lim_{t\to\infty} d_k(t) = \delta/\mu$ by Theorem 33, and $d_k(t) \le U(\delta)$ by Exercise 28. Then by the dominated convergence theorem (see the Appendix, Theorem 14) and the DRI property of h,

$$\begin{aligned}\lim_{\delta\to 0} \lim_{t\to\infty} U \star \overline{h}(t) &= \lim_{\delta\to 0} \frac{\delta}{\mu} \sum_{k=0}^{\infty} \overline{h}_k \\ &= \lim_{\delta\to 0} \frac{1}{\mu} I^{\delta}(h) = \frac{1}{\mu} \int_{\mathbb{R}_+} h(s)ds.\end{aligned}$$

This (double) limit is the same with $\overline{h}(t)$ and $I^{\delta}(h)$ replaced by $\underline{h}(t)$ and $I_{\delta}(h)$. Therefore, the upper and lower bounds in (2.71) for $U \star h(t)$ have the same limit $\frac{1}{\mu}\int_{\mathbb{R}_+} h(s)ds$, and so $U \star h(t)$ must also have this limit. This proves the assertion.

We end this section with criteria for a function to be DRI.

Proposition 88. *Any one of the following conditions is sufficient for $h(t)$ to be DRI.*
(a) *$h(t) \ge 0$ is decreasing and is Riemann integrable on $\mathbb{R}_+$.*
(b) *$h(t)$ is Riemann integrable on $[0, a]$ for each a, and $I^{\delta}(h) < \infty$ for some $\delta > 0$.*
(c) *$h(t)$ is continuous except possibly on a set of Lebesgue measure 0, and $|h(t)| \le b(t)$, where $b(t)$ is DRI.*

Proof. Suppose condition (a) holds. Since the usual Riemann integral of h on $\mathbb{R}_+$ exists, we have

$$I_{\delta}(h) \le \int_{\mathbb{R}_+} h(s)ds \le I^{\delta}(h).$$

Also, the decreasing property of $h(t)$ implies $I^{\delta}(h) - I_{\delta}(h) = \delta h(0) \to 0$ as $\delta \to 0$. These observations prove $h(t)$ is DRI.

Next, suppose (b) holds. We will write

$$I^{\delta}(h) = \mathcal{I}^{\delta}[0, a/\delta) + \mathcal{I}^{\delta}[a/\delta, \infty),$$

where $\mathcal{I}^{\delta}[x, \infty) = \delta \sum_{k=\lceil x-1 \rceil}^{\infty} \sup\{h(s) : k\delta \le s < (k+1)\delta\}$. We will use a similar expression for $I_{\delta}(h)$. Since $h(t)$ is Riemann integrable on $[0, a]$, it

follows that $\mathcal{I}^\delta[0, a/\delta)$ and $\mathcal{I}_\delta[0, a/\delta)$ both converge to $\int_0^a h(s)ds$ as $\delta \to 0$. Therefore,

$$I^\delta(h) - I_\delta(h) = o(1) + \mathcal{I}^\delta[a/\delta, \infty) - \mathcal{I}_\delta[a/\delta, \infty), \qquad \text{as } \delta \to 0. \tag{2.72}$$

Let γ be such that $I^\gamma(h) < \infty$. Then for any $\varepsilon > 0$, there is a large enough a such that $\mathcal{I}^\gamma[a/\gamma, \infty) < \varepsilon$. Then clearly, for sufficiently small δ,

$$\mathcal{I}_\delta[a/\delta, \infty) \le \mathcal{I}^\delta[a/\delta, \infty) \le \mathcal{I}^\gamma[a/\gamma, \infty) < \varepsilon.$$

Using this in (2.72), we have

$$I^\delta(h) - I_\delta(h) \le o(1) + 2\varepsilon, \qquad \text{as } \delta \to 0.$$

Since this holds for any ε, it follows that $h(t)$ is DRI.

Finally, (c) implies (b) since $I^\delta(h) \le I^\delta(b)$. Thus $h(t)$ is DRI.

2.18 Proof of Blackwell's Theorem*

This section describes a coupling proof of Blackwell's theorem. The proof is more complicated than the one we presented above for arithmetic inter-renewal times.

The classical proof of Blackwell's theorem based on analytical properties of the renewal function and integral equations is in Feller (1971). Lindvall (1977) and Athreya, McDonald and Ney (1978) gave another probabilistic proof involving "coupling" techniques. A nice review of various applications of coupling is in Lindvall (1992). A recent refinement of the coupling proof is given in Durrett (2005). The following is a sketch of his presentation when the inter-renewal time has a finite mean (he gives a different proof for the case of an infinite mean).

Let $N(t)$ be a renewal process with renewal times T_n whose inter-renewal times ξ_n have a non-arithmetic distribution and a finite mean μ. For convenience, we will write Blackwell's theorem (Theorem 33) as

$$\lim_{t\to\infty} E[N(t, t+a]] = a/\mu, \tag{2.73}$$

where $N(t, t+a] = N(t+a) - N(t)$. Now, this statement would trivially hold if $N(t)$ were a stationary renewal process, since in this case $E[N(t, t+a]]$ would equal a/μ by Proposition 75. So if one could construct a version of $N(t)$ that approximates a stationary process as close as possible, then (2.73) would be true. That is the approach in the proof that we now describe.

On the same probability space as $N(t)$, let $N'(t)$ be a stationary renewal process with renewal times T'_n, whose inter-renewal ξ'_n times for $n \ge 2$ have the same distribution as the ξ_n. The first and most subtle part of the proof

is to construct a third renewal process $N''(t)$ on the same probability space that is equal in distribution to the original process $N(t)$ and approximates the stationary process $N'(t)$. We will not describe the construction of these processes, but only specify their main properties.

In particular, for a fixed $\varepsilon > 0$, the proof begins by defining random indices ν and ν' such that $|T_\nu - T_{\nu'}| < \varepsilon$. Then a third renewal process $N''(t)$ is defined (on the same probability space) with inter-renewal times $\xi_1, \ldots, \xi_\nu, \xi'_{\nu'}, \xi'_{\nu'+1} \ldots$ This process has the following properties:

(a) $\{N''(t) : t \geq 0\} \stackrel{d}{=} \{N(t) : t \geq 0\}$ (i.e., their finite-dimensional distributions are equal).

(b) On the event $\{T_\nu \leq t\}$,

$$N'(t+\varepsilon, t+a-\varepsilon] \leq N''(t, t+a] \leq N'(t-\varepsilon, t+a+\varepsilon]. \tag{2.74}$$

This construction is an *ε-coupling* in that $N''(t)$ is a coupling of $N(t)$ that is within ε of the targeted stationary version $N'(t)$ in the sense of condition (b).

With this third renewal process in hand, the rest of the proof is as follows. Consider the expectation

$$E[N(t, t+a]] = E[N''(t, t+a]] = V_{\mathbf{1}}(t) + V_2(t), \tag{2.75}$$

where

$$V_{\mathbf{1}}(t) = E\Big[N''(t, t+a]\mathbf{1}(T_\nu \leq t)\Big], \qquad V_2(t) = E\Big[N''(t, t+a]\mathbf{1}(T_\nu > t)\Big].$$

Condition (b) and $E[N'(c, d]] = (d-c)/\mu$ (due to the stationarity) ensure

$$V_{\mathbf{1}}(t) \leq E\Big[N'(t-\varepsilon, t+a+\varepsilon]\mathbf{1}(T_\nu \leq t)\Big] \leq (a+2\varepsilon)\mu.$$

Next, observe that $E[N''(t, t+a]|T_\nu > t] \leq E[N''(a)]$, since the worse-case scenario is that there is a renewal at t. This and condition (b) yield

$$V_2(t) \leq P\{T_\nu > t\}E[N''(a)].$$

Similarly,

$$\begin{aligned} V_{\mathbf{1}}(t) &\geq E\Big[N'(t+\varepsilon, t+a-\varepsilon] - N''(t, t+a]\mathbf{1}(T_\nu > t)\Big] \\ &\geq (a-2\varepsilon)/\mu - P\{T_\nu > t\}E[N''(a)]. \end{aligned}$$

Here we take $\varepsilon < a/2$, so that $t+\varepsilon < t+a-\varepsilon$. Combining the preceding inequalities with (2.75), and using $P\{T_\nu > t\} \to 0$ as $t \to \infty$, it follows that

$$(a-2\varepsilon)/\mu + o(1) \leq E[N(t, t+a]] \leq (a+2\varepsilon)/\mu + o(1).$$

Since this is true for arbitrarily small ε, we obtain $E[N(t, t+a]] \to a/\mu$, which is Blackwell's result.

2.19 Stationary-Cycle Processes*

Most of the results above for regenerative processes also apply to a wider class of regenerative-like processes that we will now describe.

For this discussion, suppose $\{X(t) : t \geq 0\}$ is a continuous-time stochastic process with a general state space S, and $N(t)$ is a renewal process defined on the same probability space. As in Section 2.8, we let

$$\zeta_n = (\xi_n, \{X(T_{n-1} + t) : 0 \leq t < \xi_n\})$$

denote the segment of these processes on the interval $[T_{n-1}, T_n)$. Then $\{\zeta_{n+k} : k \geq 1\}$ is the *future of* $(N(t), X(t))$ *beginning at time* T_n. This is what an observer of the processes would see beginning at time T_n.

Definition 89. The process $X(t)$ is a *stationary-cycle process* over the times T_n if the future $\{\zeta_{n+k} : k \geq 1\}$ of $(N(t), X(t))$ beginning at any time T_n is independent of $T_1, \ldots, T_n$, and the distribution of this future is independent of n. Discrete-time and delayed stationary-cycle processes are defined similarly.

The defining property ensures that the segments ζ_n form a stationary sequence, whereas for a regenerative process, the segments are i.i.d. Also, for a regenerative process $X(t)$, its future $\{\zeta_{n+k} : k \geq 1\}$ beginning at any time T_n is independent of the entire past $\{\zeta_k : k \leq n\}$ (rather than only $T_1, \ldots, T_n$ as in the preceding definition).

All the strong laws of large numbers for regenerative processes in this chapter also hold for stationary-cycle processes. A law's limiting value would be a constant as usual when ζ_n is ergodic (as in Section 4.18 in Chapter 4), but the value would be random when ζ_n is not ergodic. We will not get into these details.

As in Section 2.10, one can define processes with stationary-cycle increments. Most of the results above such as the CLT have obvious extensions to these more complicated processes.

We end this section by commenting on limiting theorems for probabilities and expectations of stationary-cycle processes.

Remark 90. Theorem 45 and Corollary 46 are also true for stationary-cycle processes. This follows since such a process satisfies the crude-regeneration property in Theorem 41 leading to Theorem 45 and Corollary 46.

There are many intricate stationary-cycle processes that arise naturally from systems that involve stationary and regenerative phenomena. Here is an elementary illustration.

Example 91. Regenerations in a Stationary Environment. Consider a process $X(t) = g(Y(t), Z(t))$ where $Y(t)$ and $Z(t)$ are independent processes and g is a function on their product space. Assume $Y(t)$ is a regenerative process over the times T_n (e.g., an ergodic continuous-time Markov chain as in Chapter 4) with a metric state space S. Assume $Z(t)$ is a stationary process. One can regard $X(t) = g(Y(t), Z(t))$ as a regenerative-stationary reward process, where $g(y, z)$ is the reward rate from operating a system in state y in environment z. Now, the segments ζ_n defined above form a stationary process, and hence $X(t)$ is a stationary-cycle process.

In light of Remark 90, we can describe the limiting behavior of $X(t)$ as we did for regenerative processes. In particular, assuming for simplicity that g is real-valued and bounded, Theorem 45 for stationary-cycle processes tells us that

$$\lim_{t\to\infty} E[X(t)] = \frac{1}{\mu} E\Big[\int_0^{T_1} g(Y(s), Z(s))ds\Big].$$

2.20 Exercises

Exercise 1. Show that if X is nonnegative with distribution F, then

$$E[X] = \int_{\mathbb{R}_+} (1 - F(x))dx.$$

One approach is to use $E[X] = \int_{\mathbb{R}_+} \big(\int_0^x dy\big)dF(x)$. (For an integer-valued X, the preceding formula is $E[X] = \sum_{n=0}^{\infty} P\{X > n\}$.)

For a general X with finite mean, use $X = X^+ - X^-$ to prove

$$E[X] = \int_{\mathbb{R}_+} (1 - F(x))dx - \int_{-\infty}^{0} F(x)dx.$$

Exercise 2. *Bernoulli Process.* Consider a sequence of independent Bernoulli trials in which each trial results in a success or failure with respective probabilities p and $q = 1 - p$. Let $N(t)$ denote the number of successes in t trials, where t is an integer. Show that $N(t)$ is a discrete-time renewal process, called a Bernoulli Process. (The parameter t may denote discrete-time or any integer referring to sequential information.) Justify that the inter-renewal times have the geometric distribution $P\{\xi_1 = n\} = pq^{n-1}$, $n \geq 1$. Find the distribution and mean of $N(t)$, and do the same for the renewal time T_n. Show that the moment generating function of T_n is

$$E[e^{\alpha T_n}] = \Big(\frac{pe^{\alpha}}{1 - qe^{\alpha}}\Big)^n, \qquad 0 < \alpha < -\log q.$$

Exercise 3. Exercise 1 in Chapter 3 shows that an exponential random variable X satisfies the *memoryless property*

$$P\{X > s+t|X > s\} = P\{X > t\}, \qquad s,t > 0.$$

Prove the analogue $P\{X > \tau+t|X > \tau\} = P\{X > t\}$, for $t > 0$, where τ is a positive random variable independent of X. Show that, for a Poisson process $N(t)$ with rate λ, the forward recurrence time $B(t) = T_{N(t)+1} - t$ at time t has an exponential distribution with rate λ. Hint: condition on $T_{N(t)}$.

Consider the forward recurrence time $B(\tau)$ at a random time τ independent of the Poisson process. Show that $B(\tau)$ also has an exponential distribution with rate λ.

Exercise 4. A system consists of two components with independent lifetimes X_1 and X_2, where X_1 is exponentially distributed with rate λ, and X_2 has a uniform distribution on $[0,1]$. The components operate in parallel, and the system lifetime is $\max\{X_1, X_2\}$ (the system is operational if and only if at least one component is working). When the system fails, it is replaced by another system with an identical and independent lifetime, and this is repeated indefinitely. The number of system renewals over time forms a renewal process $N(t)$. Find the distribution and mean of the system lifetime. Find the distribution and mean of $N(t)$ (reduce your formulas as much as possible). Determine the portion of time that (a) two components are working, (b) only type 1 component is working, and (c) only type 2 component is working.

Exercise 5. *Continuation.* In the context of the preceding exercise, a typical system initially operates for a time $Y = \min\{X_1, X_2\}$ with two components and then operates for a time $Z = \max\{X_1, X_2\} - Y$ with one component. Thereupon it fails. Find the distributions and means of Y and Z. Find the distribution of Z conditioned that $X_1 > X_2$. You might want to use the memoryless property of the exponential distribution in Exercise 3. Find the distribution of Z conditioned that $X_2 > X_1$.

Exercise 6. Let $N(t)$ denote a renewal process with inter-renewal distribution F and consider the number of renewals $N(T)$ in an interval $(0,T]$ for some random time T independent of $N(t)$. For instance, $N(T)$ might represent the number of customers that arrive at a service station during a service time T. Find general expressions for the mean and distribution of $N(T)$. Evaluate these expressions for the case in which T has an exponential distribution with rate μ and $F = G^{2\star}$, where G is an exponential distribution with rate λ.

Exercise 7. Let $X(t)$ denote the cyclic renewal process in Example 8, where $F_0, \ldots, F_{K-1}$ are the sojourn distributions in states $0, 1, \ldots, K-1$. Assume $p = F_0(0) > 0$, but $F_i(0) = 0$, for $i = 1, \ldots, K-1$. Let T_n denote the times at which the process $X(t)$ jumps from state $K-1$ directly to state 1 (i.e., it

spends no time in state 0). Justify that the T_n form a delayed renewal process with inter-renewal distribution

$$F(t) = p\sum_{j=0}^{\infty} F_1 \star \cdots \star F_{K-1} \star \tilde{F}^{j\star}(t),$$

where $\tilde{F}(t) = \tilde{F}_0 \star F_1 \star \cdots \star F_{K-1}(t)$, and $\tilde{F}_0(t)$ is the conditional distribution of the sojourn time in state 0 given it is positive. Specify a formula for $\tilde{F}_0(t)$, and describe what $\tilde{F}(t)$ represents.

Exercise 8. *Large Inter-renewal Times.* Let $N(t)$ denote a renewal process with inter-renewal distribution F. Of interest are occurrences of inter-renewal times that are greater than a value c, assuming $F(c) < 1$. Let $\tilde{T}_n$ denote the subset of times T_n for which $\xi_n > c$. So $\tilde{T}_n$ equals some T_k if $\xi_k > c$. (Example 72 addresses a related problem of determining the waiting time for a gap of size c in a Poisson process.) Show that $\tilde{T}_n$ are delayed renewal times and the inter-renewal distribution has the form

$$\tilde{F}(t) = \sum_{k=0}^{\infty} F_c^{k\star} \star G(t),$$

where $F_c(t) = F(t)/F(c)$, $0 \le t \le c$ (the conditional distribution of an inter-renewal time given that it is $\le c$), and specify the distribution $G(t)$ as a function of F.

Exercise 9. *Partitioning and Thinning of a Renewal Process.* Let $N(t)$ be a renewal process with inter-renewal distribution F. Suppose each renewal time is independently assigned to be a type i renewal with probability p_i, for $i = 1, \ldots, m$, where $p_1 + \cdots + p_m = 1$. Let $N_i(t)$ denote the number of type i renewals up to time t. These processes form a partition of $N(t)$ in that $N(t) = \sum_{i=1}^{m} N_i(t)$. Each $N_i(t)$ is a thinning of $N(t)$, where p_i is the probability that a point of $N(t)$ is assigned to $N_i(t)$.

Show that $N_i(t)$ is a renewal process with inter-renewal distribution

$$F_i(t) = \sum_{k=1}^{\infty} (1-p_i)^{k-1} p_i F^{k\star}(t).$$

Show that, for $n = n_1 + \cdots n_m$,

$$\begin{aligned} &P\{N_1(t) = n_1, \ldots, N_m(t) = n_m\} \\ &\quad = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m} \Big[F^{(n)\star}(t) - F^{(n+1)\star}(t)\Big]. \end{aligned}$$

For $m = 2$, specify an F for which $N_1(t)$ and $N_2(t)$ are not independent.

Exercise 10. *Multi-type Renewals.* An infinite number of jobs are to be processed one-at-a-time by a single server. There are m types of jobs, and the

probability that any job is of type i is p_i, where $p_1 + \cdots + p_m = 1$. The service time of a type i job has a distribution F_i with mean μ_i. The service times and types of the jobs are independent. Let $N(t)$ denote the number of jobs completed by time t. Show that $N(t)$ is a renewal process and specify its inter-renewal distribution and mean. Let $N_i(t)$ denote the number of type i jobs processed up to time t. Show that $N_i(t)$ is a delayed renewal process and specify $\lim_{t\to\infty} t^{-1}N_i(t)$.

Exercise 11. *Continuation.* In the context of Exercise 10, let $X(t)$ denote the type of job being processed at time t. Find the limiting distribution of $X(t)$. Find the portion of time devoted to type i jobs.

Exercise 12. *Continuation.* Consider the multi-type renewal process with two types of renewals that have exponential distributions with rates λ_i, and type i occurs with probability p_i, , $i = 1, 2$. Show that the renewal function has the density

$$U'(t) = \frac{\lambda_1\lambda_2 + p_1p_2(\lambda_1 - \lambda_2)^2 e^{-(p_1\lambda_2+p_2\lambda_1)t}}{p_1\lambda_2 + p_2\lambda_1}, \qquad t > 0.$$

Exercise 13. *System Availability.* The status of a system is represented by an alternating renewal process $X(t)$, where the mean sojourn time in a working state 1 is μ_1 and the mean sojourn time in a non-working state 0 is μ_0. The system *availability* is measured by the portion of time it is working, which is $\lim_{t\to\infty} t^{-1}\int_0^t X(s)ds$. Determine this quantity and show that it is equal to the *cycle-availability* measured by $\lim_{n\to\infty} T_n^{-1}\int_0^{T_n} X(s)ds$.

Exercise 14. *Integrals of Renewal Processes.* Suppose $N(t)$ is a renewal process with renewal times T_n and $\mu = E[T_1]$. Prove

$$E\Big[\int_0^{T_n} N(s)\,ds\Big] = \mu n(n-1)/2.$$

For any non-random $t > 0$, it follows by Fubini's theorem that

$$E\Big[\int_0^{t} N(s)\,ds\Big] = \int_0^{t} E[N(s)]\,ds.$$

Assuming τ is an exponential random variable independent of N with rate γ, prove

$$E\Big[\int_0^{\tau} N(s)\,ds\Big] = \int_{\mathbb{R}_+} e^{-\gamma t}E[N(t)]\,dt.$$

Show that if N is a Poisson process with rate λ, then the preceding expectation equals λ/γ^2. (Integrals like these are used to model holding costs; see Section 2.12 and the next exercise.)

Exercise 15. *Continuation.* Items arrive to a service station according to a Poisson process $N(t)$ with rate λ. The items are stored until m have accumulated. Then the m items are served in a batch. The service time is exponentially distributed with rate γ. During the service, items continue to arrive. There is a cost hi per unit time of holding i customers in the system. Assume the station is empty at time 0. Find the expected cost C_1 of holding the customers until m have arrived. Find the expected cost C_2 for holding the added arrivals in the system during the service.

Exercise 16. Customers arrive to a service system according to a Poisson process with rate λ. The system can only serve one customer at a time and, while it is busy serving a customer, arriving customers are blocked from getting service (they may seek service elsewhere or simply go unserved). Assume the service times are independent with common distribution G and are independent of the arrival process. For instance, a contractor may only be able to handle one project at a time (or a vehicle can only transport one item at a time). Determine the following quantities:
(a) The portions of time the system is busy, and not busy.
(b) The number of customers per unit time that are served.
(c) The portion of customers that are blocked from service.

Exercise 17. *Delayed Renewals.* A point process $N(t)$ is an *m-step delayed renewal process* if the inter-occurrence times ξ_{m+k}, for $k \geq 1$, are independent with a common distribution F, and no other restrictions are placed on $\xi_1, \ldots, \xi_m$. That is, $N_m(t) = N(t) - N(T_m)$, for $t \geq T_m$ is a renewal process. Show that Corollary 11 and Theorem 13 hold for such processes. Use the fact that $N(t)$ is asymptotically equivalent to $N_m(t)$ in that

$$N_m(t)/N(t) = 1 - N(T_m)/N(t) \to 1, \quad \text{a.s. as } t \to \infty.$$

Exercise 18. For a point process $N(t)$ that is not simple, show that if $t^{-1}N(t) \to 1/\mu$ as $t \to \infty$, then $n^{-1}T_n \to \mu$, as $n \to \infty$. Hint: For a fixed positive constant c, note that $N((T_n - c)^+) \leq n \leq N(T_n)$. Divide these terms by T_n and take limits as $n \to \infty$.

Exercise 19. *Age Replacement Model.* An item (e.g., battery, vehicle, tool, or electronic component) whose use is needed continuously is replaced whenever it fails or reaches age a, whichever comes first. The successive items are independent and have the same lifetime distribution G. The cost of a failure is c_f dollars and the cost of a replacement at age a is c_r. Show that the average cost per unit time is

$$C(a) = [c_f G(a) + c_r(1 - G(a))] / \int_0^a (1 - G(s))ds.$$

Find the optimal age a that minimizes this average cost.

Exercise 20. *Point Processes as Jump Processes.* Consider a point process $N(t) = \sum_{k=1}^{\infty} \mathbf{1}(T_k \le t)$, where $T_1 \le T_2 \le \cdots$. It can also be formulated as an integer-valued jump process of the form

$$N(t) = \sum_{n=1}^{\infty} \nu_n \mathbf{1}(\hat{T}_n \le t),$$

where $\hat{T}_n$ are the "distinct" times at which $N(t)$ takes a jump, and ν_n is the size of the jump. That is, $\hat{T}_n = \min\{T_k : T_k > \hat{T}_{n-1}\}$, where $\hat{T}_0 = 0$, and $\nu_n = \sum_{k=1}^{\infty} \mathbf{1}(T_k = \hat{T}_n)$, $n \ge 1$.

For instance, suppose T_n are times at which data packets arrive to a computer file. Then $\hat{T}_n$ are the times at which batches of packets arrive, and at time $\hat{T}_n$, a batch of ν_n packets arrive. Suppose $\hat{T}_n$ are renewal times, and ν_n are i.i.d. and independent of the $\hat{T}_n$. Show that the number of packets that arrive per unit time is $E[\nu_1]/E[\hat{T}_1]$ a.s., provided these expectations are finite. Next, assume $\hat{T}_n$ form a Poisson process with rate λ, and ν_n has a Poisson distribution. Find $E[N(t)]$ by elementary reasoning, and then show that $N(t)$ has a Poisson distribution.

Exercise 21. *Batch Renewals.* Consider times $T_n = \sum_{k=1}^{n} \xi_k$, where the ξ_k are i.i.d. with distribution F and $F(0) = P\{\xi_k = 0\} > 0$. The associated point process $N(t)$ is a renewal process with *instantaneous renewals* (or batch renewals). In the notation of Exercise 20, $N(t) = \sum_{n=1}^{\infty} \nu_n \mathbf{1}(\hat{T}_n \le t)$, where ν_n is the number of renewals exactly at time $\hat{T}_n$. Specify the distribution of ν_n. Are the ν_n i.i.d.? Are they independent of $\hat{T}_n$? Specify the distribution of $\hat{T}_1$ in terms of F.

Exercise 22. Prove $E[T_{N(t)}] = \mu E[N(t) + 1] - E[\xi_{N(t)}]$. If $N(t)$ is a Poisson process, show that $E[T_{N(t)}] < \mu E[N(t)]$.

Exercise 23. *Little Law.* In the context of the Little law in Theorem 57, show that if L and W exist, then λ exists and $L = \lambda W$.

Exercise 24. *Superpositions of Renewal Processes.* Let $N_1(t)$ and $N_2(t)$ be independent renewal processes with the same inter-renewal distribution, and consider the sum $N(t) = N_1(t) + N_2(t)$ (sometimes called a *superposition*). Assuming that $N(t)$ is a renewal process. prove that it is a Poisson process if and only if $N_1(t)$ and $N_2(t)$ are Poisson processes.

Exercise 25. *Production-Inventory Model.* Consider a production-inventory system that produces a product at a constant rate of c units per unit time and the items are put in inventory to satisfy demands. The products may be discrete or continuous (e.g., oil, chemicals). Demands occur according to a Poisson process $N(t)$ with rate λ, and the demand quantities $X_1, X_2, \ldots$ are independent, identically distributed positive random variables with mean μ,

and are independent of the arrival times. Then the inventory level at time t would be

$$Z_x(t) = x + ct - \sum_{n=1}^{N(t)} X_n, \qquad t \geq 0,$$

where x is the initial inventory level. Consider the probability $R(x) = P\{Z_x(t) \geq 0, t \geq 0\}$ of never running out of inventory. Show that if $c < \lambda\mu$, then $R(x) = 0$ no matter how high the initial inventory level x is. Hint: apply a SLLN to show that $Z_x(t) \to -\infty$ as $t \to \infty$ if $c < \lambda\mu$, where x is fixed. Find the limit of $Z_x(t)$ as $t \to \infty$ if $c > \lambda\mu$. (The process $Z_x(t)$ is a classical model of the capital of an insurance company; see Example 73.)

Exercise 26. Let $H(t) = E[N(t) - N(t-a)\mathbf{1}(a \leq t)]$. Find a renewal equation that H satisfies.

Exercise 27. *Non-homogeneous Renewals.* Suppose $N(t)$ is a point process on $\mathbb{R}_+$ whose inter-point times $\xi_n = T_n - T_{n-1}$ are independent with distributions F_n. Assuming it is finite, prove that $E[N(t)] = \sum_{n=1}^{\infty} F_1 \star \cdots \star F_n(t)$.

Exercise 28. *Subadditivity of Renewal Function.* Prove that

$$U(t+a) \leq U(a) + U(t), \qquad a, t \geq 0.$$

Hint: Use $a \leq T_{N(a)+1}$ in the expression

$$N(t+a) - N(a) = \sum_{k=1}^{\infty} \mathbf{1}(T_{N(a)+k} \leq t + a).$$

Exercise 29. *Arithmetic Blackwell Theorem.* The proof of Theorem 33 for arithmetic inter-arrival distributions was proved under the standard condition that $p_0 = F(0) = 0$. Prove the same theorem when $0 < p_0 < 1$. Use a similar argument including the fact that renewals occur in batches and a batch size has a geometric distribution with parameter $1 - p_0$.

Exercise 30. *Elementary Renewal Theorem via Blackwell.* Prove the elementary renewal theorem (Theorem 32) by an application of Blackwell's theorem. One approach, for non-arithmetic F, is to use

$$E[N(t)] = \sum_{k=1}^{\lceil t \rceil} [U(k) - U(k-1)] + E[N(\lceil t \rceil)] - E[N(t)].$$

Then use the fact $n^{-1} \sum_{k=1}^{n} c_k \to c$ when $c_k \to c$.

Exercise 31. *Arithmetic Key Renewal Theorem.* Represent $U \star h(u + nd)$ as a sum like (2.19), and then prove Theorem 37 by applying Blackwell's theorem.

Exercise 32. Let $h(t) = \sum_{n=1}^{\infty} a_n \mathbf{1}(n-\varepsilon_n \le t < n+\varepsilon_n)$, where $a_n \to \infty$ and $1/2 > \varepsilon_n \downarrow 0$ such that $\sum_{n=1}^{\infty} a_n \varepsilon_n < \infty$. Show that h is Riemann integrable, but not DRI.

Exercise 33. Prove that a continuous function $h(t) \ge 0$ is DRI if and only if $I^{\delta}(h) < \infty$ for some $\delta > 0$.

The next eight exercises concern the renewal process trinity: the backward and forward recurrence times $A(t) = t - T_{N(t)}$, $B(t) = T_{N(t)+1} - t$, and the length $L(t) = \xi_{N(t)+1} = A(t) + B(t)$ of the renewal interval containing t. Assume the inter-renewal distribution is non-arithmetic.

Exercise 34. Draw a typical sample path for each of the processes $A(t)$, $B(t)$, and $L(t)$.

Exercise 35. Prove that $B(t)$ is a Markov process by showing it satisfies the following Markov property, for $x, y, t, u \ge 0$:

$$P\{B(t+u) \le y | B(s) : s < t, B(t) = x\} = P\{B(u) \le y | B(0) = x\}.$$

Exercise 36. Formulate a renewal equation that $P\{B(t) > x\}$ satisfies.

Exercise 37. *Bypassing a renewal equation.* Use Proposition 40 (without using a renewal equation) to prove $P\{B(t) > x\} = \int_{[0,t]} [1-F(t+x-s)]dU(s)$.

Exercise 38. Prove $E[B(t)] = \mu E[N(t)+1] - t$. Assuming F has a finite variance σ^2, prove

$$\lim_{t\to\infty} E[A(t)] = \lim_{t\to\infty} E[B(t)] = \frac{\sigma^2 + \mu^2}{2\mu}.$$

Is this limit also the mean of the limiting distribution $F_e(t) = \frac{1}{\mu}\int_0^t [1-F(s)]ds$ of $A(t)$ and $B(t)$?

Exercise 39. *Inspection Paradox.* Consider the length $L(t) = \xi_{N(t)+1}$ of the renewal interval at any time t (this is what an inspector of the process would see at time t). Prove the paradoxical result that $L(t)$ is *stochastically larger* than the length ξ_1 of a typical renewal interval; that is

$$P\{L(t) > x\} \ge P\{\xi_1 > x\}, \qquad t, x \ge 0.$$

This inequality is understandable upon observing that the first probability is for the event that a renewal interval bigger than x "covers" t, and this is more likely to happen than a fixed renewal interval being bigger than x. A consequence of this result is $E[L(t] \ge E[\xi_1]$, which is often a strict inequality.

Suppose $\mu = E[T_1]$ and $\sigma^2 = \mathrm{Var}[T_1]$ are finite. Recall from (2.34) that the limiting distribution of $L(t)$ is $\frac{1}{\mu}\int_0^x s dF(s)$. Derive the mean of this distribution (as a function of μ and σ^2), and show it is $\ge \mu$.

Show that if $N(t)$ is a Poisson process with rate λ, then

$$E[L(t)] = \lambda^{-1}[2 - (1 + \lambda t)e^{-\lambda t}].$$

In this case, $E[L(t] > E[\xi_1]$.

Exercise 40. Prove $\lim_{t\to\infty} P\{A(t)/L(t) \le x\} = x$, $0 \le x \le 1$. Prove this result with $B(t)$ in place of $A(t)$.

Exercise 41. Show that

$$\lim_{t\to\infty} E[A(t)^k B(t)^\ell (A(t) + B(t))^m] = \frac{E[T_1^{k+\ell+m}]}{\mu(k+\ell+1)\binom{k+\ell}{k}}.$$

Find the limiting covariance, $\lim_{t\to\infty} \mathrm{Cov}(A(t), B(t))$.

Exercise 42. *Delayed Versus Non-delayed Regenerations.* Let $\tilde{X}(t)$ be a real-valued delayed regenerative process over T_n. Then $X(t) = \tilde{X}(T_1+t)$, $t \ge 0$ is a regenerative process. Assuming $\tilde{X}(t)$ is bounded, show that if $\lim_{t\to\infty} E[X(t)]$ exists (such as by Theorem 45), then $E[\tilde{X}(t)]$ has the same limit. Hint: Take the limit as $t \to \infty$ of

$$E[\tilde{X}(t)] = \int_{[0,t]} E[X(t-s)]dF(s) + E[\tilde{X}(t)\mathbf{1}(T_1 > t)].$$

Exercise 43. *Dispatching System.* Items arrive at a depot (warehouse or computer) at times that form a renewal process with finite mean μ between arrivals. Whenever M items accumulate, they are instantaneously removed (dispatched) from the depot. Let $X(t)$ denote the number of items in the depot at time t. Find the limiting probability that there are p_j items in the system ($j = 0, \ldots, M-1$). Find the average number of items in the system over an infinite time horizon.

Suppose the batch size M is to be selected to minimize the average cost of running the system. The relevant costs are a cost C for dispatching the items, and a cost h per unit time for holding an item in the depot. Let $C(M)$ denote the average dispatching plus holding cost for running the system with batch size M. Find an expression for $C(M)$. Show that the value of M that minimizes $C(M)$ is an integer adjacent to the value $M^* = \sqrt{2C/h\mu}$.

Exercise 44. *Continuation.* In the context of the preceding exercise, find the average time W that a typical item waits in the system before being dispatched. Find the average waiting time $W(i)$ in the system for the ith arrival in the batch.

Exercise 45. Consider an ergodic Markov chain X_n with limiting distribution π_i. Prove

$$\lim_{n\to\infty} P\{X_n = j, X_{n+1} = \ell\} = \pi_j p_{j\ell}.$$

One can show that (X_n, X_{n+1}) is a two-dimensional Markov chain that is ergodic with the preceding limiting distribution. However, establish the limit above only with the knowledge that X_n has a limiting distribution.

Exercise 46. Items with volumes $V_1, V_2, \ldots$ are loaded on a truck one at a time until the addition of an arriving item would exceed the capacity v of the truck. Then the truck leaves to deliver the items. The number of items that can be loaded in the truck before its volume v is exceeded is

$$N(v) = \min\{n : \sum_{k=1}^{n} V_k > v\} - 1.$$

Assume the V_n are independent with identical distribution F that has a mean μ and variance σ^2. Suppose the capacity v is large compared to μ. Specify a single value that would be a good approximation for $N(v)$. What would be a good approximation for $E[N(v)]$? Specify how to approximate the distribution of $N(v)$ by a normal distribution. Assign specific numerical values for μ, σ^2, and v, and use the normal distribution to approximate the probability $P\{a \leq N(v) \leq b\}$ for a few values of a and b.

Exercise 47. *Limiting Distribution of a Cyclic Renewal Process.* Consider a cyclic renewal process $X(t)$ on the states $0, 1, \ldots, K-1$ as described in Example 8. Its inter-renewal distribution is $F = F_0 \star \cdots \star F_{K-1}$, where F_i is distribution of a sojourn time in state i having a finite mean μ_i. Assume one of the F_i is non-arithmetic. Show that F is non-arithmetic. Prove

$$\lim_{t\to\infty} P\{X(t) = i\} = \frac{\mu_i}{\mu_0 + \ldots + \mu_{K-1}}.$$

Is this limiting distribution the same as $\lim_{t\to\infty} t^{-1} E[\int_0^t \mathbf{1}(X(s) = i)ds]$, the average expected time spent in state i? State any additional assumptions needed for the existence of this limit.

Exercise 48. Consider a $G/G/1$ system as in Example 60. Let W'_n denote the length of time the nth customer waits in the queue prior to obtaining service. Determine a Little law for the average wait $W' = \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} W'_k$.

Exercise 49. *System Down Time.* Consider an alternating renewal process that represents the up and down states of a system. Suppose the up times have a distribution G with mean μ and variance σ^2, and the down times have a distribution G_0 with mean μ_0 and variance σ_0^2. Let $D(t)$ denote the length of time the system is down in the time interval $[0, t]$. Find the average expected down time $\beta = \lim_{t\to\infty} t^{-1} E[D(t)]$. Then show $(D(t) - \beta t)/t^{1/2} \xrightarrow{d} N(0, \gamma^2)$ and specify γ.

Exercise 50. *Congestion in a Running Race.* The following model was developed by Georgia Tech undergraduate students to assess the congestion in

the 10-kilometer Atlanta Road Race, which is held every July 4th. After the pack of elite runners begins the race, the rest of the runners start the race a little later as follows. The runners are partitioned into m groups, with r_k runners assigned to group k, $1 \le k \le m$, depending on their anticipated completion times (the runners in each group being about equal in ability). The groups are released every τ minutes, with group k starting the race at time $k\tau$ (the groups are ordered so that the faster runners go earlier). Although the group sizes r_k are random, assume for simplicity that they are not. Typical numbers are 10 groups of 5000 runners in each group. The aim was to design the race so that the congestion did not exceed a critical level that would force runners to walk. To do this, the students developed a model for computing the probability that the congestion would be above the critical level. (They used this model to determine reasonable group sizes and their start times under which the runners would start as soon as possible, with a low probability of runners being forced to walk.)

The students assumed the velocity of each runner is the same throughout the race, the velocities of all the runners are independent, and the velocity of each runner in group k has the same distribution F_k. The distributions F_k were based on empirical distributions from samples obtained in prior races. Using pictures of past races, it was determined that if the number of runners in an interval of length ℓ in the road was greater than b, then the runners in that interval would be forced to walk. This was based on observing pictures of congestion in past races where the runners had to slow down to a walk.

Under these assumptions, the number of runners in group k that are in an interval $[a, a+\ell]$ on the road at time t is

$$Z_a^k(t) = \sum_{n=1}^{r_k} \mathbf{1}(V_{kn}(t - k\tau) \in [a, a+\ell]),$$

where $V_{k1}, \ldots, V_{kr_k}$ are the independent velocities of the runners in group k that have the distribution F_k. Then the total number of runners that are in $[a, a+\ell]$ at time t is

$$Z_a(t) = \sum_{k=1}^{m} Z_a^k(t).$$

Specify how one would use the central limit theorem to compute the probability $P\{Z_a(t) > b\}$ that the runners in $[a, a+\ell]$ at time t would be forced to walk.

Exercise 51. *Confidence Interval.* In the context of Example 66, suppose the regenerative-increment process $Z(t)$ is not observed continuously over time, but only observed at its regeneration times T_n. In this case, the information observed up to the nth regeneration time T_n is $\{Z(t) : t \le T_n\}$. First, find the a.s. limit of $Z(T_n)/n$, and the limiting distribution of $(Z(T_n) - aT_n)/n^{1/2}$.

Then find an approximate confidence interval for the mean a analogous to that in Example 66.

Exercise 52. *Continuation.* Use the CLT in Examples 67 and 68 to obtain approximate confidence intervals for a renewal process and a Markov chain comparable to the confidence interval in Example 66.

Exercise 53. Consider a delayed renewal process $N(t)$ with initial distribution $F_e(x) = \frac{1}{\mu}\int_0^x [1 - F(s)]ds$. Prove $E[N(t)] = U \star F_e(t) = t/\mu$ by a direct evaluation of the integral representing the convolution, where $U = \sum_{n=0}^{\infty} F^{n\star}$.

Exercise 54. Justify expression (2.49), which in expanded form is

$$\sigma^2 = \mu_i^{-1}\mathrm{Var}[Z_{\nu_1} - a\nu_1] = \sum_{j\in S}\pi_j\tilde{f}(j)^2 + 2\sum_{j\in S}\pi_j\tilde{f}(j)\sum_{k\in S}\sum_{n=1}^{\infty}p_{jk}^n\tilde{f}(k).$$

First show that $E[Z_{\nu_1} - a\nu_1)] = 0$, and then and use the expansion

$$\begin{aligned}\mathrm{Var}[Z_{\nu_1} - a\nu_1] &= E_i\Big[[\sum_{n=1}^{\nu_1}\tilde{f}(X_n)]^2\Big] \\ &= E_i\Big[\sum_{n=1}^{\nu_1}\tilde{f}(X_n)^2\Big] + 2E_i\Big[\sum_{n=1}^{\nu_1}V_n\Big], \qquad (2.76)\end{aligned}$$

where $V_n = \tilde{f}(X_n)\sum_{\ell=n+1}^{\nu_1}\tilde{f}(X_\ell)$. Apply Proposition 69 from Chapter 1 to the last two expressions in (2.76)(noting that $\sum_{n=1}^{\nu_1}V_n = \sum_{n=0}^{\nu_1-1}V_n$). Use the fact that

$$E_i[V_n|X_n = j, \nu_1 \geq n] = \tilde{f}(j)h(j),$$

where $h(j) = E_j[\sum_{n=1}^{\nu_1}\tilde{f}(X_n)]$ satisfies the equation

$$h(j) = \sum_{k\in S}p_{jk}\tilde{f}(k) + \sum_{k\in S}p_{jk}h(k),$$

and hence $h(j) = \sum_{k\in S}\sum_{n=1}^{\infty}p_{jk}^n\tilde{f}(k)$.

Exercise 55. In the context of Theorem 76, show that if the distribution of the residual time $B(t)$ is independent of t, then it is a stationary process. Hint: For any s_i, x_i and t, let

$$\Gamma_t = \{B(s_1 + t) \leq x_1, \ldots, B(s_k + t) \leq x_k\}.$$

Show that $P\{\Gamma_t|B(t) = x\} = P\{\Gamma_0|B(0) = x\}$, and use this equality to prove $P(\Gamma_t)$ is independent of t.

Chapter 3
Poisson Processes

Poisson processes are used extensively in applied probability models. Their importance is due to their versatility for representing a variety of physical processes, and because a Poisson process is a natural model for a sum of many sparse point processes. The most basic Poisson process, introduced in the preceding chapter, is a renewal process on the time axis with exponential inter-renewal times. This type of process is useful for representing times at which an event occurs, such as the times at which items arrive to a network, machine components fail, emergencies occur, a stock price takes a large jump, etc. The first part of the present chapter continues the discussion of this basic Poisson process by presenting several characterizations of it, including the result that its point locations (i.e., occurrence times) on a finite time interval are equal in distribution to order-statistics from a uniform distribution on the interval.

Applications of classical Poisson processes often involve auxiliary marks or random elements associated with the event occurrence times. For instance, if items arrive to a network at times that form a Poisson process, then a typical mark for an arriving item might be a vector denoting its route in the network and its service times at the nodes on the route. A convenient approach for analyzing such marks is to consider them as part of a larger "space-time" marked Poisson process on a multidimensional space. The properties of these processes are similar to those of "spatial" Poisson processes used for modeling locations of discrete items in the plane or a Euclidean space such as cell phone calls, truck delivery points, disease centers, geological formations, particles in space, fish colonies, etc.

Following the discussion of classical Poisson processes on the time axis, we describe the structure of contemporary Poisson processes on general spaces, which includes space-time and spatial Poisson processes. The methodology for Poisson processes on general spaces involves the use of counting processes on general spaces and their representation by Laplace functionals. A Poisson process on a general space is characterized in terms of a mixed binomial

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

process. This is a generalization of the uniform order-statistic characterization of a classical Poisson process.

Several sections describe summations, partitions, translations and general transformations of Poisson processes. Included are applications of space-time Poisson processes for analyzing particle systems and stochastic networks. Next, we show that many properties of Poisson processes readily extend to several related processes; namely, Cox processes (i.e., Poisson processes with random intensities), compound Poisson processes, and cluster processes. The final results are laws of small numbers (like central limit theorems) for rare events or points, including the Poisson approximation for a binomial distribution. They justify that a Poisson process is a natural limit for a collection (or sum) of many sparse families of random points.

3.1 Poisson Processes on $\mathbb{R}_+$

As in the last chapter, we define a *point process* $N = \{N(t) : t \geq 0\}$ on $\mathbb{R}_+$ as a counting process $N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t)$, where $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$ are random points (or times) such that $T_n \to \infty$ a.s. as $n \to \infty$. The point process N is *simple* when the points are distinct ($T_0 < T_1 < \ldots$ a.s.).

We will also refer to the point process as the set of random variables $N = \{N(B) : B \in \mathcal{B}_+\}$, where

$$N(B) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \in B), \qquad B \in \mathcal{B}_+,$$

denotes the number of points in the set B, and $\mathcal{B}_+$ denotes the Borel sets in $\mathbb{R}_+$ (see Section 6.1 in the Appendix). Note that $N(B)$ is finite when B is bounded since $T_n \to \infty$ a.s. However, $N(B)$ may be infinite when the set B is not bounded, and $E[N(B)]$ may be infinite even though $N(B)$ is finite. In addition, we write

$$N(a, b] = N((a, b]) = N(b) - N(a), \qquad a \leq b.$$

In the last chapter, a renewal process with exponential inter-renewal times was said to be a Poisson process. This characterization, as we show in the next section, is equivalent to the following one.

Definition 1. A simple point process $N = \{N(t) : t \geq 0\}$ on $\mathbb{R}_+$ is a *Poisson process* with rate $\lambda > 0$ if it satisfies the following properties.

- It has *independent increments*: $N(s_1, t_1], \ldots, N(s_n, t_n]$ are independent, for $s_1 < t_1 \cdots < s_n < t_n$.
- $N(s, t]$ has a Poisson distribution with mean $\lambda(t - s)$, for any $s < t$.

A Poisson process N with rate λ is sometimes called a *homogeneous* or *time-stationary* Poisson process, or a *classical* Poisson process. Under the

preceding definition, Theorem 4 below establishes that N is also a renewal process whose inter-renewal times are independent exponentially distributed with rate λ.

A number of elementary properties of N follow from this renewal characterization. For instance, we saw in Example 5 that the time T_n of the nth renewal has the distribution

$$P\{T_n \le t\} = P\{N(t) \ge n\} = 1 - \sum_{k=0}^{n-1} (\lambda t)^k e^{-\lambda t}/k!.$$

The derivative of this expression is $f(t) = \lambda^{n+1} t^n e^{-\lambda t}/n!$, and hence T_n has a gamma distribution with parameters n and λ.

Keep in mind that all the properties of renewal processes apply to N. For instance, $t^{-1}N(t) \to \lambda$ a.s. by Corollary 11 in Chapter 2. Another important fact is that the Poisson process N is also a continuous-time Markov chain as in Chapter 4.

Some applications of Poisson processes involve only elementary properties of the processes. Here are two examples.

Example 2. Comparing Arrival Times. Two types of items arrive at a station for processing by independent Poisson processes with respective rates λ and λ'. Of interest is the probability that n λ-arrivals come before the first λ'-arrival. This probability can be expressed as

$$P\{T_n < T'\} = \Big(\frac{\lambda}{\lambda + \lambda'}\Big)^n,$$

where T_n is the time of the nth λ-arrival and T' is the time of the first λ'-arrival. Indeed, since T' has an exponential distribution with rate λ' and it is independent of T_n,

$$\begin{aligned} P\{T_n < T'\} &= \int_0^\infty P\{T' > t | T_n = t\} P\{T_n \in dt\} \\ &= \int_0^\infty e^{-\lambda' t} P\{T_n \in dt\} = E[e^{-\lambda' T_n}] = \Big(\frac{\lambda}{\lambda + \lambda'}\Big)^n. \end{aligned}$$

The last term is the Laplace transform of the gamma random variable T_n with parameters n and λ. More general results on comparing arrival times are given in Exercises 6 and 21.

Example 3. Optimal Dispatching. Consider a system in which discrete items arrive to a dispatching station according to a Poisson process N with rate λ during a fixed time interval $[0, T]$. The items might represent people to be bussed, ships to be unloaded, computer data or messages to be forwarded, material to be shipped, etc. There is a cost of h dollars per unit time of holding one item in the system. Also, at any time during the period, the items may be dispatched (or processed) at a cost of c dollars, and a dispatch

is automatically done at time T. A dispatch is performed instantaneously and all the items in the system at that time are dispatched. Consider a dispatching policy defined by a vector $(n, t_1, \ldots, t_n)$, where n is the number of dispatches to make in the period, and $t_1 < t_2 < \cdots < t_n = T$ are the times of the dispatches. The aim is to find a dispatching policy that minimizes the expected cost.

We will show that the optimal solution is to have n^* dispatches at the times $t_i^* = iT/n^*$, where

$$n^* = \begin{cases} \lfloor x \rfloor \text{ if } \lfloor x \rfloor \lceil x \rceil \geq x^2 \\ \lceil x \rceil \text{ otherwise,} \end{cases} \tag{3.1}$$

and $x = T(h\lambda/2c)^{1/2}$.

This type of policy is a "static" policy in that it is implemented at the beginning of the time period and remains in effect during the period regardless of how the items actually arrive (e.g., there may be 0 items in a dispatch at a predetermined dispatch time t_i). A static policy might be appropriate when it is not feasible or too costly to monitor the system and do real-time dispatching. An alternative is to use a "dynamic" control policy that involves deciding when to make dispatches based on the observed queue of units. Exercise 11 asks if the policy above is optimal for non-Poisson processes.

To solve the problem, we will derive expressions for the total cost and its mean, and then find optimal values of the policy parameters. Under a fixed policy $(n, t_1, \ldots, t_n)$, the total cost is

$$Z = cn + h \sum_{i=1}^{n} W_i,$$

where W_i is the amount of time that items wait in the system during the time period $(t_{i-1}, t_i]$. Since $N(a, b]$ is the number of arrivals in a time interval $(a, b]$, it follows that

$$W_i = \int_{t_{i-1}}^{t_i} N(t_{i-1}, s]\, ds.$$

Using Fubini's theorem and $E[N(a, b]] = \lambda(b - a)$,

$$\begin{aligned} E[W_i] &= \int_{t_{i-1}}^{t_i} E[N(t_{i-1}, s]]\, ds \\ &= \lambda \int_{t_{i-1}}^{t_i} (s - t_{i-1})\, ds = \lambda(t_i - t_{i-1})^2/2. \end{aligned}$$

Then the expected cost under the policy $(n, t_1, \ldots, t_n)$ is

$$f(n, t_1, \ldots, t_n) = E[Z] = cn + \frac{h\lambda}{2} \sum_{i=1}^{n} (t_i - t_{i-1})^2.$$

Therefore, the aim is to solve the optimization problem

$$\min_n \min_{t_1,\ldots,t_n} f(n, t_1, \ldots, t_n)], \tag{3.2}$$

subject to $t_{i-1} < t_i$ and $\sum_{i=1}^n (t_i - t_{i-1}) = T$.

It is well-known that the problem $\min_{x_1\ldots,x_n} \sum_{i=1}^n x_i^2$, under the constraint $\sum_{i=1}^n x_i = T$, has the solution $x_i^* = T/n$. This result follows by dynamic programming (backward induction), or by the use of Lagrange multipliers.

Applying this result to the problem (3.2), it follows that for fixed n, the subproblem $\min_{t_1,\ldots,t_n} f(n, t_1, \ldots, t_n)$ has the solution $t_i^* - t_{i-1}^* = T/n$, so that $t_i^* = iT/n$. Also, note that

$$g(n) = f(n, t_1^*, \ldots, t_n^*) = nc + h\lambda T^2/2n.$$

Then to solve (3.2), it remains to solve $\min_n g(n)$. Viewing n as a continuous variable x, the derivative $g'(x) = c - h\lambda T^2/(2x^2)$ is nondecreasing. Then $g(x)$ is convex and it is minimized at $x^* = T(h\lambda/2c)^{1/2}$. So the integer that minimizes $g(n)$ is either $\lfloor x^* \rfloor$ or $\lceil x^* \rceil$. Now $g(\lfloor x^* \rfloor) \le g(\lceil x^* \rceil)$ if and only if $\lfloor x^* \rfloor \lceil x^* \rceil \ge (x^*)^2$. This yields (3.1).

3.2 Characterizations of Classical Poisson Processes

Another way of characterizing a Poisson process is that it has independent increments and satisfies the following infinitesimal properties: the probability of a point in a small interval is directly proportional to the interval length and the probability of having more than one point in such an interval is essentially 0. These properties are the basis of differential equations for probabilities of the process whose solutions are Poisson probabilities.

This section consists of the following theorem, which covers the characterization we just mentioned and the renewal characterization as well.

Theorem 4. *For a simple point process $N = \{N(t) : t \ge 0\}$ on $\mathbb{R}_+$ and $\lambda > 0$, the following statements are equivalent.*
(a) *N is a Poisson process with rate λ.*
(b) *N is a renewal process whose inter-renewal times are exponentially distributed with rate λ.*
(c) *N has independent increments and, for any t and $h \downarrow 0$,*

$$P\{N(t, t+h] = 1\} = \lambda h + o(h), \qquad P\{N(t, t+h] \ge 2\} = o(h). \tag{3.3}$$

Proof. (a) $\Rightarrow$ (b). Assertion (b) states that $\xi_n = T_n - T_{n-1}$, $n \ge 1$, are independent and have an exponential distribution with rate λ. That is,

$$P(A_n) = e^{-\lambda \sum_{i=1}^n t_i}, \quad n \ge 1, \tag{3.4}$$

where $A_n = \{\xi_1 > t_1, \ldots, \xi_n > t_n\}$ for $t_i > 0$.

Assuming N is a Poisson process with rate λ, we will prove (3.4) by induction. It is true for $n = 1$ since

$$P(A_1) = P\{\xi_1 > t_1\} = P\{N(t_1) = 0\} = e^{-\lambda t_1}.$$

Now assume (3.4) is true for some $n - 1$. Then

$$P(A_n) = P(A_{n-1}, \xi_n > t_n) = P(A_{n-1})P\{\xi_n > t_n | A_{n-1}\}. \tag{3.5}$$

Using Poisson properties of N,

$$\begin{aligned} P\{\xi_n > t_n | A_{n-1}\} &= E\left[P\{\xi_n > t_n | A_{n-1}, T_n\}\right] \\ &= E\left[P\{N(T_n, T_n + t_n] = 0 | A_{n-1}, T_n\}\right] \\ &= P\{N(t_n) = 0\} = e^{-\lambda t_n}. \end{aligned}$$

Substituting this in (3.5), along with (3.4) for $n - 1$, yields (3.4) for n.

(b) $\Rightarrow$ (c). Suppose (b) is true. By Corollary 78 in Chapter 2, we know that N is stationary. In particular, $N(t, t + h] \stackrel{d}{=} N(h)$, for $h, t \geq 0$. Now, because the inter-renewal times are exponentially distributed with rate λ,

$$\begin{aligned} &P\{N(h) = 0\} = P\{T_1 > h\} = e^{-\lambda h}, \\ &P\{N(h) = 1\} = P\{T_1 \leq h, T_2 > h\} = \int_0^h e^{-\lambda(h-s)} \lambda e^{-\lambda s} ds = \lambda h e^{-\lambda h}. \end{aligned}$$

Then using $e^{-\lambda h} = 1 - \lambda h + o(h)$ in these expressions yields (3.3).

The proof of (c) will be complete upon showing that N has independent increments: for any $s_1 < t_1 < \cdots < s_n < t_n$,

$$N(s_1, t_1], \ldots, N(s_n, t_n], \quad n \geq 1, \text{ are independent.}$$

Proceeding by induction, the statement is obviously true for $n = 1$. Next, assume it is true for some n. Now, the statement for $n+1$ will follow by showing that $Z_n = N(s_{n+1}, t_{n+1}]$ is independent of $Y_n = (N(s_1, t_1], \ldots, N(s_n, t_n])$. Our proof will use the fact from Example 49 in Chapter 2 that the forward recurrence time $B(t) = T_{N(t)+1} - t$ at time t is exponentially distributed with rate λ, and it is independent of the renewals in $[0, t]$.

Using this property at time t_n, it follows that Z_n is conditionally independent of Y_n given $B(t_n)$, and so

$$P\{Y_n = k | Z_n\} = E[P\{Y_n = k | B(t_n), Z_n\} | Z_n] = E[P\{Y_n = k | B(t_n)\}].$$

The last term is $P\{Y_n = k\}$, which does not depend on Z_n. Thus, Y_n is independent of Z_n.

(c) $\Rightarrow$ (a). Assuming (c) is true, (a) will follow by proving $N(s,t]$ has a Poisson distribution with mean $\lambda(t-s)$, for each $s<t$. We begin with the case $s=0$ and prove

$$p_n(t) = P\{N(t)=n\} = (\lambda t)^n e^{-\lambda t}/n!, \qquad n \geq 0.$$

We will establish this by deriving differential equations for the functions $p_n(t)$ and showing that their solutions are the preceding Poisson probabilities.

Under the assumptions in (c), for $n \geq 1$,

$$\begin{aligned} p_n(t+h) = P\{&N(t)=n, N(t,t+h]=0\} \\ &+P\{N(t)=n-1, N(t,t+h]=1\} \\ &+P\{N(t+h)=n, N(t,t+h]\geq 2\}. \end{aligned}$$

Since the last probability is $\leq P\{N(t,t+h]\geq 2\} = o(h)$, and N has independent increments, the preceding is

$$p_n(t+h) = p_n(t)p_0(h) + p_{n-1}(t)p_1(h) + o(h).$$

In light of this expression, $p_n(t)$ is right-continuous on $\mathbb{R}_+$ since $p_0(h)\to 1$ and $p_1(h)\to 0$ as $h\downarrow 0$. From (3.3), we have $p_1(h)=\lambda h + o(h)$ and $p_0(h) = 1-p_1(h)-o(h)$. Substituting these in the preceding expression yields

$$(p_n(t+h)-p_n(t))/h = -\lambda p_n(t) + \lambda p_{n-1}(t) + o(h)/h.$$

Similar reasoning shows that $p_n(t)$ is left-continuous on $(0,\infty)$ and

$$(p_n(t)-p_n(t-h))/h = -\lambda p_n(t-h) + \lambda p_{n-1}(t-h) + o(h)/h.$$

Then letting $h\downarrow 0$, yields the differential equations

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \qquad n\geq 1.$$

In case $n=0$, a similar argument yields $p_0'(t) = -\lambda p_0(t)$.

To solve this family of differential-difference equations, with boundary conditions $p_n(0) = \mathbf{1}(n=0)$, first note that $p_0(t)=e^{-\lambda t}$ satisfies the last differential equation. Then using this function and induction on $n\geq 1$ it follows that $p_n(t) = (\lambda t)^n e^{-\lambda t}/n!$. This proves that $N(t)$ has a Poisson distribution with mean λt. Furthermore, by the same argument, one can show that $N(s,t]$ has a Poisson distribution with mean $\lambda(t-s)$ for any $s<t$ (in this case, one uses $p_n(t) = P\{N(s,t]=n\}$). This completes the proof that (c) implies (a).

3.3 Location of Points

Many applications of Poisson processes involve knowledge about the locations of their points. Statement (c) in Theorem 4 above suggests that the points of a Poisson process are located independently in a uniform sense on $\mathbb{R}_+$. This section gives a precise description of this property based on a multinomial characterization of Poisson processes.

We begin with an important property of Poisson processes.

Remark 5. If N is a Poisson process with rate λ, then the probability that it has a point at any fixed location t is 0 (i.e., $P\{N(\{t\}) = 1\} = 0$). This follows from Proposition 18 in Chapter 2 or from Proposition 30 below.

This remark implies that $N(a,b] \stackrel{d}{=} N(I)$, where I equals (a,b), $[a,b]$, or $[a,b)$, for $a < b$. Thus, Definition 1 is equivalent to the following one.

Definition 6. A simple point process N is a Poisson process with rate λ if and only if $N(I_1), \ldots, N(I_n)$ are independent, for disjoint finite intervals $I_1, \ldots, I_n$, and $N(I)$ has a Poisson distribution with mean $\lambda|I|$ for any finite interval I. Here $|I|$ denotes the length of I.

The next result is a characterization of a Poisson process involving a multinomial distribution (3.6) of the numbers of its points in disjoint intervals. Interestingly, (3.6) is independent of the rate λ. The analogous property for Poisson processes on general spaces is given in Theorem 28 and Example 27.

Theorem 7. *For a simple point process $N = \{N(t) : t \geq 0\}$ on $\mathbb{R}_+$ and $\lambda > 0$, the following statements are equivalent.*
(a) *N is a Poisson process with rate λ.*
(b) (Multinomial Property) *For any $t > 0$, the quantity $N(t)$ has a Poisson distribution with mean λt, and, for any disjoint intervals $I_1, \ldots, I_k$ in $[0,t]$, and nonnegative integers $n_1, \ldots, n_k$,*

$$P\{N(I_1) = n_1, \ldots, N(I_k) = n_k | N(t) = n\} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \quad (3.6)$$

where $n = n_1 + \cdots + n_k$ and $p_i = |I_i|/t$.

Proof. (a) $\Rightarrow$ (b). Suppose (a) holds. By Definition 6, N has independent Poisson increments over any disjoint intervals. Then letting $I_0 = [0,t] \backslash \cup_{i=1}^k I_i$ and $n_0 = 0$, the conditional probability in (3.6) is

$$\frac{P\{N(I_i) = n_i,\ 0 \leq i \leq k\}}{P\{N(t) = n\}} = \frac{\prod_{i=0}^k (\lambda |I_i|)^{n_i} e^{-\lambda |I_i|} / n_i!}{e^{-\lambda t} (\lambda t)^n / n!}.$$

This clearly reduces to the right-hand side of (3.6) since $\sum_{i=0}^k |I_i| = t$.

(b) $\Rightarrow$ (a). Suppose (b) holds. Fix a t and choose any $0 = t_0 < t_1 < \cdots < t_k = t$ and nonnegative integers $n_1, \ldots, n_k$ such that $n = n_1 + \cdots + n_k$. Define $I_i = (t_{i-1}, t_i]$ and $A_i = \{N(I_i) = n_i\}$. Then under the properties in (b),

$$P(\cap_{i=1}^k A_i) = P\{N(t)=n\}P\{\cap_{i=1}^k A_i | N(t)=n\}$$
$$= \prod_{i=1}^k \frac{[\lambda(t_i - t_{i-1})]^{n_i}}{n_i!} e^{-\lambda(t_i - t_{i-1})} = \prod_{i=1}^k P(A_i).$$

This proves that $N(I_1), \ldots, N(I_k)$ are independent, and $N(I_i)$ has a Poisson distribution with mean $\lambda(t_i - t_{i-1})$.

The proof of (a) will be complete upon showing that the increments $N(s_1, t_1], \ldots, N(s_k, t_k]$ are independent, for any $s_1 < t_1 < \cdots < s_k < t_k = t$. However, this independence follows because these increments are a subset of the increments $N(0, s_1], N(s_1, t_1], N(t_1, s_2], \ldots, N(s_k, t_k]$ over all the adjacent intervals, which are independent as we just proved.

A special case of (3.6) is the *binomial property*: For $I \subseteq [0, t]$ and $k \leq n$,

$$P\{N(I) = k | N(t) = n\} = \binom{n}{k} (|I|/t)^k (1 - |I|/t)^{n-k}.$$

The multinomial property also yields the joint conditional distribution of point locations in $[0, t]$ given $N(t) = n$, as shown in (3.7) below.

Theorem 8. (Order Statistic Property) *Suppose N is a Poisson process with rate λ. Then, for any disjoint intervals $I_1, \ldots, I_n$ in $[0, t]$,*

$$P\Big\{T_1 \in I_1, \ldots, T_n \in I_n \Big| N(t) = n\Big\} = \frac{n!}{t^n} \prod_{i=1}^n |I_i|. \tag{3.7}$$

Hence, the joint conditional density of $T_1, \ldots, T_n$ given $N(t) = n$ is

$$f_{T_1,\ldots,T_n}\Big(t_1, t_2, \ldots, t_n \Big| N(t) = n\Big) = \frac{n!}{t^n}, \qquad 0 < t_1 < \cdots < t_n < t. \tag{3.8}$$

The density (3.8) is the density of the order statistics of n independent uniformly distributed random variables on $[0, t]$ (see Proposition 10 below).

Proof. Expression (3.7) follows since the conditional probability in it equals $P\{N(I_i) = 1, 1 \leq i \leq n | N(t) = n\}$, which in turn equals the right-hand side of (3.7) by the multinomial property (3.6).

Next, note that (3.7), for $a_1 < b_1 < \cdots < a_n < b_n < t$, is

$$P\Big\{T_i \in (a_i, b_i], 1 \leq i \leq n \Big| N(t) = n\Big\} = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \frac{n!}{t^n} dt_1 \cdots dt_n.$$

Then the integrand $n!/t^n$ is the conditional density as asserted in (3.8).

Example 9. Marginal Distributions. From (3.7), (3.8) and Exercise 28,

$$P\{T_k \le s | N(t) = n\} = \sum_{j=k}^{n} \binom{n}{j} (s/t)^j (1 - s/t)^{n-j},$$

$$f_{T_k}(s|N(t) = n) = \frac{n!}{(k-1)!(n-k)!} (s/t)^{k-1} (1/t)(1 - s/t)^{n-k}, \quad 0 \le s \le t.$$

This conditional distribution of T_k is the same as that of tY, where Y has a beta distribution as shown in the Appendix with parameters $a = k$ and $b = n - k + 1$, and mean $a/(a + b)$. Taking advantage of this fact, it follows easily that

$$E[T_k|N(t) = n] = E[tY] = kt/(n + 1).$$

In particular, for a single point,

$$P\{T_1 \le s|N(t) = 1\} = s/t, \qquad 0 \le s \le t.$$

This is a uniform distribution on $[0, t]$.

We referred to (3.8) as the density of n order statistics of independent uniformly distributed random variables on $[0, t]$. This is justified by the following formula for the density of order statistics of a random sample with a general density.

Proposition 10. (Order Statistics) *Suppose $X_1, \ldots, X_n$ are independent continuous random variables with density f, and let $X_{(1)} < \cdots < X_{(n)}$ denote the quantities $X_1, \ldots, X_n$ in increasing order. These* order statistics *$X_{(1)}, \ldots, X_{(n)}$ have the joint density*

$$f(x_1, \ldots, x_n) = n! f(x_1) \cdots f(x_n), \qquad x_1 < \cdots < x_n. \tag{3.9}$$

Proof. Choose any $a_1 < b_1 < \cdots < a_n < b_n$, and let $I_i = (a_i, b_i]$, $1 \le i \le n$. Since $X_{(1)}, \ldots, X_{(n)}$ is equally likely to be any one of the $n!$ permutations of $X_1, \ldots, X_n$,

$$\begin{aligned} P\{X_{(i)} \in I_i,\ 1 \le i \le n\} &= n! P\{X_i \in I_i,\ 1 \le i \le n\} \\ &= n! \prod_{i=1}^{n} P\{X_i \in I_i\} \\ &= n! \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n. \end{aligned}$$

This proves the density formula (3.9).

3.4 Functions of Point Locations

Typical quantities of interest for a Poisson process N in a time interval $[0,t]$ are deterministic or random functions of the point locations $T_1,\ldots,T_{N(t)}$. A classic example is $\sum_{n=1}^{N(t)} f(T_n)$, where $f:\mathbb{R}_+\to\mathbb{R}$. This section shows how to analyze such functions in terms of random samples.

The following result is an immediate consequence of Theorem 8.

Corollary 11. (Order Statistic Tool) *Let N be a Poisson process with rate λ, and, for each $n\geq 1$, let h_n be a function from $\mathbb{R}_+^n$ to some Euclidean or more general space S, and let $h_0\in S$. Then, for $t>0$,*

$$h_{N(t)}(T_1,\ldots,T_{N(t)}) \stackrel{d}{=} h_\kappa(X_{(1)},\ldots,X_{(\kappa)}),$$

where $X_{(1)}<\cdots<X_{(n)}$ are the n order statistics associated with independent random variables $X_1,\ldots,X_n$ that are uniformly distributed on $[0,t]$ for each n, and κ is a Poisson random variable with mean λt, independent of the X_i. Furthermore, if each $h_n(x_1,\ldots,x_n)$ is symmetric (meaning it is the same for any permutation of $x_1,\ldots,x_n$), then

$$h_{N(t)}(T_1,\ldots,T_{N(t)}) \stackrel{d}{=} h_\kappa(X_1,\ldots,X_\kappa). \tag{3.10}$$

These expressions enable one to analyze a function of the random-length, "dependent" variables $T_1,\ldots,T_{N(t)}$ by the simpler mixed random sample $X_1,\ldots,X_\kappa$. The ideas here are related to the characterization of a Poisson process by mixed binomial or sample processes in Theorem 28 below.

As an example, for $f:\mathbb{R}_+\to\mathbb{R}$,

$$\sum_{n=1}^{N(t)} f(T_n) \stackrel{d}{=} \sum_{n=1}^{\kappa} f(X_n).$$

In this case, $h_n(x_1,\ldots,x_n)=\sum_{i=1}^n f(x_i)$ is symmetric. Here is another example involving random functions.

Proposition 12. (Random Sums) *Suppose N is a Poisson process with rate λ, and define*

$$Z(t)=\sum_{n=1}^{N(t)} f(T_n,Y_n), \tag{3.11}$$

where $Y_1,Y_2\ldots$ are i.i.d. random elements in a space S, independent of N, and $f:\mathbb{R}_+\times S\to\mathbb{R}$. Assume $\phi(\alpha,t)=E[e^{\alpha f(t,Y_1)}]$ exists for α in a neighborhood of 0 and $t\in\mathbb{R}$. Then the moment generating function of $Z(t)$ is

$$E[e^{\alpha Z(t)}]=e^{-\lambda t(1-g_t(\alpha))}, \tag{3.12}$$

where $g_t(\alpha)=t^{-1}\int_0^t \phi(\alpha,s)\,ds$. Hence,

$$E[Z(t)] = \lambda \int_0^t E[f(s, Y_1)]\, ds. \tag{3.13}$$

Proof. First note that

$$\begin{aligned} E[e^{\alpha Z(t)}] &= E\Big[E[e^{\alpha Z(t)} | N(s), s \le t]\Big] \\ &= E\Big[\prod_{n=1}^{N(t)} E[e^{\alpha f(T_n, Y_n)} | N(s), s \le t]\Big] \;=\; E\Big[\prod_{n=1}^{N(t)} \phi(\alpha, T_n)\Big]. \end{aligned}$$

Applying (3.10) with $h_n(x_1, \ldots, x_n) = \prod_{i=1}^n \phi(\alpha, x_i)$ to the last expression,

$$E[e^{\alpha Z(t)}] = E\Big[\prod_{n=1}^{\kappa} \phi(\alpha, X_n)\Big].$$

Then conditioning on κ, which is independent of the X_n's, we have

$$E[e^{\alpha Z(t)}] = E\Big[\prod_{n=1}^{\kappa} E[\phi(\alpha, X_n)]\Big].$$

Clearly $E[\phi(\alpha, X_n)] = g_t(\alpha)$, since X_n is uniformly distributed on $[0, t]$. Using this along with the independence of the X_n's and the well-known Poisson generating function $E[\beta^\kappa] = e^{-\lambda t(1-\beta)}$, we obtain

$$E[e^{\alpha Z(t)}] = E[g_t(\alpha)^\kappa] = e^{-\lambda t(1-g_t(\alpha))}.$$

This proves (3.12). In addition, (3.13) follows, since the derivative of (3.12) with respect to α at $\alpha = 0$ is $E[Z(t)] = t\lambda g_t'(0) e^{-\lambda t(1-g_t(0))}$, where $g_t(0) = 1$ and $g_t'(0) = \phi'(0) = t^{-1} \int_0^t E[f(x, Y_1)]\, dx$.

Example 13. Discounted Cash Flows. A special case of the sum (3.11) is

$$Z(t) = \sum_{n=1}^{N(t)} Y_n e^{-\gamma T_n},$$

where $\gamma > 0$. This is a standard model for discounted costs or revenues, where γ is a deterministic discount rate. For instance, suppose that sales of a product occur at times that form a Poisson process N with rate λ, and the amount of revenue from the nth sale is a random variable Y_n, independent of N. Then the total discounted revenue up to time t is given by $Z(t)$. By (3.12) and (3.13), the moment generating function and mean of $Z(t)$ are

$$\begin{aligned} E[e^{\alpha Z(t)}] &= \exp\Big\{ -\lambda t(1 - t^{-1} \int_0^t E[e^{\alpha e^{-\gamma x} Y_1}]\, dx)\Big\} \\ E[Z(t)] &= \lambda E[Y_1](1 - e^{-\gamma t})/\gamma. \end{aligned}$$

Example 14. Compound Poisson Process. Another example of (3.11) is $Z(t) = \sum_{n=1}^{N(t)} Y_n$. This is like the preceding discounted cash flow, but without discounting. Letting $\gamma \to 0$ in the preceding example, it follows that

$$E[e^{\alpha Z(t)}] = e^{-\lambda t(1-E[e^{\alpha Y_1}])},$$
$$E[Z(t)] = \lambda t E[Y_1].$$

This moment generating function of $Z(t)$ is that of a compound Poisson distribution with rate λt and distribution $F(y) = P\{Y_1 \le y\}$, which is

$$P\{Z(t) \le z\} = \sum_{n=0}^{\infty} e^{-\lambda t}(\lambda t)^n F^{n\star}(z)/n!, \qquad z \in \mathbb{R}.$$

The process $\{Z(t) : t \ge 0\}$ is called a *compound Poisson process*; further properties of it are in Section 3.15.

3.5 Poisson Processes on General Spaces

Applications of classical Poisson processes on $\mathbb{R}_+$ typically require knowledge of Poisson processes on Euclidean and general spaces. In addition, there are many spatial systems of points that can be modeled by Poisson processes. Accordingly, the rest of this chapter will focus on basics of point processes and Poisson processes on general spaces. This section introduces the terminology we will use.

A "point process" is a counting process that represents a random set of points in a space. The usual spaces are the real line, the plane, the multidimensional Euclidean space $\mathbb{R}^d$, or, more generally, a complete, separable metric space (a Polish space). Following the standard convention, we will discuss point processes on a polish space S. The exposition will be understandable by thinking of S as an Euclidean space. We let $\mathcal{S}$ denote the family of Borel sets of S (see the Appendix), and let $\hat{\mathcal{S}}$ denote the family of *bounded* Borel sets (a set is bounded if it is contained in a compact set). We refer to S simply as a *space*, and denote other spaces of this type by S', $\tilde{S}$, etc.

We begin with an informal description of a point process. A random set of points in S is a countable set of S-valued random elements $\{X_n\}$ such that only a finite number of the points are in any bounded set. Then

$$N(B) = \sum_n \delta_{X_n}(B), \quad B \in \mathcal{S},$$

denotes the number of points in B, where $\delta_x(B) = \mathbf{1}(x \in B)$ (a Dirac measure with unit mass at x). This counting measure N, as we will define below, is a *point process* on S with point locations $\{X_n\}$.

Note that N takes values in the set $\mathbb{M}$ of all counting measures ν on $(S, \mathcal{S})$ that are *locally finite* ($\nu(B) < \infty$, for bounded sets B). Each measure in $\mathbb{M}$ has the form

$$\nu(B) = \sum_n \delta_{x_n}(B), \quad B \in \mathcal{S}, \tag{3.14}$$

where $\{x_1, \ldots, x_k\}$ is its associated set of points, for $0 \le k \le \infty$. There may be more than one point at a location, and the order of the subscripts on the locations is invariant under permutations. The set $\mathbb{M}$ is endowed with the σ-field $\mathcal{M}$ generated by the sets $\{\nu \in \mathbb{M} : \nu(B) = n\}$, for $B \in \mathcal{S}$ and $n \ge 0$.

We are now ready for a formal definition.

Definition 15. A *point process* on a space S is a measurable map N from a probability space $(\Omega, \mathcal{F}, P)$ to the space $(\mathbb{M}, \mathcal{M})$. The quantity $N(B)$ is the number of points in the set $B \in \mathcal{S}$. By the formulation (3.14),

$$N(B) = \sum_n \delta_{X_n}(B), \quad B \in \mathcal{S}, \tag{3.15}$$

where the X_n denote the *locations* of the points of N.

For the following discussion, assume that N is a point process on the space S. Technical properties of the spaces S and $\mathbb{M}$ are not used explicitly in the sequel. One can simply think of N as a counting process on $S = \mathbb{R}^d$ that is locally finite ($N(B) < \infty$ a.s. for bounded sets B). The *probability distribution* of the point process N (i.e., $P\{N \in \cdot\}$) is determined by its finite-dimensional distributions

$$P\{N(B_1) = n_1, \ldots, N(B_k) = n_k\}, \qquad B_1, \ldots, B_k \in \hat{\mathcal{S}}. \tag{3.16}$$

In other words, two point processes N and N' on S are equal in distribution, denoted by $N \stackrel{d}{=} N'$, if their finite-dimensional distributions are equal:

$$(N(B_1) \ldots, N(B_k)) \stackrel{d}{=} (N'(B_1) \ldots, N'(B_k)), \qquad B_1, \ldots, B_k \in \hat{\mathcal{S}}.$$

In constructing a point process, it suffices to define the probabilities (3.16) on sets B_i that generate $\mathcal{S}$. When $S = \mathbb{R}^d$, "rectangles" of the form[1] $(a, b]$ generate $\mathcal{S}$.

The *intensity measure* (or mean measure) of the point process N is

$$\mu(B) = E[N(B)], \qquad B \in \mathcal{S}.$$

Note that $\mu(B)$ may be infinite, even if B is bounded. When $S = \mathbb{R}^d$, the intensity is sometimes of the form $\mu(B) = \int_B \lambda(x)dx$, where $\lambda(x)$ is the rate of N at the location x and dx denotes the Lebesgue measure. We call $\lambda(x)$ the location-dependent *rate function* of N.

[1] Here $x = (x_1, \ldots, x_d)$ is a typical vector in $\mathbb{R}^d$ and $(a, b] = \{x \in \mathbb{R}^d : a_i < x_i \le b_i,\ 1 \le i \le d\}$ is a half-open interval in $\mathbb{R}^d$.

Our main focus hereafter will be on Poisson point processes.

Definition 16. A point process N on a space S is a *Poisson process with intensity measure* μ that is locally finite if the following conditions are satisfied.
• N has *independent increments*: The quantities $N(B_1), \ldots, N(B_n)$ are independent for disjoint sets $B_1, \ldots, B_n$ in $\hat{\mathcal{S}}$.
• For each $B \in \hat{\mathcal{S}}$, the quantity $N(B)$ is a Poisson random variable with mean $\mu(B)$.

This definition uses the fact that $N(B) = 0$ a.s. when $\mu(B) = 0$. Note that the number of points $N(\{x\})$ exactly at x has a Poisson distribution with mean $\mu(\{x\})$; so $N(\{x\}) = 0$ a.s. when $\mu(\{x\}) = 0$. From the definition, it follows that the finite-dimensional distributions of a Poisson process are uniquely determined by its intensity measure, and vice versa. That is, two Poisson processes N and N' on S are equal in distribution if and only if their intensity measures are equal.

Do Poisson processes exist? In other words, does there exist a point process on a probability space that satisfies the properties in Definition 16? We will establish the existence later when we show in Theorem 29 below that a Poisson process can be characterized by independent random elements, which do exist.

The family of Poisson processes on $\mathbb{R}_+$ contain the classical ones.

Example 17. Poisson Processes on $\mathbb{R}_+$. A Poisson process N on $\mathbb{R}_+$ (or on $\mathbb{R}$) with intensity measure μ is sometimes called a *non-homogeneous* Poisson process. We denote its point locations (as we have been doing) by $0 < T_1 \leq T_2 \leq \ldots$ instead of X_n, and call them "times" when appropriate. Here there may be more than one point at a location. In this setting, $N(B) = \sum_n \delta_{T_n}(B)$ has a Poisson distribution with mean $\mu(B)$. We also write $N(t) = N(0, t]$, for $t > 0$, and $N(a, b] = N((a, b])$ and $\mu(a, b] = \mu((a, b])$, for $a < b$. When $\mu(B) = \int_B \lambda(t)dt$, we say N is a Poisson process with *rate function* $\lambda(t)$. In case $\mu(t) = E[N(t)] = \lambda t$, for some $\lambda > 0$, then N is a classical Poisson process with rate λ, consistent with Definition 1. It is sometimes called a *homogeneous* or stationary Poisson process with rate λ.

Results in the preceding sections for homogeneous Poisson processes have obvious analogues for nonhomogeneous processes. For instance, N satisfies the multinomial property (3.6) with $p_i = \mu(B_i)/\mu(0, t]$, and Exercise 29 describes its order statistic property.

3.6 Integrals and Laplace Functionals of Poisson Processes

Laplace transforms are useful for identifying distributions of nonnegative random variables and for establishing convergence in distribution of random

variables. The analogous tool for point processes is a Laplace functional. This section covers a few properties of Laplace functionals and related integrals of point processes. These are preliminaries needed to establish the existence of Poisson processes, the topic of the next section. The use of Laplace transforms and functionals for establishing the convergence of random variables and point processes are covered later in Sections 3.16 and 3.17.

We begin with a little review. Recall that the Laplace transform of a nonnegative random variable X with distribution F is

$$\hat{F}_X(\alpha) = E[e^{-\alpha X}] = \int_{\mathbb{R}_+} e^{-\alpha x} dF(x), \qquad \alpha \geq 0.$$

This function uniquely determines the distribution of X in that $X \stackrel{d}{=} Y$ if and only if $\hat{F}_X(\cdot) = \hat{F}_Y(\cdot)$. For instance, the Laplace transform of a Poisson random variable X with mean λ is

$$\hat{F}_X(\alpha) = \sum_{n=0}^{\infty} e^{-\alpha n} e^{-\lambda} \lambda^n / n! = e^{-\lambda(1-e^{-\alpha})}.$$

Now, if Y is a nonnegative integer-valued random variable with $E[e^{-\alpha Y}] = e^{-\lambda(1-e^{-\alpha})}$, then Y has a Poisson distribution with mean λ. Here is an example for sums.

Example 18. Sums of Independent Poisson Random Variables. Let $Y_1, \ldots, Y_n$ be independent Poisson random variables with respective means $\mu_1, \ldots, \mu_n$. Then $\sum_{i=1}^n Y_i$ has a Poisson distribution with mean $\mu = \sum_{i=1}^n \mu_i$ (which we assume is finite when $n = \infty$). To see this result, note that by the independence of the Y_i and the form of their Laplace transforms,

$$E[e^{-\alpha \sum_{i=1}^n Y_i}] = \prod_{i=1}^n E[e^{-\alpha Y_i}] = e^{-\mu(1-e^{-\alpha})}.$$

We recognize this as being the Laplace transform of a Poisson distribution with mean μ, and so $\sum_{i=1}^n Y_i$ has this distribution.

The rest of this section covers analogous properties for Laplace functionals of point processes. Consider a point process $N = \sum_n \delta_{X_n}$ on a space S. The "integral" of a function $f : S \to \mathbb{R}_+$ with respect to N is simply the sum

$$Nf = \int_S f(x) N(dx) = \sum_n f(X_n),$$

provided it is finite. It is finite when f has a *compact support*, meaning that $\{x : f(x) > 0\}$ is contained in a compact set. Similarly, the integral of $f : S \to \mathbb{R}_+$ with respect to a measure μ will be denoted by

$$\mu f = \int_S f(x)\mu(dx).$$

We will often use integrals of functions f in the set $C_K^+(S)$ of all continuous functions $f : S \to \mathbb{R}_+$ with compact support.

Definition 19. The *Laplace functional* of the point process N is

$$E[e^{-Nf}] = E\Big[\exp\{-\int_S f(x)N(dx)\}\Big], \qquad f : S \to \mathbb{R}_+.$$

The function f is a "variable" of this expectation (just as the parameter α is a variable in a Laplace transform $E[e^{-\alpha X}]$).

The following result contains the basic property that the Laplace functional of a point process uniquely determines its distribution (the proof is in [61]). It also justifies that a Laplace functional is uniquely defined on the set $C_K^+(S)$ viewed as "test" functions.

Theorem 20. *For point processes N and N' on S, each one of the following statements is equivalent to $N \stackrel{d}{=} N'$.*
(a) $Nf \stackrel{d}{=} N'f, \qquad f \in C_K^+(S).$
(b) $E[e^{-Nf}] = E[e^{-N'f}], \quad f \in C_K^+(S).$

Laplace functionals are often more convenient to use than finite-dimensional distributions in deriving the distribution of a point process constructed as a function of random variables or point processes. A standard approach for establishing that a point process is Poisson is to verify that its Laplace functional has the following form; this also yields its intensity measure.

Theorem 21. (Poisson Laplace Functional) *For a Poisson process N on S with intensity measure μ, and $f : S \to \mathbb{R}_+$,*

$$E[e^{-Nf}] = \exp[-\int_S (1 - e^{-f(x)})\mu(dx)].$$

Proof. First consider the simple function $f(x) = \sum_{i=1}^k a_i \mathbf{1}(x \in B_i)$, for some nonnegative $a_1, \ldots, a_k$ and disjoint $B_1, \ldots, B_k$ in $\hat{\mathcal{S}}$. Then

$$Nf = \sum_{i=1}^k a_i \int_S \mathbf{1}(x \in B_i)N(dx) = \sum_{i=1}^k a_i N(B_i).$$

Using this and the independence of the $N(B_i)$, we have

$$\begin{aligned} E[e^{-Nf}] = \prod_{i=1}^k E[e^{-a_i N(B_i)}] \;&=\; \exp[-\sum_{i=1}^k \mu(B_i)(1 - e^{-a_i})] \\ &= \exp[-\int_S (1 - e^{-f(x)})\mu(dx)]. \end{aligned}$$

Next, for any continuous $f : S \to \mathbb{R}_+$, there exist simple functions $f_n \uparrow f$ (e.g., $f_n(x) = n \wedge (\lfloor 2^n f(x) \rfloor / 2^n)$). Then by the monotone convergence theorem (see the Appendix) and the first part of this proof,

$$\begin{aligned} E[e^{-Nf}] &= \lim_{n\to\infty} E[e^{-Nf_n}] = \lim_{n\to\infty} \exp[-\int_S (1 - e^{-f_n(x)})\mu(dx)] \\ &= \exp[-\int_S (1 - e^{-f(x)})\mu(dx)]. \end{aligned}$$

This completes the proof.

Recall that Example 18 uses Laplace transforms to prove that a sum of independent Poisson random variables is Poisson. Here is an analogous result for a sum of Poisson processes.

Theorem 22. (Sums of Independent Poisson Processes) *Let $N_1, \ldots, N_n$ denote independent Poisson processes on S with respective intensity measures $\mu_1, \ldots, \mu_n$. Then their* sum (or superposition) *$N = \sum_{i=1}^n N_i$ is a Poisson process with intensity measure $\mu = \sum_{i=1}^n \mu_i$. This is also true for $n = \infty$ provided μ is locally finite.*

Proof. One can prove this, as suggested in Exercise 14, by verifying that N satisfies the defining properties of a Poisson process. Another approach, using Laplace functionals and Theorem 21, is to verify that

$$E[e^{-Nf}] = e^{-\mu h}, \qquad f \in C_K^+(S),$$

where $h(x) = 1 - e^{-f(x)}$. But this follows since by the independence of the N_i and the form of their Laplace functionals in Theorem 21,

$$E[e^{-Nf}] = \prod_{i=1}^n E[e^{-N_i f}] = \prod_{i=1}^n e^{-\mu_i h} = e^{-\mu h}.$$

Example 23. A company that produces a household cleaning fluid has a bottle-filling production line that occasionally has to stop for repair due to imperfections in the bottles or due to worker errors. There are four types of line stoppages: (1) minor stop (under 30 minutes) due to bottle imperfection, (2) major stop (over 30 minutes) due to bottle imperfection, (3) minor stop due to worker error, and (4) major stop due to worker error. These four types of stoppages occur according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_4$. Then by Theorem 22, line stops due to any of these causes occur according to a Poisson process with rate $\lambda_1 + \cdots + \lambda_4$. Similarly, minor stops occur according to a Poisson process with rate $\lambda_1 + \lambda_3$, and major stops occur according to a Poisson process with rate $\lambda_2 + \lambda_4$. A model like this was used in a study of a bottling plant by Georgia Tech students.

We end this section with more insight on integrals $Nf = \sum_n f(X_n)$ with respect to a point process N. Expression (3.17) below says the mean of such an integral equals the corresponding integral with respect to the intensity measure. Theorem 22 in Chapter 2 for renewal processes is a special case. The variance of Nf has the nice form (3.18), when N is Poisson.

Theorem 24. *Let $N = \sum_n \delta_{X_n}$ be a point process on S with intensity measure μ. For any $f : S \to \mathbb{R}$,*

$$E\Big[\sum_n f(X_n)\Big] = \int_S f(x)\mu(dx), \tag{3.17}$$

provided the integral exists. That is, $E[Nf] = \mu f$. If in addition, N is a Poisson process, then

$$Var\Big[\sum_n f(X_n)\Big] = \int_S f(x)^2\mu(dx), \tag{3.18}$$

provided the integral exists. That is, $Var[Nf] = \mu f^2$.

Proof. The proof of (3.17) is similar to that of Theorem 22 or Theorem 21. Namely, first one shows $E[Nf] = \mu f$ is true when f is a simple function, and then monotone convergence yields the equality for general f, which is a monotone limit of simple functions.

To prove (3.18), note that by Theorem 21, we have

$$E[e^{-\alpha Nf}] = e^{-h(\alpha)}, \tag{3.19}$$

where $h(\alpha) = \int_S (1 - e^{-\alpha f(x)})\mu(dx)$. The derivative of this expression at $\alpha = 0$, yields

$$E[Nf] = h'(0)e^{-h(0)} = \mu f.$$

Furthermore, taking the second derivative of (3.19) at $\alpha = 0$, and using $h(0) = 1$ and $h'(0) = E[Nf]$, we obtain

$$\begin{aligned} E[(Nf)^2] &= \lim_{\alpha \downarrow 0} \Big[(h'(\alpha))^2 e^{-h(\alpha)} - h''(\alpha)e^{-h(\alpha)}\Big] \\ &= \lim_{\alpha \downarrow 0} \Big[\int_S f(x)^2 e^{-\alpha f(x)}\mu(dx) + (h'(\alpha))^2\Big] \\ &= \int_S f(x)^2\mu(dx) + (E[Nf])^2. \end{aligned}$$

This proves (3.18).

Example 25. Suppose N is a Poisson process on $\mathbb{R}_+$ with intensity measure $\mu(B) = \int_B e^{-at}dt$, and consider

$$Z = \int_0^u ce^{-bt} N(dt) = \sum_{n=1}^{N(u)} ce^{-bT_n}.$$

Here a, b, c, u are nonnegative constants. By Theorem 24,

$$E[Z] = c\int_0^u e^{-bt}e^{-at}dt = \frac{c[1-e^{-u(a+b)}]}{a+b},$$
$$\mathrm{Var}[Z] = c^2\int_0^u e^{-2bt}e^{-at}dt = \frac{c^2[1-e^{-u(a+2b)}]}{a+2b}.$$

3.7 Poisson Processes as Sample Processes

Section 3.4 showed that a Poisson process on $\mathbb{R}_+$ can be characterized by an order-statistic property of its point locations. This section presents an analogous characterization of a Poisson process on a general space in terms of sample processes. Using this result, we establish the existence of Poisson processes. Sample processes are also useful as models by themselves as well as building blocks for identifying Poisson processes.

A sample process is a fundamental point process defined as follows. Suppose that $X_1, X_2, \ldots$ are i.i.d. random elements in the space S with distribution $F(B) = P\{X_1 \in B\}$. The point process

$$N = \sum_{i=1}^{n} \delta_{X_i}$$

on S is a *sample process* for n samples from F.

Clearly, the number of samples $N(B)$ in a set B has a binomial distribution with parameters n and $\mu(B)$ (N is also called a *binomial point process*). In what follows, we consider such a sample process in which the sample size is a random variable that is independent of the samples; the resulting process is called a *mixed sample process*.

We first consider Poisson processes with finite intensities.

Theorem 26. *Suppose N is a point process on S with an intensity measure μ such that $0 < \mu(S) < \infty$. The N is a Poisson process if and only if*

$$N \stackrel{d}{=} \tilde{N} = \sum_{i=1}^{\kappa} \delta_{\tilde{X}_i},$$

where $\tilde{N}$ is a mixed sample process for κ samples from the distribution $\mu(\cdot)/\mu(S)$, and κ has a Poisson distribution with mean $\mu(S)$.

Proof. To prove the assertion, it suffices to verify that the Laplace functional of $\tilde{N}$ is the same as that in Theorem 21 for a Poisson process with intensity measure μ. That is,

$$E[e^{-\tilde{N}f}] = \exp[-\int_S (1 - e^{-f(x)})\mu(dx)], \qquad f \in C_K^+(S).$$

But this follows, since using the generating function $E[z^\kappa] = e^{-\mu(S)(1-z)}$ and the fact that $\tilde{X}_1$ has the distribution $\mu(\cdot)/\mu(S)$,

$$\begin{aligned} E[e^{-\tilde{N}f}] &= E\Big[E[e^{-\sum_{i=1}^{\kappa} f(\tilde{X}_i)}|\kappa]\Big] = E\Big[\Big(E[e^{-f(\tilde{X}_1)}]\Big)^{\kappa}\Big] \\ &= \exp[-\mu(S)(1 - E[e^{-f(\tilde{X}_1)}])] \\ &= \exp[-\int_S (1 - e^{-f(x)})\mu(dx)]. \end{aligned}$$

Example 27. Fires occur in a region S of a city at locations $X_1, X_2, \ldots$ that are independent with distribution F. Then the spatial locations of n fires in S is given by the sample process $N = \sum_{i=1}^n \delta_{X_i}$. In particular, the number of fires in a region $B \in \mathcal{S}$ has a binomial distribution with parameters n and $F(B)$. Furthermore, the numbers of fires in $B_1, \ldots, B_k$ in $\mathcal{S}$ that form a partition of S, have the multinomial distribution, for $n = n_1 + \cdots + n_k$,

$$P\{N(B_1) = n_1, \ldots, N(B_k) = n_k\} = \frac{n!}{n_1! \cdots n_k!} F(B_1)^{n_1} \cdots F(B_k)^{n_k}.$$

Next, suppose the number of fires in a year is a Poisson random variable κ with mean λ that is independent of the fire locations. Then the fires in S are represented by $\tilde{N} = \sum_{i=1}^{\kappa} \delta_{X_i}$, which is Poisson process with intensity measure $\lambda F(\cdot)$ by Theorem 26. In particular, the number of fires $\tilde{N}(B)$ in B has a Poisson distribution with mean $\lambda F(B)$.

Theorem 26 has the following extension that says any Poisson process on a bounded set is equal to a sample process on that set.

Theorem 28. *A point process N on S with a locally finite intensity measure μ is a Poisson process if and only if, for each $B \in \hat{\mathcal{S}}$ with $\mu(B) > 0$,*

$$N(\cdot \cap B) \stackrel{d}{=} \tilde{N}(\cdot),$$

where $\tilde{N}$ is a mixed sample process of κ samples from the probability measure $\mu(\cdot \cap B)/\mu(B)$, and κ has a Poisson distribution with mean $\mu(B)$.

Proof. This follows by applying Theorem 26 to each bounded set B, and using the fact that the distribution of a Poisson process is determined (via Laplace functionals) by its distribution on bounded sets (the supports of the functions in the Laplace functional).

The preceding characterization of a Poisson process yields the following result.

Theorem 29. (Existence of Poisson Processes) *For any locally finite measure μ on S, there exists a Poisson process N on S with intensity measure μ.*

Proof. First note that a sample process for a random-sized sample exists since it is a function of an infinite collection of independent random variables, which exist by Corollary 6 in the Appendix (i.e., one can construct a probability space and the independent random variables on it). Then a Poisson process N with a "finite" intensity measure μ exists since it is a sample process by Theorem 26.

Next, consider the case when μ is infinite. Choose bounded sets $B_1, B_2, \ldots$ in $\mathcal{S}$ that partition S such that $\mu(B_n) < \infty$. By the preceding part of the proof, there exists a Poisson process N_n on S, for each n, with intensity $\mu_n(\cdot) = \mu(\cdot \cap B_n)$. By Theorem 6 in the Appendix, we can define these N_n on a common probability space so that they are independent. Then define $N = \sum_n N_n$. By Theorem 22, N is a Poisson process on S with intensity $\sum_n \mu_n = \mu$.

We end this section with a criterion for a Poisson process to be simple. A point process N on S is *simple* if $P\{N(\{x\}) \leq 1, x \in S\} = 1$ (i.e., its point locations are distinct).

Proposition 30. *A Poisson process N with intensity measure μ is simple if and only if $\mu(\{x\}) = 0$, $x \in S$. Hence any Poisson process on an Euclidean space is simple if its intensity has the form $\mu(B) = \int_B \lambda(x)dx$, for some rate function $\lambda(x)$.*

Proof. In light of Theorem 28, it suffices to prove this when μ is finite. In this case, $N \stackrel{d}{=} \tilde{N}$, where $\tilde{N}$ is a mixed sample process as in Theorem 26. Now

$$P(N \text{ is simple}) = P(\tilde{N} \text{ is simple}) = \sum_{n=2}^{\infty} P(D_n) P\{\kappa = n\},$$

where $D_n = \{\tilde{X}_1, \ldots, \tilde{X}_n \text{ are distinct}\}$. Then N is simple if and only if $P(D_n) = 1$, for each $n \geq 2$. The latter statement is true, by Exercise 39, if and only if $\mu(\{x\})/\mu(S) = 0$, $x \in S$. Hence N is simple if and only if $\mu(\{x\}) = 0$, $x \in S$.

3.8 Deterministic Transformations of Poisson Processes

As we will see, many point processes involving complex phenomena or systems can be represented by functions of Poisson processes. In these settings,

a Poisson process is typically the basic data that defines or initializes a system, and various characteristics of the system are deterministic or random functions of the Poisson process. In particular, if part of a system's information is a point process (e.g., production completion times, or a random function of the original Poisson points), it is natural to know whether that point process is a Poisson process.

A basic issue in this regard is: If the point locations of a Poisson processes are mapped to some space by a deterministic or random mapping, then do the resulting new points also form a Poisson Process? This issue in a variety of contexts is the underlying theme for the next six sections.

We begin in this section by considering deterministic maps such as the following one.

Example 31. Suppose N is a Poisson process in the nonnegative quadrant $S = \mathbb{R}_+^2$ of the plane with intensity measure μ. Let $N'(r)$ denote the number of points of N within a distance r from the origin. We can represent N' as a mapping of N in which a point (x, y) of N is mapped to its distance $g(x, y) = \sqrt{x^2 + y^2}$ from the origin. Then,

$$N'(r) = N(\{(x, y) \in S : \, g(x, y) \le r\}).$$

In this example, N' simply records some characteristic (the distance to the origin) of each point in the so-called data N.

By Theorem 32 below, N' is a Poisson process on $\mathbb{R}_+$ with mean measure $E[N'(r)] = \mu(\{(x, y) \in S : \, g(x, y) \le r\})$. This mean is clearly finite for each r. In case the Poisson process N is homogeneous with a constant rate λ, then $E[N'(r)] = \lambda\pi r^2$.

We will now show that any general mapping of a Poisson process, such as the one above, results in a new Poisson process. Suppose N is a Poisson process on S with intensity measure μ. Consider a transformation of N in which its points in S are mapped to a space S' (possibly S) by the rule that a point of N located at $x \in S$ is mapped to the location $g(x) \in S'$, where $g : S \to S'$. Then the number of points mapped into $B \in \mathcal{S}'$ is

$$N'(B) = \sum_n \delta_{g(X_n)}(B), \qquad B \in \mathcal{S}'.$$

This N' is a point process on S', provided it is locally finite. In this case, we say that N' is a *transformation of N under the map g.*

A more complete representation of this transformation is given by the point process M on the product space $S \times S'$ defined[2] by

[2] To define a random measure M on a product space, it suffices to define it on product sets $A \times B$ (this also highlights the sets separately). For the case at hand, $M(A \times B) = \sum_n \delta_{(X_n, g(X_n))}(A \times B)$ automatically implies $M(C) = \sum_n \delta_{(X_n, g(X_n))(C)}$, for $C \in \mathcal{S} \times \mathcal{S}'$. Mean measures $E[M(A \times B)]$ are also defined on product sets.

$$M(A \times B) = \sum_n \delta_{(X_n, g(X_n))}(A \times B), \qquad A \in \mathcal{S},\ B \in \mathcal{S}'.$$

This is the number of points of N in $A \in \mathcal{S}$ that are mapped into $B \in \mathcal{S}'$. So M contains the original process $N(\cdot) = M(\cdot \times S')$ as well as the transformed process $N'(\cdot) = M(S \times \cdot)$. The M is an example of a "marked" point process, where $g(X_n)$ is a "mark" associated with X_n. General marked point processes are discussed in the next sections.

A better understanding of the processes M and N' is provided by the inverse of g, which is

$$g^{-1}(B) = \{x \in S : g(x) \in B\}, \qquad B \in \mathcal{S}'.$$

Namely, using $\delta_{g(X_n)}(B) = \delta_{X_n}(g^{-1}(B))$, it follows that, for $A \in S$, $B \in \mathcal{S}'$,

$$M(A \times B) = N(A \cap g^{-1}(B)), \qquad N'(B) = N(g^{-1}(B)). \tag{3.20}$$

In other words, M and the transformed process N' behave like the original process N on part of its space. Because of this, M and N' inherit the Poisson property of N as follows.

Theorem 32. *The marked point process M defined by (3.20) is a Poisson process with*

$$E[M(A \times B)] = \mu(A \cap g^{-1}(B)), \quad A \in \mathcal{S},\ B \in \mathcal{S}'. \tag{3.21}$$

Hence the transformed process N' is a Poisson process on S' with intensity $E[N'(B)] = \mu(g^{-1}(B))$, $B \in \mathcal{S}'$, provided this measure is locally finite.

Proof. This result is a special case of Theorem 36 below, which is proved by showing that the Laplace transform of M is the same as that for a Poisson process with intensity (3.21).

Here is another proof that illustrates the sample-process characterization of a Poisson process. For the case $\mu(S) < \infty$, we have

$$N \stackrel{d}{=} \sum_{m=1}^{\kappa} \delta_{\tilde{X}_n}, \qquad M \stackrel{d}{=} \sum_{m=1}^{\kappa} \delta_{(\tilde{X}_n, g(\tilde{X}_n))}.$$

The first sum is a sample process of κ samples from $\mu(\cdot)/\mu(S)$ and κ has a Poisson distribution with mean $\mu(S)$. The second sum, due to the definition of M and the form of N, is a sample process on $S \times S'$ of κ samples from the distribution

$$F(A \times B) = P\{\tilde{X}_1 \in A, g(\tilde{X}_1) \in B\} = \mu(A \cap g^{-1}(B))/\mu(S).$$

Therefore, M is a Poisson process with intensity measure given by (3.21). This statement is also true for the case $\mu(S) = \infty$ by Theorem 28, since the preceding argument also applies to each bounded set in $S \times S'$.

In addition, since M is Poisson and $N'(\cdot) = M(S \times \cdot)$ is M on part of its space, it follows that N' is Poisson with intensity $E[M(S \times B)] = \mu(g^{-1}(B))$, when this measure is locally finite.

A basic property of a Poisson process N on a product space $S_1 \times S_2$ is that the projection $N(S_1 \times \cdot)$ on S_2 is also a Poisson process provided its intensity $E[N(S_1 \times \cdot)]$ is locally finite. This follows immediately from the definition of a Poisson process. Here is an extension of this fact.

Example 33. Projections of a Poisson Process. Let $N = \sum_n \delta_{\mathbf{X}_n}$ denote a Poisson process on $S \subseteq S_1 \times \cdots S_m$, with intensity μ, where $\mathbf{X}_n = (X_{1n}, \ldots, X_{mn})$. Let $N_i = \sum_n \delta_{X_{in}}$ denote the projection of N on the subspace $S_i = \{x_i : \mathbf{x} \in S\}$, and let $M_i = \sum_n \delta_{(\mathbf{X}_n, X_{in})}$ be the marked point process that describes the points in the domain and range of the mapping of N by the projection map $g_i(\mathbf{x}) = x_i$.

Then M_i is Poisson with intensity $\mu_i(A \times B) = \mu\{\mathbf{x} \in A : x_i \in B\}$ by Theorem 32. Hence $N_i(\cdot) = M_i(S \times \cdot)$ is a Poisson process with intensity $E[N_i(B)] = \mu\{\mathbf{x} \in S : x_i \in B\}$, provided this intensity is locally finite.

For instance, suppose N is a homogeneous Poisson process on $\mathbb{R}_+^m$ with rate λ. Then $N_i(0, b] = \infty$ a.s., but M_i still gives insights on the projection.

Next, consider the more general projection $N_I = \sum_n \delta_{g_I(\mathbf{X}_n)}$ on the space $S_I = \{g_I(\mathbf{x}) : \mathbf{x} \in S\}$, where $g_I(\mathbf{x}) = (x_i : i \in I)$, for $I \subseteq \{1, \ldots, m\}$. The related process $M_I = \sum_n \delta_{(\mathbf{X}_n, g_I(\mathbf{X}_n))}$ on $S \times S_I$ is Poisson by Theorem 32. Hence $N_I(\cdot) = M(S \times \cdot)$ is Poisson with $E[N_I(B)] = \mu\{\mathbf{x} \in S : g_I(\mathbf{x}) \in B\}$, provided this is locally finite.

Example 34. Let $N = \sum_n \delta_{(X_n, Y_n)}$ denote a Poisson process on the unit disc S in $\mathbb{R}^2$ with rate function $\lambda(x, y)$. Consider the projection of N on the interval $S' = [-1, 1]$, which is described by the process $N' = \sum_n \delta_{X_n}$ on S'. By Example 33, N' is Poisson with

$$E[N'(a, b]] = \int_a^b \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \lambda(x, y)\, dy\, dx.$$

More generally, projections on $S' = [-1, 1]$ from points located in sets like $A_u = \{(x, y) \in S : y \geq u\}$, $u \in (0, 1]$ are described by $M = \sum_n \delta_{((X_n, Y_n), X_n)}$ on $S \times S'$, which is Poisson with

$$E[M(A_u \times (a, b])] = \int_a^b \int_u^1 \lambda(x, y)\, dy\, dx.$$

Next, consider the transformation $\overline{N}$ of N under which a point in the unit disc S is mapped to the closest point on the unit circle C. To define $\overline{N}$, we represent a point in S by its polar coordinates (r, θ), where $(x, y) = (r\cos\theta, r\sin\theta)$, and view $N = \sum_n \delta_{(R_n, \Theta_n)}$ as a Poisson process on $S = \{(r, \theta) \in [0, 1] \times [0, 2\pi)\}$ with rate function $\lambda(r\cos\theta, r\sin\theta)$. The unit circle can be expressed as $C = [0, 2\pi)$, since each point on it has the form $(1, \theta)$.

The transformation under consideration maps a point at (r, θ) to $(1, \theta)$ (i.e., to $\theta \in C$), and so the transformed process is $\overline{N} = \sum_n \delta_{\Theta_n}$, which is simply the projection of N on the coordinate set C. Therefore $\overline{N}$ is Poisson with

$$E[\overline{N}(B)] = \int_B \int_0^1 \lambda(r\cos\theta, r\sin\theta) dr\, d\theta.$$

See Exercise 35 for more details on these processes.

3.9 Marked and Space-Time Poisson Processes

The last section showed how a deterministic transformation of a Poisson process can be represented by a marked Poisson process. We now extend this idea to random transformations that are represented by marked Poisson processes. An important class of marked Poisson processes are space-time Poisson processes.

The focus of this section is on a transformation of a Poisson process on a space S in which each of its points is independently assigned a random mark in a space S' depending only on the particular point location. The distributions of the marks will be determined by probability kernels.

A mark assigned to a point at $x \in S$, will take a value in a set $B \in \mathcal{S}'$ according to a *probability kernel* $p(x, B)$ from $S \to S'$. Such a kernel is a function $p : S \times \mathcal{S}' \to [0, 1]$ such that $p(\cdot\, B)$ is a measurable function on S and $p(x, \cdot)$ is a probability measure on S'. Our interest will be in modeling the initial points as well as the marks by a marked point process on $S \times S'$. The formal definition is as follows.

Definition 35. Let $N = \sum_n \delta_{X_n}$ be a Poisson process on S with intensity μ. Let $M = \sum_n \delta_{(X_n, Y_n)}$ be a point process on $S \times S'$ such that

$$P\{Y_n \in B | N\} = p(X_n, B), \qquad B \in \mathcal{S}',\ n \le N(S),$$

where $p(x, B)$ is a probability kernel from S to S'. The Y_n are *p-marks of the* X_n, and the point process M of the initial points and their marks is a *p-marked Poisson process* associated with N.

The M is a *space-time* Poisson process when it is defined on $\mathbb{R}_+ \times S'$ (or $\mathbb{R} \times S'$) and $\mathbb{R}_+$ represents the time axis. The mean measure of M is given by (3.22) below.

Calling M a Poisson process in this definition is justified by the next result, which is an extension of Theorem 32 for deterministic marks.

Theorem 36. *The point process* $M = \sum_n \delta_{(X_n, Y_n)}$ *in Definition 35 is a Poisson process on* $S \times S'$ *with intensity measure* μ_M *defined by*

$$\mu_M(A \times B) = \int_A p(x, B)\mu(dx), \qquad A \in \mathcal{S},\ B \in \mathcal{S}'. \tag{3.22}$$

Hence, the point process of marks $N' = \sum_n \delta_{Y_n}$ is a Poisson process on S' with intensity $\mu'(B) = \int_S p(x, B)\mu(dx)$, $B \in \mathcal{S}'$, provided this is locally finite.

Proof. It suffices by Theorem 21 to show that the Laplace functional of M is the same as that for a Poisson process on $S \times S'$ with intensity measure μ_M. Since Y_n are p-marks of the X_n,

$$\begin{aligned} E[e^{-Mf}] &= E\Big[E[e^{-\sum_n f(X_n,Y_n)}|N]\Big] \\ &= E\Big[\prod_n \int_{S'} e^{-f(X_n,y)}p(X_n, dy)\Big] \\ &= E\Big[\exp\Big\{\sum_n \log \int_{S'} e^{-f(X_n,y)}p(X_n, dy)\Big\}\Big]. \end{aligned}$$

Letting $h(x) = -\log \int_{S'} e^{-f(x,y)}p(x, dy)$ in the last line, and using the Laplace functional of N as in Theorem 21, we have

$$\begin{aligned} E[e^{-Mf}] &= E[e^{-\int_{S'} h(x)N(dx)}] \\ &= \exp[-\int_S (1 - e^{-h(x)})\mu(dx)] \\ &= \exp[-\int_{S\times S'} (1 - e^{-f(x,y)})\mu_M(dx,\ dy)]. \end{aligned}$$

The last line by Theorem 21 is the Laplace functional of a Poisson process with mean measure μ_M and hence M is such a Poisson process.

In addition $N'(B) = M(S \times B)$ is M on part of its state space, and so N' is Poisson with intensity $\mu'(B) = \mu_M(S \times B) = \int_S p(x, B)\mu(dx)$.

Here are some examples. Further applications are in the next sections and in Exercises 47, 49 and 32.

Example 37. Marked Tornados. Suppose $N = \sum_n \delta_{X_n}$ is a Poisson process on a region S of a country that represents locations of tornados that might occur in a year. For simplicity, assume its intensity $\mu(S)$ is finite. Additional information about the tornados is naturally recorded by marks Y_n associated with the tornados at the respective locations X_n. For instance, Y_n might record the cost to repair the damage, the number of deaths, or a vector of auxiliary information for tornado X_n. Assume the Y_n take values in a space S' and are p-marks of the X_n. Then the tornado information is conveniently represented by the p-marked Poisson processes

$$M = \sum_n \delta_{(X_n,Y_n)}, \qquad N' = \sum_n \delta_{Y_n}.$$

Example 38. Maxima of Marks. Suppose that events occur over the time axis $\mathbb{R}_+$ according to a Poisson process $N = \sum_n \delta_{T_n}$ with intensity measure μ. The event that occurs at time T_n produces a real-valued random variable Y_n. Assume the Y_n are p-marks of the T_n, where $p(t, \cdot)$ is the distribution of a mark at time t. Consider the cumulative maxima process

$$V(t) = \max_{n \le N(t)} Y_n, \qquad t \ge 0,$$

where $V(t) = 0$ when $N(t) = 0$. For instance, this could be the maxima of heights of solar flares on a region of the sun, where the flares occur at times that form a Poisson process.

One can obtain information about this maxima process via the space-time Poisson process $M = \sum_n \delta_{(T_n, Y_n)}$ on $\mathbb{R}_+ \times \mathbb{R}$ with intensity

$$E[M((0,t] \times B)] = \int_{(0,t]} p(s, B)\mu(ds).$$

For instance, the event $\{V(t) \le y\}$ equals $\{M((0,t] \times (y, \infty)) = 0\}$, and so

$$P\{V(t) \le y\} = e^{-\int_{(0,t]} p(s,(y,\infty))\mu(ds)}.$$

Also, if $\mu(t) = \lambda t$ and $p(t, \cdot) = G(\cdot)$, independent of t, then $V(t)$ is a continuous-time Markov process with transition probabilities

$$\begin{aligned} P\{V(s+t) \le y | V(s) = x\} &= \mathbf{1}(x \le y) P\{M((s, s+t] \times (y, \infty)) = 0\} \\ &= \mathbf{1}(x \le y) e^{-\lambda t(1-G(y))}. \end{aligned}$$

Example 39. Serial Marking of a Poisson Process. Theorem 36 for a single marking of a Poisson process extends to a series of markings as follows. Starting with a Poisson process $N = \sum_n \delta_{X_n}$, if Y_n are p-marks of X_n, then $M = \sum_n \delta_{(X_n, Y_n)}$ is a Poisson process. Similarly, if Y_n' are p'-marks of (X_n, Y_n), then $M' = \sum_n (X_n, Y_n, Y_n')$ is again a Poisson process. These marking steps can be continued several times, with the end result being a Poisson process from which one can "read off" many results. In addition to serial markings in applications, they are useful for proving results for compound Poisson processes as discussed in Section 3.15.

3.10 Partitions and Translations of Poisson Processes

In this section, we continue our study of transformations of Poisson processes by considering partitioning and translations of the points of a Poisson process. We show how these two types of transformations can be modeled as marked Poisson processes.

We begin with a special kind of partitioning.

Example 40. Thinning of a Poisson Process. Let N be a Poisson process on S with intensity μ. Suppose the points of N are deleted according to the rule that a point at x is retained with probability $p(x)$, and the point is deleted with probability $1 - p(x)$. Let N_1 and N_2 denote the resulting processes of retained and deleted points, respectively, where $N = N_1 + N_2$. By Proposition 41 below, N_1 and N_2 are independent Poisson processes with respective mean measures

$$E[N_1(A)] = \int_A p(x)\mu(dx), \quad E[N_2(A)] = \int_A (1 - p(x))\mu(dx), \quad A \in \mathcal{S}.$$

Interestingly, N_1 and N_2 are independent even though $N = N_1 + N_2$.

As an example, suppose a web site that sells products has visitors arriving to it according to a Poisson process N with rate λ. Suppose p percent of these visitors buy a product, which means that each visitor independently buys a product with probability p. Then from the preceding result, the times of sales form a Poisson process with rate $p\lambda$, and the visits without sales occur according to a Poisson process with rate $(1 - p)\lambda$.

The preceding thinning model is a special case of the following partitioning procedure for decomposing a point process into several subprocesses. Consider a Poisson process N on S with intensity μ. Suppose N is partitioned into a countable family of processes N_i, $i \in I$, on S by the following rule.
Partitioning Rule: A point of N at x is assigned to subprocess N_i with probability $p(x, i)$, where $\sum_{i \in I} p(x, i) = 1$.

The processes N_i form a partition of N in that $N = \sum_{i \in I} N_i$.

Proposition 41. (Partitioning of a Poisson Process) *The subprocesses N_i, $i \in I$, of the Poisson process N are independent Poisson processes with intensities*

$$E[N_i(B)] = \int_B p(x, i)\mu(dx), \quad B \in \mathcal{S},\ i \in I.$$

Proof. Let $M(B \times \{i\})$ denote the number of points of N in B that are assigned to N_i. That is, $M(B \times \{i\}) = N_i(B)$. Clearly, M is a marked Poisson process on $S \times I$ associated with N, where the marks have the distribution $p(x, i)$. Since M has independent increments and the subprocesses N_i represent M on the disjoint subsets $S \times \{i\}$, for $i \in I$, they are independent Poisson processes. Furthermore,

$$E[N_i(B)] = E[M(B \times \{i\})] = \int_B p(x, i)\mu(dx).$$

The preceding result for partitions is the opposite of the result that a sum of independent Poisson processes is also Poisson (recall Theorem 22). Here is a typical partition model.

Example 42. Suppose telephone calls in a region S of the USA occur according to a space-time Poisson process M on $\mathbb{R}_+ \times S$, where $M((0,t] \times B)$ denotes the number of calls connected in the subregion $B \subseteq S$ in the time interval $(0,t]$, and $E[M((0,t] \times B)] = \lambda t \mu(B)$. There are three types of calls: (1) Long distance calls outside the USA. (2) Long distance calls within the USA. (3) Local calls. The calls are independent, and a call at time t and location x is a type i call with probability $p(t,x;i)$, $i = 1,2,3$. Then by Proposition 41, the number of type i calls occur according to a space-time Poisson process M_i with

$$E[M_i((0,t] \times B)] = \lambda \int_0^t \int_B p(s,x;i)\mu(dx)ds.$$

Furthermore, M_1, M_2, M_3 are independent and $M = M_1 + M_2 + M_3$.

The next two examples illustrate indirect uses of thinning.

Example 43. Simulating a Non-Homogeneous Poisson Process. Suppose that $N = \sum_n \delta_{X_n}$ denotes a Poisson process on a bounded Euclidean space S with mean measure $\Lambda(A) = \int_A \lambda(x)\,dx$ and rate function $\lambda(x)$.

Let us represent N as a thinning of a homogeneous Poisson process $\bar{N} = \sum_n \delta_{\bar{X}_n}$ on S with rate 1. Accordingly, suppose that Y_n are location-dependent marks of $\bar{X}_n$ with $P\{Y_n = 1|\bar{N}\} = p(\bar{X}_n) = 1 - P\{Y_n = 0|\bar{N}\}$ where $p(x) = \lambda(x)/\Lambda(S)$. These marks form a $p(x)$-thinning of $\bar{N}$ and the resulting process is

$$\sum_n Y_n \delta_{\bar{X}_n} \stackrel{d}{=} N.$$

This thinning representation of N justifies the following procedure from [79] for simulating N via a realization of $\bar{N}$. Construct a realization of the point locations of $\bar{N}$, say $\bar{x}_1, \ldots, \bar{x}_m$ in S. Next, independently thin the points such that $\bar{x}_n$ is retained with probability $p(\bar{x}_n)$ and it is deleted otherwise, for $1 \le n \le m$. Then the retained points $x_1, \ldots, x_\ell$ form a realization of N, because of its representation above.

Example 44. Terminating Poisson Process. Suppose that errors in a software package occur (while it is running) at times T_n that form a Poisson process on $\mathbb{R}_+$ with rate λ. Each error is detected with probability $1-p$ independent of everything else. Then the time to detect the first error is $T_\nu = \sum_{i=1}^{\nu} X_i$, where $X_1, X_2, \ldots$ are the exponential times between errors and ν is the number of errors until the first one is detected. The ν is a random variable independent of the X_n with the geometric distribution $P\{\nu = n\} = (1-p)p^{n-1}$, $n \ge 1$. The T_ν can be viewed as the time at which the Poisson process terminates. Exercise 5 shows that T_ν is exponentially distributed with rate $(1-p)\lambda$, since it is a geometric sum of exponential variables.

An alternate proof is to consider the process of detected errors as a thinning of the Poisson error process, where $1-p$ is the probability of retaining a point. The resulting thinned process is a Poisson process with rate $(1-p)\lambda$, and T_ν

is the time to its first point. This proves T_ν is exponentially distributed with rate $(1-p)\lambda$.

Splitting and merging of flows in a network, as we now show, are typical examples of partitioning and summing of point processes.

Example 45. Routing in a Graph. Consider the directed graph shown in Figure 3.1 in which units are routed in the directions of the arrows. Let $N_{ij}(t)$ denote the number of units that are routed on the arc from node i to node j in the time interval $(0,t]$. Assume that items enter the graph by independent Poisson processes N_{0j}, $j=1,2,3$ on $\mathbb{R}_+$ with respective rates λ_{0j}, $j=1,2,3$. Upon entering the graph, each item is routed independently through the graph according to the probabilities on the arcs, and there are no delays at the nodes (travel through the graph is instantaneous). For instance, an item entering node 3 is routed to node 5 or node 6 with respective probabilities p_{35} and p_{36}, where $p_{35}+p_{36}=1$.

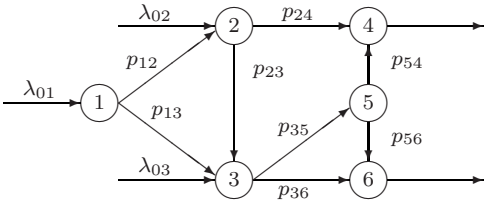

Fig. 3.1 Partitioning and Merging of Flows

Our results on partitions and sums of Poisson processes yield the following properties. First note that flows N_{12} and N_{13} are independent Poisson processes with rates $\lambda_{12}=p_{12}\lambda_{01}$ and $\lambda_{13}=p_{13}\lambda_{01}$, since they form a partition of N_{01}. Next, the flow into node 2 is the sum $N_{02}+N_{12}$ (of independent flows) and hence is Poisson with rate $\lambda_{02}+p_{12}\lambda_{01}$. Similar properties extend to the other flows in the graph. Specifically, each flow N_{jk} from j to k is a Poisson process, and one can evaluate their rates λ_{jk} in the obvious way. For instance, knowing λ_{12} and λ_{13} as mentioned above,

$$\begin{aligned} \lambda_{23} &= p_{23}(\lambda_{02}+\lambda_{12}), & \lambda_{36} &= p_{36}(\lambda_{03}+\lambda_{13}+\lambda_{23}),\\ \lambda_{35} &= p_{35}(\lambda_{03}+\lambda_{13}+\lambda_{23}), & \lambda_{60} &= \lambda_{36}+p_{56}\lambda_{35}. \end{aligned}$$

Also, some of the flows are independent (denoted by $\perp$), while some are not independent (denoted by $\not\perp$). Examples are

$$\begin{aligned} &N_{12}\perp N_{13}, \quad N_{36}\perp N_{56}, \quad N_{36}\perp N_{24}, \quad N_{13}\perp N_{24}\\ &N_{12}\not\perp N_{24}, \quad N_{35}\not\perp N_{40}, \quad N_{13}\not\perp N_{60}, \quad N_{23}\not\perp N_{40}. \end{aligned}$$

In addition, the flow $N_i=\sum_j N_{ji}$ through each node i is a Poisson process with intensity $\sum_j \lambda_{ji}$. Clearly all the N_i's are dependent. If the arc between

5 and 4 did not exist, however, then N_4 would be independent of N_3, N_5, and N_6.

We end this section with another natural transformation of a Poisson process involving translating its points within the same space. Suppose that N is a Poisson process on $S = \mathbb{R}^d$ with intensity measure μ. Assume that a point of N at x is independently translated to another location $x+Y$ by a random vector Y in S that has a distribution $G_x(\cdot)$. That is, x is mapped into a set $B \subseteq S$ by a probability kernel $p(x, B) = G_x(B - x)$, where $B - x = \{y - x : y \in B\}$. Let $M(A \times B)$ denote the number of points of N in A that are translated into B. Then the process $N'(B) = M(S \times B)$ denotes the number of points of N translated into B. The definition of marked Poisson processes yields the following result.

Proposition 46. (Translation of a Poisson process) *The translation processes M and N' defined above are Poisson processes and, for $A, B \subseteq \mathbb{R}^d_+$,*

$$E[M(A \times B)] = \int_A G_x(B - x)\mu(dx), \quad E[N'(B)] = \int_S G_x(B - x)\mu(dx).$$

Example 47. Trees in a Forest. The locations (X_n, Y_n) of a certain type of tree in a forest form a Poisson process with intensity measure μ. Suppose the height of a tree at a location (x, y) has a distribution $G_{x,y}(\cdot)$. That is, the height Z_n of the tree at location (X_n, Y_n) is a mark, and $M = \sum_n \delta_{(X_n,Y_n,Z_n)}$ forms a marked Poisson process with

$$E[M(A \times B \times (0, b])] = \int_A \int_B G_{x,y}(b)\mu(dx\, dy).$$

After several years of growth, it is anticipated that the increase in height for a tree has a distribution $H_{(x,y,z)}(\cdot)$, where (x, y) is the location and z is the original height. In other words, the increases Z'_n are p-marks of (X_n, Y_n, Z_n) with $p((x, y, z), \cdot) = H_{(x,y,z)}(\cdot)$. Then the collection of trees is depicted by the point process $M' = \sum_n \delta_{(X_n,Y_n,Z_n+Z'_n)}$. By Proposition 46, M' is a Poisson process with

$$E[M'(A \times B \times (0, b])] = \int_A \int_B \int_{\mathbb{R}_+} H_{(x,y,z)}(b - z)G_{x,y}(dz)\mu(dx\, dy).$$

In the preceding example, a little more realism could be added by considering the possibility that while some trees grow as indicated, other trees may die according to a location-dependent thinning. Then one would have a combined translation–thinning transformation. Similarly, complicated systems might involve transformations involving a combination of translations, thinnings, partitions, deterministic maps, and random transformations.

3.11 Markov/Poisson Processes

In this section, we discuss a discrete-time Markov chain whose state at each time is a spatial Poisson process. This Markov/Poisson process is formulated by successive transformations of Poisson processes. This model reveals an interesting property of invariant measures of Markov chains.

We will describe a Markov/Poisson process in the context of a particle system. Consider a family of particles that move about in a space S (e.g., $\mathbb{R}^d$, or a countable set) as follows. At time 0, the particles are located in S according to a point process N_0 on S. Thereafter, each particle moves independently in S at discrete times as if it were a Markov chain[3] on the space S with the one-step transition kernel $p(x, B)$. That is, a particle located at x at time n moves into a set B at time $n+1$ with probability $p(x, B)$. Then a particle in state x at time 0 will be in a set B at time n with probability $p^n(x, B)$. These n-step probabilities are defined by the recursion[4]

$$p^n(x, B) = \int_S p^{n-1}(y, B)p(x, dy), \qquad n \geq 1.$$

As in the setting of countable state space Markov chains, a measure μ on S is an *invariant measure* of $p(x, B)$ if

$$\int_S p(x, B)\mu(dx) = \mu(B), \qquad B \in \mathcal{S}. \tag{3.23}$$

Consider the point process N_n on S that represents the locations of the points at time n. The sequence N_n is a discrete-time Markov chain that takes values in the set of counting measures on S. This follows since the point locations of N_{n+1} depend on $N_0, \ldots, N_n$ only through the point locations of N_n and the one-step transition probabilities (which are functions of the $p(x, B)$ that do not depend on n).

As above, we will analyze the Markov chain N_n via the marked point processes M_n on S^2, where $M_n(A \times B)$ denotes the number of particles in A at time 0 that are in B at time n. The sequence M_n is a Markov chain for the same reason that N_n is.

Theorem 48. *Suppose N_0 is a Poisson process with intensity measure μ. Then each M_n is a Poisson process on S^2 with*

$$E[M_n(A \times B)] = \int_A p^n(x, B)\mu(dx), \qquad A, B \in \mathcal{S},$$

[3] The definition of a Markov chain in Chapter 1 extends readily to uncountable state spaces, as is the case here. This example should be understandable by thinking of the probability kernel as a transition probability for a countable state space and interpreting the notions of invariant and stationary distributions as one would for a countable state space.

[4] When S is countable, the matrix $(p^n(x, y))$ is the nth product of the matrix $(p(x, y))$.

and N_n is a Poisson process with $E[N_n(B)] = \int_S p^n(x,B)\mu(dx)$. If in addition, the intensity μ for N_0 is an invariant measure of $p(x,B)$, then $\{N_n : n \geq 0\}$ is a stationary Markov chain, and each N_n is a Poisson process on S with intensity μ.

Proof. Think of M_n as a transformation of N_0 in which a point of N_0 at x is independently mapped into a point of M_n in a set $A \times B$ with probability kernel $r(x, A \times B) = \mathbf{1}(x \in A)p^n(x,B)$. Then by Theorem 36, M_n is a r-marked Poisson process associated with N_0 and its intensity is given by

$$E[M_n(A \times B)] = \int_{S \times S'} r(x, A \times B)\mu(dx) = \int_A p^n(x,B)\mu(dx).$$

Furthermore, N_n is also Poisson and its intensity as shown is locally finite since μ is. This proves the first assertion.

Next, suppose that μ is an invariant measure for N_0. Then an induction argument using (3.23) shows that $\int_S p^n(x,B)\mu(dx) = \mu(B)$, $B \in \mathcal{S}$, $n \geq 1$. Thus $E[N_n(B)] = \mu(B)$, and so the Markov chain N_n is stationary.

Consider the stationary Markov/Poisson Particle process N_n described in the preceding theorem. Being a stationary Markov chain, the distribution of N_0 (a Poisson process on S with intensity μ) is the stationary and limiting distribution of the chain. This distribution, as we saw in Chapter 1, describes many performance measures of the particle system. Here is an illustration.

Example 49. Set with no particles. Let us see what we can say about how likely it is that there are no particles in a set $B \in \mathcal{S}$ when $\mu(S) < \infty$. From the stationary nature of the Markov chain N_n, it is clear that the portion of time that a set B contains no particles is $P\{N_0(B) = 0\} = e^{-\mu(B)}$.

Next, consider the average duration of time $W(B)$ that the set B is empty; $W(B) = \lim_{n\to\infty} n^{-1} \sum_{k=1}^n W_k(B)$, where $W_1(B), W_2(B), \ldots$ are the successive durations of (discrete) time that B is empty. We will determine $W(B)$ by the Little law in Theorem 57 in Chapter 2 for the artificial queueing process $Q_n = \mathbf{1}(N_n(B) = 0)$ (this is 1 when B is empty and 0 otherwise).

The average queue length (in discrete time) is

$$L(B) = \lim_{n\to\infty} n^{-1} \sum_{k=1}^n Q_k = P\{N_0(B) = 0\} = e^{-\mu(B)} \quad a.s.$$

Also, the rate at which the queue becomes empty is

$$\begin{aligned}\lambda(B) &= \lim_{n\to\infty} n^{-1} \sum_{k=1}^n \mathbf{1}(N_{k-1}(B) > 0, N_k(B) = 0) \\ &= P\{N_0(B) > 0, N_1(B) = 0\} \quad a.s.\end{aligned}$$

Using the stationarity of N_n,

$$\begin{aligned}\lambda(B) &= P\{N_1(B) = 0\} - P\{N_0(B) = 0,\ N_1(B) = 0\} \\ &= e^{-\mu(B)} - e^{-\mu(B)} P\{N_1(B) = 0 | N_0(B) = 0\}.\end{aligned}$$

By the sample process representation of the Poisson process N_0 on B^c, a typical point in B^c is located in a set $C \subseteq B^c$ with probability $\mu(C)/\mu(B^c)$. Then the probability that a typical point in B^c does not enter B in the next step is $r = \int_{B^c} p(x, B^c)\mu(dx)/\mu(B^c)$. Using this,

$$\begin{aligned}P\{N_1(B) = 0 | N_0(B) = 0\} &= E[P\{N_1(B) = 0 | N_0(B) = 0, N_0(B^c)\}] \\ &= E[r^{N_0(B^c)}] \ = \ e^{-\mu(B^c)(1-r)}.\end{aligned}$$

Substituting this expression in the preceding display yields

$$\lambda(B) = e^{-\mu(B)}[1 - e^{-\int_{B^c} p(x,B)\mu(dx)}].$$

Since this quantity is positive, Theorem 57 in Chapter 2 ensures that the limit $W(B)$ exists and $L(B) = \lambda(B)W(B)$. Consequently,

$$W(B) = [1 - e^{-\int_{B^c} p(x,B)\mu(dx)}]^{-1}.$$

3.12 Poisson Input-Output Systems

This and the next section show how one can use marked Poisson processes to model many processing systems with Poisson input, and arrival-dependent service times, and no queueing. Here we describe $M_t/G_t/\infty$ systems, which are time-dependent versions of a classical $M/G/\infty$ system. The results involve formulating processes of interest as functions of a space-time Poisson input process, and then characterizing several system features by applying the results above for translations, projections and random transformations.

Consider a general processing system that operates as follows. Items arrive at times that form a Poisson process on $\mathbb{R}_+$ with intensity measure μ. An item that arrives at time t spends a random amount of time in the system that has a distribution $G_t(\cdot)$ and then departs. This sojourn time is independent of the other items in the system and everything else. The items may arrive in batches at any time t for which $\mu(\{t\}) > 0$; the batch size has a Poisson distribution with mean $\mu(\{t\})$. The items in this batch may not depart at the same time since their sojourn times are independent.

Let $Q(t)$ denote the quantity of items in the system at time t that arrived after time 0. We are not considering items that may be in the system at time 0. The process $\{Q(t) : t \geq 0\}$ is an $M_t/G_t/\infty$ process with time-dependent arrivals and services.

The process $Q(t)$ is a typical model for the quantity of items in a service system with a large number of parallel servers (envisioned as infinite servers)

in which there is essentially no queueing prior to service. For instance, $Q(t)$ could be the number of: (1) Computers being used in a wireless network with a high capacity. (2) Groups of people dining in a cafeteria. (3) Vehicles in a parking lot. (4) Patients in a hospital. (5) Calls being processed in a call center.

To analyze the process $Q(t)$, the first step is to define it by the system data. The data is represented by the marked point process $M = \sum_n \delta_{(T_n, V_n)}$ on $\mathbb{R}^2_+$, where T_n is the arrival time of the nth item and V_n is its sojourn or service time. The V_n are location-dependent marks of the T_n with the distribution $p(t, B) = G_t(B)$, and M is a space-time Poisson process with $E[M(A \times B)] = \int_A G_s(B)\mu(ds)$.

Since the quantity $Q(t)$ is a function of the arrival times T_n and departure times $T_n + V_n$ of the items, let us consider the marked point process of these arrival/departure times, which is

$$N = \sum_n \delta_{(T_n, T_n+V_n)}, \qquad \text{on } S = \{(t, u) \in \mathbb{R}^2_+ : u \geq t\}.$$

This process N is a transformation of M by the map $g(t, v) = (t, t+v)$. Then N is a space-time Poisson process by Theorem 32. In particular, the quantity $N((a, b] \times (c, d])$, where $b \leq c$, is the number of items that arrive in $(a, b]$ and depart in $(c, d]$, and its mean is

$$E[N((a, b] \times (c, d])] = \int_{(a,b]} [G_s(d - s) - G_s(c - s)]\mu(ds).$$

Using the preceding notation, the quantity of items in the system at time t is defined by

$$Q(t) = \sum_n \mathbf{1}(T_n \leq t, T_n + V_n > t) = N((0, t] \times (t, \infty]).$$

Since N is a Poisson process, it follows that $Q(t)$ has a Poisson distribution with

$$E[Q(t)] = \int_{(0,t]} [1 - G_s(t - s)]\mu(ds). \tag{3.24}$$

Although the distribution of each $Q(t)$ is Poisson, the entire process is not.

In addition to analyzing the number of items in the system, one may want information about the departure process. This is useful when the departures form an arrival process into another service system. Now, the total number of departures in $(0, t]$ is $D(t) = \sum_n \mathbf{1}(T_n + V_n \leq t)$. That is, D is the projection of N on its second coordinate, and so D is a Poisson process with

$$E[D(t)] = E[N(t) - Q(t)] = \int_{(0,t]} G_s(t - s)\mu(ds).$$

Example 50. $M/G/\infty$ System. Consider the special case of the preceding system in which the Poisson arrival process is homogeneous with rate λ and the service distribution $G_t(\cdot) = G(\cdot)$ is independent of t. This is a classical $M/G/\infty$ system with arrival rate λ and service distribution G. In this case, $Q(t)$ has a Poisson distribution and from (3.24) (with $u = t - s$),

$$E[Q(t)] = \lambda \int_0^t [1 - G(u)]du.$$

Suppose that G has a mean α. Then it follows by Exercise 32 that the limiting distribution of $Q(t)$ is Poisson with rate $\lambda\alpha$, as $t \to \infty$.

In addition, the total number of departures $D(t)$ in the time interval $(0, t]$ is a Poisson process with

$$E[D(t)] = \lambda \int_0^t G(s)ds.$$

An abstraction of the $M_t/G_t/\infty$ system we just discussed is as follows.

Example 51. Poisson Input-Output-Mobility Model. Consider a system in which items enter a space S at times $T_1 \leq T_2 \leq \ldots$ that form a Poisson process with intensity measure μ. The nth item that arrives at time T_n moves in the space S for a while and then exists the system (by entering the outside state 0). The movement is determined by a stochastic process $Y_n = \{Y_n(t) : t \geq 0\}$ with state space $S \cup \{0\}$, where the outside 0 is an absorbing state ($Y_n(t) = 0$ for all $t > \inf\{s : Y_n(s) = 0\}$). In particular, the nth item enters S at the location $Y_n(0)$, and, at time $t > T_n$ its location is $Y_n(t - T_n)$. Let $\mathbb{Y}$ denote a function space that contains the sample paths of Y_n (e.g., $\mathbb{Y}$ could be a space of real-valued functions that are continuous, or piece-wise constant).

Assume the Y_n are location-dependent marks of T_n with distribution $p(t, \cdot)$, which is the conditional distribution of the process Y_n starting at time t. For simplicity, assume Y_n depends on t only through its initial value $Y_n(0)$ (the entry point in S of the nth point), whose distribution is denoted by $F_t(\cdot)$. Then conditioning on $Y_n(0)$,

$$p(t, \cdot) = \int_S P\{Y_n \in \cdot | Y_n(0) = x\} F_t(dx). \tag{3.25}$$

In other words, the system data is the process $M = \sum_n \delta_{(T_n, Y_n)}$ on $\mathbb{R}_+ \times \mathbb{Y}$, which is a space-time Poisson process by Theorem 32.

Now, the number of items in the set $B \in \mathcal{S}$ at time t is given by

$$N_t(B) = \sum_n \mathbf{1}(T_n \leq t, Y_n(t - T_n) \in B) = \sum_n \delta_{g_t(T_n, Y_n)}(B),$$

where $g_t(s,y) = y(t-s)$ and $y \in \mathbb{Y}$. Since N_t is a transformation of the Poisson process M by the map g_t, it follows by Theorem 32 that N_t is a Poisson process on S for each fixed t, and from (3.25),

$$E[N_t(B)] = \int_{(0,t]} \int_S P^{t-s}(x,B) F_s(dx) \mu(ds),$$

where $P^t(x,B) = P\{Y_n(t) \in B | Y_n(0) = x\}$.

Next, note that the number of departures from the set B in the time interval $(a,b]$ is

$$D((a,b] \times B) = \sum_n \mathbf{1}(h(T_n, Y_n) \in (a,b] \times B),$$

where $h(s,y) = (s, y(t-s))$. This D is a transformation of the Poisson process M by the map h, and so by Theorem 32, D is a space-time Poisson process on $\mathbb{R}_+ \times S$ with

$$E[D((0,t] \times B)] = \int_{(0,t]} \int_B P^{t-s}(x, \{0\}) F_s(dx) \mu(ds).$$

3.13 Network of $M_t/G_t/\infty$ Stations

In this section, we show how the ideas in the preceding section extend to the analysis of flows in a stochastic network of $M_t/G_t/\infty$ stations. The network dynamics are determined by marks of Poisson processes, and the analysis amounts to formulating appropriate Poisson processes that represent parameters of interest, and then specifying their intensity measures.

Consider a network of m service stations (or nodes) that operate as follows. Items enter the network at times $T_1 \le T_2 \le \ldots$ that form a Poisson process with intensity measure μ. The nth item entering the network at time T_n selects, or is assigned, a random route $\mathbf{R}_n = (R_{n1}, \ldots, R_{nL_n})$ through the network, where $R_{nk} \in \{1, \ldots, m\}$ denotes the kth node the item visits, and the length $1 \le L_n \le \infty$ may be random and depend on the R_{nk}. After visiting the last node R_{nL_n} on its route, the item exits the network and enters node 0 ("outside" the network) and stays there forever.

Associated with the nth arrival is a vector of nonnegative sojourn (or visit) times $\mathbf{V}_n = (V_{n1}, \ldots, V_{nL_n})$, where V_{nk} is the item's sojourn time at node R_{nk}. The time at which the item departs from node R_{nk} is

$$\tau_{nk} = T_n + \sum_{j=1}^{k} V_{nj}, \qquad k \le L_n,$$

where τ_{nL_n} is the time at which the item exits the network.

The main assumption is that the route and waiting time vectors $Y_n = (\mathbf{R}_n, \mathbf{V}_n)$ are marks of the arrival times T_n. This implies there are no interactions among the items that affect their waiting times, so each node operates like an $M_t/G_t/\infty$ system. As above, we consider only those items that enter the network "after" time 0. In summary, the system data is represented by the space-time Poisson process $M = \sum_n \delta_{(T_n, Y_n)}$.

Many features of the network are expressible by space-time Poisson processes of the form

$$N_t = \sum_n \delta_{(T_n, g_t(T_n, Y_n))}, \tag{3.26}$$

$$E[N_t((a,b] \times B)] = \int_{(a,b]} P\{g_t(T_n, Y_n)) \in B | T_n = s\} \mu(ds). \tag{3.27}$$

One need only define the function g_t for the application at hand; in some cases, g_t and N_t do not depend on t. Note that each N_t is a Poisson process since it is a deterministic transformation of the Poisson process M.

Typical uses of these space-time Poisson processes are as follows.

Locations of Items at Time t. The space-time Poisson process describing where the items are located is

$$N_t((a,b] \times B) = \#\text{ of items that arrive in the time interval } (a,b] \text{ that are in } B \subseteq \{0,1,\ldots,m\} \text{ at time } t.$$

The location of the item at time t that arrives at time T_n is

$$g_t(T_n, Y_n) = \begin{cases} 0 & \text{if } \tau_{nL_n} \le t \\ R_{nk} & \text{if } \tau_{n(k-1)} \le t < \tau_{nk}, \text{ for some } k \le L_n. \end{cases} \tag{3.28}$$

Consequently, the quantities $Q_i(t) = N_t((0,t] \times \{i\})$, $1 \le i \le m$, at the nodes at a fixed time t are independent Poisson random variables with

$$E[Q_i(t)] = \int_{(0,t]} P\{g_t(T_n, Y_n) = i | T_n = s\} \mu(ds). \tag{3.29}$$

Departure Process. The space-time Poisson process describing the times at which items exit the network is

$$N((a,b] \times B) = \#\text{ of items arriving in } (a,b] \text{ whose exit time from the network is in } B \subseteq \mathbb{R}_+.$$

The item arriving at T_n exits the network a time $g(T_n, Y_n) = \tau_{nL_n}$.

Usually, the mean values (3.27) of the space-time Poisson processes would be determined by the distributions of the routes and sojourn times of the items. The routes depend on the structure of the network and the nature of

the items and services. A standard assumption is that the routes are independent and Markovian, where p_{jk} denotes the probability of an item moving to node k upon departing from node j. Then the probability of a particular route $(r_1, \ldots, r_\ell)$ of nonrandom length ℓ is $p_{0r_1} \cdots p_{r_\ell 0}$. Another convention is that there are several types of items and all items of the same type take the same route. In this case, the probability of a route is the probability that the item entering the network is the type that takes that route.

The simplest sojourn times at a node are those that are i.i.d., depending on the node and independent of everything else. Then the sums of sojourn times are characterized by convolutions of the distributions. The next level of generality is that the service times are independent at the nodes, but their distributions may depend on the route as well as the node. An example of dependent service times is that an item entering a certain subset of routes is initially assigned a service time according to some distribution and then that time is its service time at "each" node on its route.

Here is a typical example of a network.

Example 52. Acyclic Network. Consider the stochastic network shown in Figure 3.2 that operates as described above with the following particular properties. Items arrive at the nodes 1, 2 and 3 from outside according to independent Poisson processes with respective rates λ_1, λ_2, λ_3. The sojourn or service times at the nodes are independent random variables, and the sojourn times at node i have the distribution $G_i(\cdot)$. When an item ends its sojourn at node 1, it departs and enters node 2 with probability p_{12}, or it enters node 3 with probability $p_{13} = 1 - p_{12}$. Analogously, departures from node 2 enter node 3 with probability p_{23}, or enter node 4 with probability $p_{24} = 1 - p_{23}$. Also, departures from node 3 enter node 5 ($p_{35} = 1$), and departures from nodes 4 and 5 exit the network.

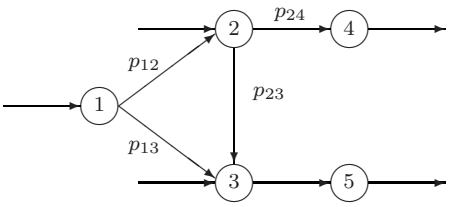

Fig. 3.2 Acyclic Network

The times $T_1 < T_2 < \ldots$ at which items enter the network from outside form a Poisson process with rate $\lambda = \lambda_1 + \lambda_2 + \lambda_3$, since this process is the sum of the three independent Poisson processes flowing into nodes 1, 2 and 3. The probability that an arrival at any time T_n enters node i is λ_i/λ. This is the probability that the exponential time of an arrival at i is smaller than those exponential arrival times at the other nodes; see Exercise 2. The item that arrives at time T_n traverses a route $\mathbf{R}_n = (R_{n1}, \ldots, R_{n\ell_n})$ in $\mathcal{R}$ (the

set of all routes), and its sojourn times at the ℓ_n nodes on the route are $\mathbf{V}_n = (V_{n1}, \ldots, V_{n\ell_n})$. The joint distribution of these marks $Y_n = (\mathbf{R}_n, \mathbf{V}_n)$ as functions of the network data λ_i, G_i and p_{ij} is

$$P\{\mathbf{R}_n = \mathbf{r}, \mathbf{V}_n \le \mathbf{v} | T_n\} = p(\mathbf{r}) \prod_{k=1}^{\ell} G_{r_k}(v_k),$$

where $p(\mathbf{r}) = (\lambda_{r_1}/\lambda) p_{r_1 r_2} \cdots p_{r_{\ell-1} r_\ell}$ is the probability of route $\mathbf{r} = (r_1, \ldots, r_\ell)$.

To analyze the quantity of items on the routes as well as at the nodes, let us consider the space-time point process

$$N_t((a,b] \times B) = \#\text{ of items arriving in } (a,b] \text{ whose route and node location } (\mathbf{r}, i) \text{ is in } B \subseteq \mathcal{R} \times \{0, 1, \ldots, m\} \text{ at time } t.$$

As in (3.26), N_t is a Poisson process on $\mathbb{R}_+ \times \mathcal{R} \times \{0, 1, \ldots, m\}$, for fixed t, where $g_t(T_n, Y_n) = (\mathbf{R}_n, h_t(T_n, Y_n))$ and $h_t(T_n, Y_n)$ is defined by (3.28). The item that enters at T_n is at node $h_t(T_n, Y_n)$ at time t.

In particular, the quantity of items

$$Q_i(t) = \sum_{\mathbf{r} \in \mathcal{R}_i} N_t((a,b] \times \{\mathbf{r}\} \times \{i\})$$

at node i at time t has a Poisson distribution. The sum here is over all routes in the set of routes $\mathcal{R}_i$ that contain node i. Also, for a fixed t, since N_t has independent increments, it follows that $Q_i(t)$ is independent of $Q_j(t)$ if $\mathcal{R}_i$ and $\mathcal{R}_j$ are disjoint. For instance, $Q_4(t)$ is independent of $Q_3(t)$ and $Q_5(t)$.

The next step is to evaluate the intensity of N_t. Let $P^u_{\mathbf{r}}(i)$ denote the conditional probability that an item is at node i given that it is on route $\mathbf{r}$ for a time u since it entered the network. By the independence of the sojourn times,

$$P^u_{\mathbf{r}}(i) = \begin{cases} G_{r_1} \star \cdots \star G_{r_\ell}(u) & \text{if } i = 0 \\ G_{r_1} \star \cdots \star G_{r_{k-1}} \star \overline{G}_{r_k}(u) & \text{if } r_k = i \ne 0, \text{ for some } k \le \ell. \end{cases}$$

Here $\overline{G}(t) = 1 - G(t)$. For instance, the conditional probability that an item is at node 3 at time t, given that it enters route $\mathbf{r} = (1, 2, 3, 5)$ at time s, is

$$\begin{aligned} G_1 \star G_2 \star \overline{G}_3(t-s) &= P\{\tau_{n2} \le t < \tau_{n3} | \mathbf{R}_n = \mathbf{r}, T_n = s\} \\ &= P\{h_t(T_n, Y_n) = 3 | \mathbf{R}_n = \mathbf{r}, T_n = s\}. \end{aligned}$$

Then from (3.27), it follows that

$$E[N_t((a,b] \times \{\mathbf{r}\} \times \{i\})] = \lambda p(\mathbf{r}) \int_a^b P^{t-s}_{\mathbf{r}}(i) ds.$$

The last integral equals $\int_a^b P^u_{\mathbf{r}}(i)du$, under the change-of-variable $u = t - s$. For example, the number of items arriving in $(0, t]$ that are on route $\mathbf{r} = (1, 2, 3, 5)$ and in node 3 at time t has a Poisson distribution with mean

$$E[N_t((0,t] \times \{\mathbf{r}\} \times \{3\})] = \lambda_1 p_{12} p_{23} \int_0^t G_1 \star G_2 \star \overline{G}_3(u)du.$$

The process N_t yields considerable information about numbers of items at nodes and on routes as well. For instance, the quantity $Q_3(t)$ of items at node 3 at time t is the sum of the quantities of items on the routes in $\mathcal{R}_3 = \{(3,5), (2,3,5), (1,3,5), (1,2,3,5)\}$, all the routes containing node 3. Then $Q_3(t)$ has a Poisson distribution and its mean is

$$E[Q_3(t)] = \int_0^t \Big[\lambda_3 \overline{G}_3(u) + \lambda_2 p_{23} G_2 \star \overline{G}_3(u) + \lambda_1 p_{13} G_1 \star \overline{G}_3(u) + \lambda_1 p_{12} p_{23} G_1 \star G_2 \star \overline{G}_3(u)\Big] du.$$

The term in brackets is $\lambda \sum_{\mathbf{r} \in \mathcal{R}_3} p(\mathbf{r}) P^u_{\mathbf{r}}(3)$.

Similarly, the quantity of items on route $\mathbf{r}$ at time t is

$$Q_{\mathbf{r}}(t) = N_t((0,t] \times \{\mathbf{r}\} \times \{r_1, \ldots, r_\ell\}).$$

This quantity, being part of the Poisson process N_t, has a Poisson distribution whose mean is easy to calculate. For instance,

$$E[Q_{(2,4)}(t)] = \int_0^t \lambda_2 p_{24} [\overline{G}_2(u) + G_2 \star \overline{G}_4(u)] du.$$

Let us now consider the departure times of the items from the nodes, which are depicted by the process

$$N((a,b] \times B) = \#\text{ of items arriving in } (a,b] \text{ whose departure times from the 5 nodes are in } B \subseteq \mathbb{R}^5_+.$$

The departure times are well-defined since an item cannot visit a node more than once. Now, N is a space-time Poisson process as in (3.26) and (3.27), and the departure times at the 5 nodes are given by

$$g(T_n, Y_n) = (g_1(T_n, Y_n), \ldots, g_5(T_n, Y_n)),$$

where $g_i(T_n, Y_n) = \sum_{k=1}^{\ell_n} \tau_{nk} \mathbf{1}(R_{nk} = i)$, the departure time from node i of the item that enters at T_n.

In particular, the departure process at node i is

$$D_i(t) = N((0,t] \times \{(t_1, \ldots, t_5) \in \mathbb{R}^5_+ : t_i \le t\}), \qquad t \ge 0.$$

Now, D_i is a Poisson process since it is the projection on the ith departure-time coordinate of the Poisson process N. Its mean is

$$E[D_i(t)] = \lambda \int_0^t P\{g_i(T_n, Y_n) \le t | T_n = s\} ds. \tag{3.30}$$

For instance,

$$E[D_3(t)] = \int_0^t \Big[\lambda_3 G_3(u) + \lambda_2 p_{23} G_2 \star G_3(u) + \lambda_1 p_{13} G_1 \star G_3(u) + \lambda_1 p_{12} p_{23} G_1 \star G_2 \star G_3(u)\Big] du.$$

Similarly to the independence of quantities at the nodes, processes D_i and D_j are independent if $\mathcal{R}_i$ and $\mathcal{R}_j$ are disjoint. For instance, D_4 is independent of D_3 and D_5.

3.14 Cox Processes

This section describes a Poisson process with a random intensity measure. The random intensity might represent a random environment or field that influences the Poisson locations of points. Because the intensity is random, the resulting process is rather general, but we will show that many of its properties can be characterized by features of the parent Poisson process and the intensity process. Chapter 5 covers similar material for Brownian motion in a random environment.

Suppose that N is a point process on a space S and η is a random measure on S that is locally finite a.s. and is defined on the same probability space as N. The N is a *Cox process directed by* η if, conditioned on η, the N is a Poisson process with conditional intensity measure η. Equivalently, the conditional Laplace functional of N given η is

$$E[e^{-Nf} \,|\, \eta] = \exp\{-\int_S (1 - e^{-f(x)})\eta(dx)\}, \quad \text{a.s.} \qquad f \in C_K^+(S). \tag{3.31}$$

In particular,

$$P\{N(B) = n \,|\, \eta\} = e^{-\eta(B)}\eta(B)^n/n!,$$

and taking expectations,

$$P\{N(B) = n\} = E\Big[e^{-\eta(B)}\eta(B)^n/n!\Big].$$

Note also that $E[N(B)] = E[\eta(B)]$.

A Cox process is sometimes called a conditional Poisson process, a doubly stochastic Poisson process, or a Poisson process in a randomly changing environment. Several characterizations of Cox processes appear in [61].

An important observation is that a Cox process on $\mathbb{R}_+$ can be characterized as a homogeneous Poisson process with a random time parameter.

Remark 53. A point process N on $\mathbb{R}_+$ is a Cox process directed by η if and only if $N(\cdot) \stackrel{d}{=} N_1(\eta'(\cdot))$, where N_1 and η' are defined on a common probability space such that N_1 is a Poisson process with rate 1 and $\eta' \stackrel{d}{=} \eta$. This follows by the definition above and consideration of the Laplace functionals of the processes. In case η is strictly increasing, one can show as in Exercise 43 in Chapter 5 that $N(\cdot) = N_1(\eta(\cdot))$ a.s., where N_1 is defined on the same space as X and η.

In some instances, one can formulate a Cox process as a transformation of a Poisson process that has one more layer of randomness than those above. For instance, let N be a Poisson process on $\mathbb{R}_+$ with intensity measure μ. Consider a transformation of N in which a point of N at t is mapped to a location $\gamma(t)$, where $\gamma(t)$ is a stochastic process on $\mathbb{R}_+$ that is independent of N. Then the transformed process $N'(B) = M(\gamma^{-1}(B))$ is a Cox process directed by $\eta(B) = \int_{\mathbb{R}_+} \mathbf{1}(\gamma(t) \in B)\mu(dt)$, provided this is a.s. locally finite. This follows since, by Theorem 32, N' is Poisson when $\gamma(t)$ is deterministic.

As a second example, let N be a Poisson process on $\mathbb{R}_+$ with intensity measure μ. Assume that N is partitioned into m subprocesses $N_1, \ldots, N_m$ by the rule that a point of N at t is assigned to the subprocess with the label $\alpha(t)$, where $\alpha(t)$ is a stochastic process on $\{1, \ldots, m\}$ that is independent of N. Then as in Proposition 41, $N_1, \ldots, N_m$ are conditionally independent Poisson processes given $\alpha(\cdot)$. Hence each N_i is a Cox process directed by

$$\eta(B) = \int_B \mathbf{1}(\alpha(t) = i)\mu(dt).$$

Of course, the N_i are not independent since they all depend on α.

Because Cox processes are essentially Poisson processes with an extra expectation to account for the randomized mean measure, most results for Poisson processes have counterparts for Cox processes. Here is one instance.

Example 54. If $N_1, \ldots, N_m$ are Cox processes on S directed by $\eta_1, \ldots, \eta_m$, respectively, and $(N_1, \eta_1), \ldots, (N_m, \eta_m)$ are independent, then $N = N_1 + \cdots N_m$ is a Cox process directed by $\eta = \eta_1 + \cdots \eta_m$.

Here is an illustration of a Cox input process for a $M_t/G_t/\infty$ system.

Example 55. Regenerative-Modulated Poisson Process. In computer and telecommunications systems, a standard model for the occurrences of an event in time is a Cox process N on $\mathbb{R}_+$ directed by $\eta(t) = \int_0^t f(Y(s))\, ds$, where $Y(t)$ is a regenerative process on a countable state space S that models a changing environment in which events occur. For instance, a flow of data

may be Poisson, but dependent on an environment (type or source of the data, congestion in a network, etc.) that is changing according to $Y(t)$.

Since the Cox process has the form $N(t) = N_1(\eta(t))$, its behavior far out in time is related to the limiting behavior of $Y(t)$. For instance, suppose $Y(t)$ has a limiting distribution p on S. Then the limiting average of N is

$$t^{-1}N(t) \to \lambda = \sum_{i \in S} f(i)p_i, \quad \text{a.s. as } t \to \infty.$$

This follows since by the SLLNs for Poisson processes and regenerative processes, $N_1(t)/t \to 1$ and $\eta(t)/t \to \lambda$ a.s., and so

$$t^{-1}N(t) = (\eta(t)/t)N_1(\eta(t))/\eta(t) \to \lambda, \quad \text{a.s. as } t \to \infty.$$

Now, consider a variation of the $M_t/G/\infty$ system in which items arrive for service according to the preceding Cox process N, and G is the distribution of the independent service times. The system data is $M = \sum_n \delta_{(T_n, V_n)}$ on $\mathbb{R}^2_+$, where T_n is the arrival time of the nth item, and V_n is its sojourn or service time. Analogously to marked Poisson processes, M is a space-time Cox process directed by η with

$$E[M([0,t) \times [0,v))|\eta] = \int_{(0,t]} G_s(v) f(Y(s)) ds.$$

Here $d\eta(s) = f(Y(s))ds$.

Consider the quantity of items $Q(t) = \sum_n \mathbf{1}(T_n + V_n > t)$ in the system at time t. As in (3.24), the "conditional" distribution of $Q(t)$ given η is Poisson with

$$E[Q(t)|\eta] = \int_0^t [1 - G(t-s)] f(Y(s)) ds.$$

Furthermore, under the additional assumptions that $Y(t)$ is stationary and G has a mean α, we have

$$\lim_{t \to \infty} P\{Q(t) = n\} = \sum_{i \in S} P\{Y(0) = i\} (f(i)\alpha)^n e^{-f(i)\alpha}/n!. \tag{3.32}$$

This limit is a conditional Poisson distribution with random mean $\alpha f(Y(0))$.

To prove (3.32), note that

$$P\{Q(t) = n\} = E\Big[(E[Q(t)|\eta])^n e^{-E[Q(t)|\eta]}/n!\Big],$$

$$E[Q(t)|\eta] \stackrel{d}{=} f(Y(0)) \int_0^t [1 - G(u)] du \stackrel{d}{\to} \alpha f(Y(0)).$$

In light of these properties, (3.32) follows from the dominated convergence theorem for convergence in distribution (see the Appendix).

3.15 Compound Poisson Processes

If a Poisson process has real-valued marks at its points, then the cumulative value of the marks in time (or in a space) is a compound Poisson process. This section is a brief description of such processes.

We first consider a classical compound Poisson process in time (as mentioned in Example 14).

Definition 56. Let $N(t)$ be a homogeneous Poisson process on $\mathbb{R}_+$ with rate λ, and let Y_n be real-valued random variables that are i.i.d. with distribution F and are independent of N. The stochastic process

$$Z(t) = \sum_{n=1}^{N(t)} Y_n, \qquad t \geq 0,$$

is a *compound Poisson process* with rate λ and distribution F.

The name comes from the fact that $Z(t)$ has a compound Poisson distribution with rate λt and distribution F:

$$P\{Z(t) \leq z\} = \sum_{n=0}^{\infty} e^{-\lambda t}(\lambda t)^n F^{n\star}(z)/n!, \qquad z \in \mathbb{R}. \tag{3.33}$$

Indeed, condition on $N(t)$ and use $P\{Z(t) \leq z | N(t) = n\} = F^{n\star}(z)$. Similar conditioning on $N(t)$ yields

$$E[Z(t)] = \lambda t E[Y_1], \qquad \mathrm{Var}[Z(t)] = \lambda t E[Y_1^2],$$

provided these moments exist. Finally, if the moment generating function $\phi(\alpha) = E[e^{\alpha Y_1}]$ exists for some α in a neighborhood of 0, then

$$E[e^{\alpha Z(t)}] = e^{-\lambda t(1-\phi(\alpha))}.$$

The stationary independent increments of the Poisson process N and the i.i.d. property of its marks yield the following result.

Theorem 57. *A compound Poisson process $\{Z(t) : t \geq 0\}$ has stationary, independent increments: $Z(t_1) - Z(s_1), \ldots, Z(t_n) - Z(s_n)$, for $s_1 < t_1 < \cdots s_n < t_n$, are independent; and*

$$Z(s+t) - Z(s) \stackrel{d}{=} Z(t), \qquad s, t \geq 0.$$

Proof. Using the process $M = \sum_n \delta_{(T_n, Y_n)}$, we can write

$$Z(t) = \sum_n Y_n \mathbf{1}(T_n \leq t) = \int_{\mathbb{R}} y M((0,t] \times dy).$$

Under the assumptions, M is a space-time Poisson process with

$$E[M((s, s+t] \times B)] = \lambda t F(B).$$

Now, for any $s_1 < t_1 < \ldots s_n < t_n$, consider the increments

$$Z(t_i) - Z(s_i) = \int_{\mathbb{R}} yM((s_i, t_i] \times dy), \qquad 1 \le i \le n.$$

They are independent since the point processes $M((s_i, t_i] \times \cdot)$, for $1 \le i \le n$, on $\mathbb{R}$ are independent, because M has independent increments.

Next, note that $E[M((s, s+t] \times B)] = E[M((0, t] \times B)]$. Since M is Poisson and its distribution is uniquely determined by its intensity, it follows that $M((s, s+t] \times \cdot)] \stackrel{d}{=} M((0, t] \times \cdot)$ Consequently,

$$Z(s+t) - Z(s) = \int_{\mathbb{R}} yM((s, s+t] \times dy) \stackrel{d}{=} \int_{\mathbb{R}} yM((0, t] \times dy) = Z(t).$$

Hence $Z(t)$ has stationary increments.

The classical compound Poisson process described above has several natural extensions. For instance, instead of the Y_n being independent of N, suppose Y_n are p-marks of T_n. Also, assume N has a general intensity μ. Then the process $Z(t) = \sum_{n=1}^{N(t)} Y_n$, for $t \ge 0$, is a *location-dependent compound Poisson process* with intensity measure μ and distribution $p(t, \cdot)$. Many of its properties follow directly from the fact that $M = \sum_n \delta_{(T_n, Y_n)}$ is a Poisson process. For instance, see Exercises 54 and 55. Also, since M is Poisson, the results above for Poisson processes extend to compound Poisson processes by using a p-marking of M, which would be a "second" marking of N as mentioned in Example 39. Exercise 54 illustrates these ideas for partitions of compound Poisson processes.

There are other relatives of compound Poisson processes of the form $M(A) = \sum_n Y_n \delta_{X_n}(A)$, where $N = \sum_n \delta_{X_n}$ is a Poisson process on a general space, and the marks Y_n are random vectors, matrices, or elements of a group with an addition operation. Here is an example when Y_n are point processes.

Example 58. Poisson Cluster Processes. Let $N = \sum_n \delta_{X_n}$ denote a Poisson process on a general space S with intensity measure μ. Suppose that each point X_n generates a cluster of points in a space S' that are represented by a point process N'_n. Assume the N'_n are point processes on a space S' that are i.i.d. and independent of N. Then the number of points from the processes N'_n in a set B that are generated by points of N in the set A is

$$M(A \times B) = \sum_n N'_n(B) \delta_{X_n}(A).$$

This defines a point process M on $S \times S'$ called a *marked cluster process* generated by the Poisson process N; the $M(S \times \cdot)$, provided it is locally finite, is simply the cluster process on S'.

Since $M(A \times B) = \sum_{n=0}^{N(A)} N'_n(B)$, it follows by conditioning on N that $M(A \times B)$ has the compound Poisson distribution

$$P\{M(A \times B) \le n\} = \sum_{k=0}^{\infty} e^{-\mu(A)} \mu(A)^k F^{k\star}(n; B)/k!,$$

where $F(n; B) = P\{N'_1(B) \le n\}$. Also, $E[M(A \times B)] = E[N(A)]E[N'_1(B)]$ and

$$\mathrm{Var}[M(A \times B)] = E[N(A)]\mathrm{Var}[N'_1(B)].$$

More general cluster processes, where the N'_n are marks of X_n, are analyzed in Exercise 57.

3.16 Poisson Law of Rare Events

Poisson processes are natural models for rare events in time, or rare points in a space. This is partly due to a law of rarely occurring events in which a sum of thin or rarefied point processes converges in distribution to Poisson process. A classical case for random variables is a Binomial random variable converging to a Poisson random variable as in Example 59 below. In this section, we present a generalization of this result that gives conditions under which a sum of many rare indicator random variables converges to a Poisson random variable. Analogous results under which a sum of point processes converges to a Poisson process are in the next section.

Here is a classical example of the Poisson law of rare events.

Example 59. Binomial Convergence to Poisson. Suppose $Y_{n1}, \ldots, Y_{nn}$ are independent Bernoulli random variables with $P\{Y_{ni} = 1\} = p_n$. Then $Z_n = \sum_{i=1}^{n} Y_{ni}$ has a binomial distribution with parameters n and p_n. If $np_n \to \mu > 0$ as $n \to \infty$, then $Z_n \xrightarrow{d} Z$, where Z has a Poisson distribution with mean μ. This is a special case of the following result.

Theorem 60. (Poisson Law of Rare Events) *Suppose $Y_{n1}, Y_{n2}, \ldots$, for $n \ge 1$, are a countable number of independent random variables that take values* 0 *or* 1*, and satisfy the* uniformly null *property*

$$\sup_i P\{Y_{ni} = 1\} \to 0, \quad \text{as } n \to \infty. \tag{3.34}$$

Let Z be a Poisson random variable with mean μ. Then as $n \to \infty$,

$$Z_n = \textstyle\sum_i Y_{ni} \xrightarrow{d} Z \quad \text{if and only if} \quad \sum_i P\{Y_{ni} = 1\} \to \mu.$$

Proof. We will use the property of Laplace transforms that $Z_n \xrightarrow{d} Z$ if and only if $E[e^{-\alpha Z_n}] \to E[e^{-\alpha Z}]$. By the independence of the Y_{ni},

$$E[e^{-\alpha Z_n}] = \prod_i E[e^{-\alpha Y_{ni}}] = \prod_i (1 - c_{ni}), \qquad \alpha \geq 0,$$

where $c_{ni} = E[1 - e^{-\alpha Y_{ni}}]$. Also, $E[e^{-\alpha Z}] = e^{-c}$, where $c = \mu(1 - e^{-\alpha})$, since Z has a Poisson distribution with mean μ. From these observations,

$$Z_n \xrightarrow{d} Z \quad \Longleftrightarrow \quad E[e^{-\alpha Z_n}] \to e^{-c} \quad \Longleftrightarrow \quad \prod_i (1 - c_{ni}) \to e^{-c}. \tag{3.35}$$

Moreover, under the assumption (3.34) and $c_{ni} \leq (1 - e^{-\alpha}) < 1$, it follows by Lemma 61 below for $c_{ni} = c\mu^{-1} P\{Y_{ni} = 1\}$, that

$$\prod_i (1 - c_{ni}) \to e^{-c} \quad \Longleftrightarrow \quad \sum_i c_{ni} \to c \quad \Longleftrightarrow \quad \sum_i P\{Y_{ni} = 1\} \to \mu.$$

Combining this string of equivalences with (3.35) proves the assertion.

The preceding proof of the Poisson convergence boils down to the following result on the convergence of real numbers.

Lemma 61. *Suppose $c_{n1}, c_{n2}, \ldots$, for $n \geq 1$, are a countable (possibly finite) number of real numbers in $(0, a]$, where $a < 1$, that satisfy the* uniformly null *property*

$$\sup_i c_{ni} \to 0 \qquad \text{as } n \to \infty. \tag{3.36}$$

Then, for any $c > 0$,

$$\lim_{n\to\infty} \prod_i (1 - c_{ni}) = e^{-c} \quad \text{if and only if} \quad \lim_{n\to\infty} \sum_i c_{ni} = c.$$

Proof. The assertion is equivalent to

$$\overline{s}_n = -\sum_i \log(1 - c_{ni}) \to c \quad \text{if and only if} \quad s_n = \sum_i c_{ni} \to c. \tag{3.37}$$

Since $\log(1 - c_{ni}) = -\sum_{m=1}^{\infty} c_{ni}^m / m$, the difference in these sums is

$$d_n = \overline{s}_n - s_n = \sum_i c_{ni}^2 \sum_{m=2}^{\infty} c_{ni}^{m-2}/m.$$

Using $c_{ni} \leq a$ and $\alpha_n = \sup_i c_{ni}$, we have

$$d_n \leq \frac{1}{1-a} \sum_i c_{ni}^2 \leq \frac{\alpha_n s_n}{1-a} \leq \frac{\alpha_n \overline{s}_n}{1-a}. \tag{3.38}$$

Now, if $\overline{s}_n \to c$, then from (3.38) and (3.36) we have $d_n \to 0$, and hence $s_n = \overline{s}_n - d_n \to c$. Similarly, if $s_n \to c$, then $\overline{s}_n = s_n + d_n \to c$. These observations prove (3.37).

3.17 Poisson Convergence Theorems*

This section contains Poisson convergence theorems for sequences of point processes. These results are extensions of the Poisson law of rare events in Theorem 60 above. The main theorem here is that a sum of many independent sparse point processes converges to a Poisson process. Consequently, certain sums of renewal processes and rare transformations of a point process converge to a Poisson process. Also included are examples justifying that Poisson processes are reasonable approximations for thinnings and partitions of a point process.

We will use the following notion of weak convergence, which is reviewed in the Appendix. Suppose $\mu, \mu_1, \mu_2, \ldots$ are probability measures on S. The probabilities μ_n *converge weakly* to μ as $n \to \infty$, denoted by $\mu_n \xrightarrow{w} \mu$, if $\mu_n f \to \mu f$, as $n \to \infty$, for each bounded continuous function $f : S \to \mathbb{R}$ (recall $\mu f = \int_S f(x)\mu(dx)$). This is equivalent to

$$\lim_{n\to\infty} \mu_n(B) = \mu(B), \qquad B \in \hat{\mathcal{S}}_\mu, \tag{3.39}$$

where $\hat{\mathcal{S}}_\mu = \{B \in \hat{\mathcal{S}} : \mu(\partial B) = 0\}$, the set of all bounded sets whose boundary has μ-measure 0.

A sequence of random elements in a metric space converges in distribution to a random element if their distributions converge weakly. In particular, a sequence of point processes N_n on S *converges in distribution to* N as $n \to \infty$, denoted by $N_n \xrightarrow{d} N$, if $P\{N_n \in \cdot\} \xrightarrow{w} P\{N \in \cdot\}$. This weak convergence is equivalent to the convergence of the finite-dimensional distributions (condition (ii) in the next theorem).

A few points in our analysis use the slightly more general notion of vague convergence of measures. Suppose $\mu, \mu_1, \mu_2, \ldots$ are locally finite measures on S. The measures μ_n *converge vaguely* to μ, denoted by $\mu_n \xrightarrow{v} \mu$, if

$$\mu_n f \to \mu f, \quad \text{as } n \to \infty, \text{ for each } f \in C_K^+(S).$$

This is equivalent to (3.39) if all the measures are probability measures, so vague convergence, in this case, is the same as weak convergence.

The following are several equivalent conditions for point processes to converge in distribution. Here $\hat{\mathcal{S}}_N = \{B \in \hat{\mathcal{S}} : N(\partial B) = 0 \text{ a.s.}\}$.

Theorem 62. *For point processes $N, N_1, N_2, \ldots$ on S, the following statements are equivalent as $n \to \infty$.*

(i) $N_n \xrightarrow{d} N$.
(ii) $(N_n(B_1), \ldots, N_n(B_k)) \xrightarrow{d} (N(B_1), \ldots, N(B_k)), \quad B_1, \ldots, B_k \in \hat{\mathcal{S}}_N$.
(iii) $N_n f \xrightarrow{d} Nf, \qquad f \in C_K^+(S)$.
(iv) $E[e^{-N_n f}] \to E[e^{-Nf}], \qquad f \in C_K^+(S)$.

A proof of this result is in [60]. Condition (ii) says the finite-dimensional distributions of N_n converge to those of N. When S is an Euclidean space, the sets B_i can be replaced by bounded rectangles. Condition (iii) relates the convergence in distribution of integrals with respect to point processes to the convergence of the processes. The convergence (iv) of Laplace functionals is a convenient tool for proving $N_n \xrightarrow{d} N$, when the functionals can be factored conveniently (as in the proof of Theorem 66 below).

Here is an elementary but useful fact. It justifies the convergence of a Poisson process when its intensity converges (see Exercise 54 in Chapter 4).

Proposition 63. (Convergence of Poisson Processes) *For each $n \geq 1$, suppose N_n is a Poisson process on a space S with intensity measure μ_n. If $\mu_n \xrightarrow{v} \mu$ and μ is locally finite, then $N_n \xrightarrow{d} N$, where N is a Poisson process with intensity μ.*

Proof. From Theorem 21, we know that $E[e^{-N_n f}] = e^{-\mu_n h}$, for $f \in C_K^+(S)$, where $h(x) = 1 - e^{-f(x)}$. By Theorem 62, we have $\mu_n h \xrightarrow{v} \mu h$, and so

$$E[e^{-N_n f}] = e^{-\mu_n h} \to e^{-\mu h} = E[e^{-Nf}].$$

Thus, $N_n \xrightarrow{d} N$ by Theorem 62.

We are now ready to consider the convergence of sums of point processes. Here is a motivating example.

Example 64. Consider a sum $N(t) = \sum_{i=1}^n N_i(t)$, for $t \geq 0$, where $N_1, \ldots, N_n$ are independent renewal processes. Of course, N is generally not a renewal process. However, suppose the times between renewals for each process N_i tend to be large (i.e., $F_i(t)$ is small, where F_i is the inter-renewal distribution). Consequently, each contribution $N_i(a, b]$ to $N(a, b]$ would tend to be 0. In other words, each N_i rarely contributes a point to N on bounded intervals. However, if the number n of these contributions is large, it might be reasonable to approximate N by a Poisson process with intensity $E[N(t)] = \sum_{i=1}^n E[N_i(t)]$.

A Poisson convergence theorem justifying such an approximation is as follows. The opposite situation in which $N_i(a, b]$ tends to be large is addressed in Exercise 59.

Theorem 65. (Sums of Renewal Processes) *For $n \geq 1$, let $N_n(t) = \sum_i N_{ni}(t)$ be a point process on $\mathbb{R}_+$, where $N_{n1}, N_{n2}, \ldots$ is a finite or countable number of independent renewal processes and N_{ni} has inter-renewal distribution F_{ni}. Assume the inter-renewal times are uniformly rare in that*

$$\lim_{n\to\infty} \sup_i F_{ni}(t) = 0, \qquad t \geq 0.$$

Let N be a Poisson process on $\mathbb{R}_+$ with intensity measure μ. Then $N_n \xrightarrow{d} N$, as $n \to \infty$ if and only if, for each t with $\mu(\{t\}) = 0$,

$$\lim_{n\to\infty} \sum_i F_{ni}(t) = \mu(t). \tag{3.40}$$

Proof. This result follows by Theorem 66 below, since

$$\begin{aligned}\sum_i P\{N_{ni}(t) \geq 2\} &= \sum_i \int_{(0,t]} F_{ni}(t-s) F_{ni}(ds) \\ &\leq \sup_i F_{ni}(t) \sum_i F_{ni}(t),\end{aligned}$$

and (3.43) is the same as (3.40) because $P\{N_{ni}(t) \geq 1\} = F_{ni}(t)$.

The next result is a general Poisson convergence theorem for sums of uniformly rare point processes. Suppose that

$$N_n = \sum_i N_{ni}, \qquad n \geq 1,$$

is a point process on a space S, where $N_{n1}, N_{n2}, \ldots$ is a countable number of independent point processes on S. Assume the point processes N_{ni} are *uniformly null*, meaning that

$$\lim_{n\to\infty} \sup_i P\{N_{ni}(B) \geq 1\} = 0, \qquad B \in \hat{\mathcal{S}}. \tag{3.41}$$

Let N be a Poisson process on S with intensity measure μ.

Theorem 66. (Grigelionis) *For the processes defined above, $N_n \xrightarrow{d} N$, as $n \to \infty$ if and only if*

$$\lim_{n\to\infty} \sum_i P\{N_{ni}(B) \geq 2\} = 0, \qquad B \in \hat{\mathcal{S}}, \tag{3.42}$$

$$\lim_{n\to\infty} \sum_i P\{N_{ni}(B) \geq 1\} = \mu(B), \qquad B \in \hat{\mathcal{S}}_\mu. \tag{3.43}$$

Proof. The convergence $N_n \xrightarrow{d} N$ is equivalent, by Theorem 62, to

$$E[e^{-N_n f}] \to E[e^{-Nf}], \qquad f \in C_K^+(S). \tag{3.44}$$

Using the independence of the N_{ni} and letting $c_{ni} = E[1 - e^{-N_{ni} f}]$, we have

$$E[e^{-N_n f}] = \prod_i E[e^{-N_{ni} f}] = \prod_i (1 - c_{ni}).$$

Also, by Theorem 21, $E[e^{-Nf}] = e^{-\mu h}$, where $h(x) = 1 - e^{-f(x)}$. Combining these observations, it follows that (3.44) is equivalent to

$$\prod_i (1 - c_{ni}) \to e^{-\mu h}, \qquad f \in C_K^+(S). \tag{3.45}$$

Keep in mind that c_{ni} is a function of f.

We will complete the proof by applying Lemma 61 to establish that (3.42) and (3.43) are necessary and sufficient for (3.45). First note that

$$c_{ni} = E[1 - e^{-N_{ni}f}] \le P\{N_{ni}(S_f) \ge 1\},$$

where S_f is the support of f. Then (3.41) implies

$$\sup_i c_{ni} \le \sup_i P\{N_{ni}(S_f) \ge 1\} \to 0.$$

In light of this property, we can assume c_{ni} are in $(0, a]$ for some $a < 1$. Then Lemma 61 says that (3.45) is equivalent to

$$\sum_i E[1 - e^{-N_{ni}f}] = \sum_i c_{ni} \to \mu h, \qquad f \in C_K^+(S). \tag{3.46}$$

Therefore, it remains to show that (3.42) and (3.43) are necessary and sufficient for (3.46).

To prove that (3.42) and (3.43) imply (3.46), consider

$$\begin{aligned} \sum_i c_{ni} = \sum_i E[(1 - e^{-N_{ni}f})\mathbf{1}(N_{ni}(S_f) = 1)] \\ + \sum_i E[(1 - e^{-N_{ni}f})\mathbf{1}(N_{ni}(S_f) \ge 2)]. \end{aligned} \tag{3.47}$$

The last sum is bounded by $\sum_i P\{N_{ni}(S_f) \ge 2)\}$ which converges to 0 by assumption (3.42). The first sum on the right-hand side in (3.47) equals $\eta_n h$, where

$$\eta_n(B) = \sum_i E[N_{ni}(B \cap S_f)\mathbf{1}(N_{ni}(S_f) = 1)] = \sum_i P\{N_{ni}(B \cap S_f) = 1\}.$$

The last sum has the same form as the sum in (3.42) minus the one in (3.43), and so these assumptions imply $\eta_n \xrightarrow{v} \mu$, which yields $\eta_n h \to \mu h$. Using the preceding observations in (3.47) proves that (3.42) and (3.43) are sufficient for (3.46).

Conversely, suppose (3.46) is true. Applying this property to the function $f(x) = -\mathbf{1}(x \in B) \log s$, where $B \in \hat{\mathcal{S}}$ and $s \in [0, 1]$, we have

$$H_n(s) = \sum_i E[1 - s^{N_{ni}(B)}] \to (1 - s)\mu(B), \tag{3.48}$$

since $h(x) = 1 - e^{-f(x)} = (1-s)\mathbf{1}(x \in B)$. Then (3.43) follows since

$$\sum_i P\{N_{ni}(B) \geq 1\} = H_n(0) \to \mu(B). \tag{3.49}$$

Next, consider the factorization

$$\begin{aligned} H_n(s) &= \sum_i [1 - \sum_{m=0}^{\infty} s^m P\{N_{ni}(B) = m\}] \\ &= (1-s)H_n(0) + \sum_i \sum_{m=2}^{\infty} (s - s^m) P\{N_{ni}(B) = m\}. \end{aligned}$$

This expression along with (3.48) and (3.49) yield

$$(s - s^2) \sum_i P\{N_{ni}(B) \geq 2\} \leq H_n(s) - (1-s)H_n(0) \to 0.$$

Thus (3.46) is true. These observations prove that (3.46) implies (3.42) and (3.43), which completes the proof.

Theorem 66 justifies Poisson limits for sums of independent renewal processes (Theorem 65 and Exercise 59). Although Theorem 66 is for sums of independent point processes, it also applies to certain sums of conditionally independent point processes. We will consider the convergence of such sums and their application to partitioning and thinning of point processes, following a motivating example.

Example 67. A Thinned Process. Let N be a point process on $\mathbb{R}_+$ (e.g., a renewal process) that satisfies $t^{-1}N(t) \xrightarrow{d} \lambda$ as $t \to \infty$, where λ is a positive constant. Suppose each point of N is independently retained with probability p and deleted with probability $1-p$. Let $N_p(t)$ denote the number of retained points in $(0, t]$. When p is very small, the retained points are rare and so it appears that it would be appropriate to approximate the p-thinning N_p of N by a Poisson process.

Let us see why. As $p \to 0$, clearly N_p would converge to 0. However, the thinned process $N_p(p^{-1}t)$ with its time scale magnified by p^{-1} converges in distribution to a Poisson process with rate λ. This convergence is a special case of Corollary 70 below, which applies to general partitions of a point process. Based on this convergence it follows that it is reasonable to approximate N_p by a Poisson process with rate $p\lambda$ when p is small.

For the next result, suppose that $N_n = \sum_j \delta_{X_{nj}}$ is a sequence of point processes on a space S with intensity measures μ_n. Let M_n be a marked p_n-transformation of N_n on $S \times S'$. We will specify conditions for the convergence of M_n to a Poisson process, and then apply this to partitions of a point process.

We will use the conditional mean measure of M_n given N_n, which is

$$\eta_n(A \times B) = E[M_n(A \times B)|N_n] = \sum_j p_n(X_{nj}, B)\mathbf{1}(X_{nj} \in A)$$
$$= \int_A p_n(x, B)N_n(dx), \quad A \in \mathcal{S},\ B \in \mathcal{S}'.$$

The convergence in distribution of these random mean measures η_n would be consistent with the convergence of M_n. For such random measures, the convergence $\eta_n \xrightarrow{d} \eta$ is analogous to convergence in distribution of point processes; equivalent statements for this are in Theorem 62 (with η in place of N).

A natural prerequisite for M_n to converge is that the transformations should be uniformly null. Accordingly, we will use the condition

$$\lim_{n\to\infty} \sup_{x\in A} p_n(x, B) = 0, \quad A \in \hat{\mathcal{S}},\ B \in \hat{\mathcal{S}}'. \tag{3.50}$$

Theorem 68. (Poisson Limit of Rare Transformations) *Suppose the sequence M_n of marked p_n-transformations of N_n satisfies (3.50). Also, assume $\eta_n \xrightarrow{d} \mu$ as $n \to \infty$, where μ is a (non-random) locally finite measure on $S \times S'$. Then $M_n \xrightarrow{d} M$ as $n \to \infty$, where M is a Poisson process on $S \times S'$ with intensity measure μ.*

Proof. We can write $M_n = \sum_i M_{ni}$, where $M_{ni} = \delta_{X_{ni},Y_{ni}}$ and Y_{ni} are p_n-marks of the X_{ni}, for $i \geq 1$. Although the point processes M_{ni}, $i \geq 1$, are not independent, they are conditionally independent given N_n. Clearly $P\{M_{ni}(B) \geq 2|N_n\} = 0$ and, under assumption (3.50),

$$\sup_i P\{M_{ni}(A \times B) \geq 1|N_n\} \leq \sup_{x\in A} p_n(x, B) \to 0, \quad A \in \hat{\mathcal{S}},\ B \in \hat{\mathcal{S}}'.$$

Also, using $\eta_n \xrightarrow{d} \mu$, we have, for $A \times B \in \widehat{\mathcal{S} \times \mathcal{S}'}_\mu$,

$$\sum_i P\{M_{ni}(A \times B) \geq 1|N_n\} = E[M_n(A \times B)|N_n]$$
$$= \eta_n(A \times B) \xrightarrow{d} \mu(A \times B).$$

Applying Theorem 66 to the conditional distribution of M_n given N_n, and using Theorem 62, it follows that

$$E[e^{-M_n f}|N_n] \xrightarrow{d} E[e^{-Mf}], \qquad f \in C_K^+(S).$$

Taking expectations of this and using the dominated convergence theorem for convergence in distribution (Theorem 17 in the Appendix), we have $E[e^{-M_n f}] \to E[e^{-Mf}]$. Thus, $M_n \xrightarrow{d} M$ by Theorem 62.

Example 69. Poisson Limits of Partitions. Let $N(t)$ be a point process on $\mathbb{R}_+$. Suppose N is partitioned as in Proposition 41 by the following rule: Each point of N is assigned to subprocess $i \in I$ (a countable set) with probability $p(i)$, independent of everything else, where $\sum_{i\in I} p(i) = 1$. Then $N = \sum_{i\in I} N_i$, where N_i denotes the ith subprocesses in the partition.

We address the issue of finding conditions under which the subprocesses $\{N_i : i \in I_0\}$, for a subset $I_0 \subseteq I$, are approximately independent Poisson processes. A natural condition for this is that the points of N would rarely be assigned to the subprocesses in I_0, but would mostly be assigned to the other subprocesses. The thinning in Example 67 is such a partition consisting of two subprocesses, where $I_0 = \{0\}$ and $I = \{0, 1\}$.

To ensure that the subprocesses in I_0 are sparse, we assume the partitioning probabilities are functions of n such that $p_n(i) \to 0$, for $i \in I_0$ (I is necessarily infinite when $I_0 = I$). Denote the ith subprocess by $N_{ni}(t)$. Its conditional mean given N is $E[N_{ni}(t)|N] = p_n(i)N(t)$. This mean converges to 0, which would not lead to a non-zero limit of N_{ni}.

To obtain a non-zero limit, a normalization of the processes N_{ni} is in order. Accordingly, assume there is a positive constant λ such that

$$t^{-1}N(t) \xrightarrow{d} \lambda. \tag{3.51}$$

This ensures that $N(t) \to \infty$ and that the points of N appear at a positive rate out to infinity. Next, assume the partitioning is uniformly rare on I_0: there exist positive constants $a_n \to \infty$ and $r(i)$, such that

$$\lim_{n\to\infty} a_n p_n(i) = r(i), \qquad i \in I_0. \tag{3.52}$$

Under the preceding assumptions, it is natural to consider the convergence of the point process

$$\hat{N}_{ni}(t) = N_{ni}(a_n t), \qquad i \in I_0.$$

This is a normalization of the partition-processes N_{ni} under a rescaling of time so that a_n is the new unit of time. The $\hat{N}_{ni}$ on a "fixed" interval $(0, t]$ represents subprocess i on the interval $(0, a_n t]$, which becomes larger as $n \to \infty$. The choice of a_n for the time unit is because, as $n \to \infty$,

$$E[\hat{N}_{ni}(t)|N] = a_n p_n(i)(N(a_n t)/a_n) \xrightarrow{d} r(i)\lambda, \qquad i \in I_0.$$

The following result describes the Poisson limits of the subprocesses. Interestingly, the processes $\hat{N}_{ni}$ for $i \in I_0$ are dependent for each n but in the limit they are independent. This convergence theorem justifies that the partition-processes N_{ni}, $i \in I_0$, for large n are approximately independent Poisson processes on $\mathbb{R}_+$ with respective rates $a_n p_n(i)\lambda \approx r(i)\lambda$, $i \in I_0$.

Corollary 70. *Under assumptions (3.51) and (3.52),*

$$(\hat{N}_{ni} : i \in I_0) \xrightarrow{d} (N_i : i \in I_0), \quad \text{as } n \to \infty, \tag{3.53}$$

where the limiting processes are independent homogeneous Poisson processes with respective rates $r(i)\lambda$, $i \in I_0$.

Proof. The partition of N we are studying is a special p_n-transformation of the process $N(a_n \cdot)$ with $p_n(t, B) = \sum_{i \in B} p_n(i)$. Specifically, the number of the $N(a_n t)$ points assigned to subprocess $i \in I_0$ is

$$\hat{N}_{ni}(t) = M_n((0, t] \times \{i\}),$$

where M_n is a marked p_n-transformation on $\mathbb{R}_+ \times I_0$ of $N(a_n \cdot)$ as in Theorem 68. For each $t \geq 0$ and $B \subseteq I_0$,

$$\sup_{s \leq t} p_n(s, B) = \sum_{i \in B} p_n(i) \to 0.$$

Furthermore, under assumptions (3.51) and (3.52),

$$\begin{aligned} \eta_n((0, t] \times B) \;&=\; E[M_n((0, t] \times B) | N(a_n \cdot)] \;=\; a_n \sum_{i \in B} p_n(i)(N(a_n t)/a_n) \\ &\xrightarrow{d} \eta((0, t] \times B) \;=\; \sum_{i \in B} r(i)\lambda t. \end{aligned}$$

Thus, the assumptions of Theorem 68 are satisfied, and so $M_n \xrightarrow{d} M$, where M is a Poisson process with $E[M((0, t] \times B)] = \sum_{i \in B} r(i)\lambda t$. Hence, assertion (3.53) follows since $\hat{N}_{ni}(t) = M_n((0, t] \times \{i\})$.

3.18 Exercises

The first nine exercises concern properties of exponential random variables that arise naturally in modeling. Each of these exercises has an analogue for geometric distributions (the rate of an exponential distribution is analogous to the probability (or parameter) p in a geometric distribution $p(1-p)^{n-1}$).

Exercise 1. *Memoryless Property.* Show that an exponential random variable X satisfies

$$P\{X > s + t | X > s\} = P\{X > t\}, \qquad s, t > 0.$$

Explain why this is called a memoryless property. Show that a nonnegative continuous random variable has an exponential distribution if and only if it satisfies the memoryless property. Hint: Use the fact that a continuous nonincreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies $f(s+t) = f(s)f(t)$, $s,\ t \geq 0$, if and only if f has the form $f(t) = e^{-ct}$ for some $c \geq 0$.

Exercise 2. *Minima of Exponential Random Variables.* Let $X_1, \ldots, X_m$ be independent exponential random variables with respective rates $\lambda_1, \ldots, \lambda_m$, and define

$$Y = \min_{1 \le i \le m} X_i, \qquad \nu = \text{argmin}_{1 \le i \le m} X_i.$$

Show that Y has an exponential distribution with rate $\lambda = \sum_{i=1}^m \lambda_i$, and

$$P\{\nu = i\} = \lambda_i / \lambda.$$

Show that

$$P\{X_1 < X_2 < \cdots < X_m\} = \prod_{k=1}^{m-1} \frac{\lambda_k}{\sum_{j=k}^m \lambda_j}.$$

Exercise 3. *Continuation.* In the context of the preceding exercise, prove that ν and Y are independent, and that

$$E[\min\{X_1, X_2\} | X_1 \le X_2] = 1/(\lambda_1 + \lambda_2),$$

$$P\{X_j - Y > x_j, j \ne i \,|\, \nu = i\} = \exp(-\sum_{j \ne i} \lambda_j x_j).$$

Show that $P\{\nu \in I | \nu \in J\} = \sum_{i \in I} \lambda_i / \sum_{j \in J} \lambda_j$, for $I \subseteq J$ in $\{1, \ldots, m\}$.

Exercise 4. Prove that a nonnegative distribution F is an exponential distribution if and only if $F(t) = \frac{1}{\mu} \int_0^t [1 - F(s)] ds$ for some $\mu > 0$. In that case, μ is the mean of F.

Exercise 5. *Geometric Sum of Exponential Random Variables.* Suppose $X_1, X_2, \ldots$ are independent exponentially distributed with rate λ, and ν is a random variable independent of the X_n with the geometric distribution $P\{N = n\} = p^{n-1}(1-p)$, $n \ge 1$. Prove that

$$P\{\sum_{i=1}^{\nu} X_i > t\} = e^{-(1-p)\lambda t}$$

by using Laplace transforms. Also, prove it by using the representation

$$P\{\sum_{i=1}^{\nu} X_i > t\} = P\{N(t) < \nu\},$$

and conditioning on $N(t)$, where $N(t)$ is a Poisson process with rate λ independent of ν. In addition, prove that

$$P\{\nu > m, \sum_{i=1}^{\nu - m} X_i > t\} = p^m e^{-(1-p)\lambda t}.$$

Exercise 6. Let $X_1, \ldots, X_n, Y$ be independent exponentially distributed random variables with respective rates $\lambda_1, \ldots, \lambda_n, \mu$. Show that

$$P\{\sum_{i=1}^{n} X_i < Y\} = \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \mu} = \prod_{i=1}^{n} P\{X_i < Y\}.$$

Show that if $N(t)$ is a Poisson process with rate λ, and T is an independent exponential random variable with rate μ, then $P\{N(T) > n\} = \Big(\lambda/(\lambda+\mu)\Big)^n$.

Exercise 7. *Exponential Series.* Let $X_1, X_2, \ldots$ be independent exponentially distributed random variables with respective rates $\lambda_1, \lambda_2, \ldots$ Show that

$$\sum_{n=1}^{\infty} X_n < \infty \text{ a.s.} \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} \lambda_n^{-1} < \infty.$$

(This property is used to prove Proposition 5 in the next chapter.) Hint: Use Laplace transforms and the property of products that, for $a_n \in (0, 1)$,

$$\prod_{n=1}^{\infty} a_n > 0 \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} (1 - a_n) < \infty.$$

Exercise 8. Show that if X is an exponential random variable with rate λ, then for any $h : \mathbb{R}_+ \to \mathbb{R}$,

$$E[h(X)] = \lambda E\Big[\int_0^X h(u)du\Big],$$

provided the expectations exist. (This property is used to prove Theorem 52 in the next chapter.)

Exercise 9. Suppose that F_i, $i \in I$, is a finite collection of exponential distributions with respective rates λ_i, $i \in I$, that are distinct. For $J \subseteq I$, let F_J denote the convolution of the distributions F_j, $j \in J$. Show that, for any real numbers a_i, and subsets J_i of I, for $j \in I$,

$$\sum_{k \in I} a_k F_{J_k}(x) = \sum_{i \in I} c_i F_i(x), \tag{3.54}$$

where

$$c_i = \lambda_i \sum_{k \in I} a_k 1(i \in J_k) \prod_{j \in J_k \setminus \{i\}} \frac{\lambda_j}{\lambda_j - \lambda_i}.$$

This assertion also holds when I is infinite, under the additional assumption that the summations exist. Hint: Use the fact that the Laplace transform of the left-hand side of (3.54) is

$$L(u) = \sum_{k \in I} a_k \prod_{j \in J_k} \frac{\lambda_j}{\lambda_j + u}.$$

A standard partial sum expansion of this sum of products is

$$L(u) = \sum_{i \in I} c_i \frac{\lambda_i}{\lambda_i + u}, \tag{3.55}$$

where $c_i = (\lambda_i + u)L(u)|_{u=-\lambda_i}$.

Exercise 10. A space station requires the continual use of two systems whose lifetimes are independent exponentially distributed random variables X_1 and X_2 with respective rates λ_1 and λ_2. When system 2 fails, it is replaced by a spare system whose lifetime X_3 is exponentially distributed with rate λ_3, independent of the other systems. Find the distribution of the time $Y = \min\{X_1, X_2 + X_3\}$ at which one of the systems becomes inoperative. Find the probability that system 1 will fail before system 2 (with its spare) fails.

Exercise 11. *Dispatching.* In Example 3, what properties of the Poisson process are not needed to obtain the optimal dispatching policy? What is the optimal policy when the arrival process $N(t)$ is a simple, stationary point process with $N(t) = \lambda t$, such as a stationary renewal process?

Exercise 12. *Waiting to be Dispatched.* Items arrive to a dispatching station according to a Poisson process with rate λ, and all items in the system will be dispatched at a time t. Example 3 shows that the expected time items wait before being dispatched at time t is $E[\int_0^t N(s)ds] = \lambda^2/2$. Suppose there is a cost hw^2 for holding an item in the system for a time w. Then the total holding cost in $(0, t]$ is $C = \sum_{n \geq 1} h(t - T_n)^2$. Find $E[C]$.

Exercise 13. Requests for a product arrive to a storage facility according to a Poisson process with rate λ per hour. Given that n requests are made in a t-hour time interval, find the probability that at least k requests were made in the first hour. Is this conditional probability different if the beginning of the one-hour period is chosen according to a probability density $f(s)$ on the interval $[0, t - 1]$?

Exercise 14. From Theorem 22, we know that the sum $N = N_1 + \cdots + N_n$ of independent Poisson processes is Poisson. Prove this statement by verifying that N satisfies the defining properties of a Poisson process.

Exercise 15. *Meeting of Vehicles.* Vehicles enter a one-mile stretch of a two-way highway at both ends by independent Poisson processes and move at 60 miles per hour to the opposite end. Let λ and μ denote the rate of Poisson arrivals at the two ends labeled 0 and 1. Assuming the highway is empty at time 0, show that the probability is $(\lambda e^{-\mu} + \mu e^{-\lambda})/(\lambda + \mu)$ that the first vehicle that enters at either end does not encounter another vehicle coming from the other direction during the one-mile stretch.

Show that this probability is $(\lambda e^{-\mu} + \mu e^{-(\lambda+2\mu)})/(\lambda+\mu)$ for such an encounter avoidance for the first vehicle to arrive from end 0. Hint: For this second problem, let $N_1(t)$ denote the Poisson process of arrivals at end 1, and let T denote the time of the first arrival at end 0. Then $N_1(T)$ denotes the number of arrivals at 1 before the first arrival at 0. Use its distribution from Exercise 6.

Exercise 16. Show that a simple point process N on $\mathbb{R}_+$ is a Poisson process with rate λ if and only if N has independent increments, $E[N(1)] = \lambda$, and, for any t and n, the conditional joint density of $T_1, \ldots, T_n$ given $N(t) = n$ is the same as the density of the order statistics of n independent uniformly distributed random variables on $[0, t]$.

Exercise 17. *Shot Noise Process.* Suppose that shocks (or pulses) to a system occur at times that form a Poisson process with rate λ. The shock at time T_n has a magnitude Y_n and this decays exponentially over time with rate γ. Then the cumulative effect of the shocks at time t is

$$Z(t) = \sum_{n=1}^{N(t)} Y_n e^{-\gamma(t-T_n)}.$$

Assume $Y_1, Y_2, \ldots$ are i.i.d. independent of N with mean μ and variance σ^2. Find expressions for the mean and variance of $Z(t)$ in terms of μ and σ^2.

Exercise 18. *Randomly Discounted Cash Flows.* In the context of Example 13, consider the generalization

$$Z(t) = \sum_{n=1}^{N(t)} Y_n e^{-\gamma_n T_n},$$

where $\gamma_1, \gamma_2, \ldots$ are independent nonnegative discount rates with distribution G that are independent of N and the Y_n's. Show that

$$E[Z(t)] = \lambda E[Y_1] \int_0^t \int_{\mathbb{R}_+} e^{-\gamma x} G(d\gamma) dx.$$

Exercise 19. Calls arrive to an operator at a call center at times that form a Poisson process $N(t)$ with rate λ. The time τ devoted to a typical call has an exponential distribution with rate μ, and it is independent of N. Then $N(\tau)$ is the number of calls that arrive while the operator is busy answering a call. Find the Laplace transform and variance of $N(\tau)$. Find $P\{N(\tau) \leq 1\}$.

Exercise 20. For a Poisson process N with rate λ, show that

$$E[T_\ell - T_k | N(t) = n] = (\ell - k)/(n+1), \qquad k < \ell \leq n,\ t > 0.$$

Find an expression for $E[t - T_k | N(t) = n]$.

Exercise 21. Two types of items arrive at a station for processing according to independent Poisson processes with respective rates λ and λ'. Let T_m and T'_n denote the times of the mth and nth arrivals from the two respective processes. Show that $P\{T_m < T'_n\} = P\{Y \geq m\}$, where Y has a binomial distribution with parameters $m + n - 1$ and $\lambda/(\lambda + \lambda')$.

Exercise 22. Consider a set of N jobs that are assigned to m workers for processing. Each job is randomly assigned to worker i with probability p_i, for $i = 1, \ldots, m$. Let N_i denote the number of items assigned to worker i, so that $N = N_1 + \cdots + N_m$. Suppose N has a Poisson distribution with mean λ. Describe the joint distribution of $N_1, \ldots, N_m$.

Exercise 23. Patients at an emergency room are categorized into m types. Assume the arrivals of the m types of patients occur at times that form independent homogeneous Poisson processes with respective rates $\lambda_1, \ldots, \lambda_m$.
(a) Find the probability that a type 1 patient arrives before a type 2 patient.
(b) What is the probability that the next patient to arrive after a specified time is of type 1?
(c) Find the probability that in the next 5 arrivals, there are exactly 3 type 1 patients.
(d) Find the probability that 3 type 1 patients arrive before the first type 2 patient.
(e) Find the probability that the next patient to arrive is of type 1, 2 or 3.

Exercise 24. *Dynamic Servicing.* Customers randomly request service at a manufacturing facility during an eight-hour day according to a Poisson process with intensity λ_t. The requested orders are satisfied as soon as possible, but may be delayed due to machine workloads, worker schedules, machine availability, etc. Past history shows that a request at time t will be satisfied either: (1) That day. (2) The next day. (3) Some time later. The request at time t is satisfied under scenario i with probability $p_i(t)$, and the expected revenue for such an order is r_i, where $i = 1, 2, 3$.
(a) Find the distribution of the number of requests in a day that are satisfied under each scenario i, where $i = 1, 2, 3$.
(b) Find the daily expected revenue for satisfying the customers, and find the variance of this revenue.
(This is an actual model of customer requests for orders of paper labels produced by a company.)

Exercise 25. Requests for a product (information or service) arrive from m cities at times that form independent Poisson processes with rates $\lambda_1, \ldots, \lambda_m$. Given that there are n requests from the cities in the time interval $(0, t]$, find the conditional probability that n_1 are from city 1 and n_2 are from city 2.

Exercise 26. At the end of a production shift, it is anticipated that there will be N jobs left to be processed, where N has a Poisson distribution with mean μ. Suppose the jobs are processed in parallel and the times to complete them are independent with a distribution G. Let $Q(t)$ denote the number of jobs in the system at time t, and let $D(t)$ denote the number of jobs completed in $(0, t]$. Find the distributions of $Q(t)$ and $D(t)$. Is D a Poisson process?

Answer this question when the jobs are processed serially (one at a time) and G is an exponential distribution with rate λ.

Exercise 27. Let $X_{(1)} \le \cdots \le X_{(n)}$ denote the order statistics from a random sample of size n from an exponential distribution with rate λ. Consider the distances between points $D_1 = X_{(1)}$, and $D_k = X_{(k)} - X_{(k-1)}$, $2 \le k \le n$. Show that these distances are independent, and that D_k has an exponential distribution with rate $(n - k + 1)\lambda$.

Exercise 28. Let $X_{(1)} \le \cdots \le X_{(n)}$ denote the order statistics of a random sample from a continuous distribution $F(x)$ with density $f(x)$. Show that the distribution and density of $X_{(k)}$ are

$$P\{X_{(k)} \le x\} = \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j},$$

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} f(x)(1 - F(x))^{n-k}, \quad x \in \mathbb{R}.$$

Exercise 29. *Point Locations for Non-Homogeneous Poisson Processes.* Consider a Poisson process N on $\mathbb{R}_+$ with rate function $r(x)$, where $r(x) > 0$ for each x. Prove the following *Order Statistic Property*:
The conditional density of $T_1, \ldots, T_n$ given $N(t) = n$ is

$$f_{T_1,\ldots,T_n}\Big(t_1, t_2, \ldots, t_n \Big| N(t) = n\Big) = n! f(t_1) \cdots f(t_n),$$

for $0 < t_1 < \cdots < t_n < t$, where $f(s) = r(s)/\mu(0, t]$. This is the joint density of the order statistics of n i.i.d. random variables on $[0, t]$ with density $f(s)$.

Exercise 30. Consider a Poisson process N on $\mathbb{R}_+$ with point locations $T_1 < T_2 < \ldots$ and rate function $r(t) = 3t^2$.
(a) Show that $W_n = T_n - T_{n-1}$, $n \ge 1$, are dependent.
(b) Find the distributions of T_1 and T_n.
(c) Find the distribution of W_n.

Exercise 31. Suppose N is a Poisson process on $\mathbb{R}^2$ with rate function $\lambda(x, y)$ at the location (x, y). For instance, N could represent locations of certain types of animal nests, diseased trees, land mines, auto accidents, houses of certain types of people, flaws on a surface, potholes, etc. Let D_n denote the

distance from the origin to the n-th nearest point of N.
(a) Find an expression for the distribution and mean of D_1.
(b) Find an expression for the distribution of D_n when $\lambda(x, y) = \lambda$.
(c) Are the differences $D_n - D_{n-1}$ independent (as they are for Poisson interpoint distances on $\mathbb{R}$)?
(d) Suppose there is a point located at (x^*, y^*). What is the distribution of the distance to the nearest point?
(e) Is $\lambda(x, y) = 1/(x^2 + y^2)^{1/2}$ a valid rate function for N to be a Poisson process under our definition?
(f) Specify a rate function $\lambda(x, y)$ under which $P\{N(\mathbb{R}^2) < \infty\} = 1$.

Exercise 32. Let M denote a Poisson process on $\mathbb{R}^d_+$ with intensity μ. Show that $N(t) = M((0, t]^d)$, for $t \geq 0$, is a Poisson process on $\mathbb{R}_+$ with $E[N(t)] = \mu((0, t]^d)$. This fact for $d = 2$ provides an alternate approach to proving the departure process in an $M_t/G_t/\infty$ system is Poisson; see Section 3.12.

Exercise 33. *Highway Model.* Vehicles enter an infinite highway denoted by $\mathbb{R}$ at times that form a Poisson process N on the time axis $\mathbb{R}_+$ with intensity measure μ. For simplicity, assume the highway is empty at time 0. The vehicle arriving at time T_n enters at a location X_n on the highway $\mathbb{R}$ and moves on it with a velocity V_n for a time τ_n and then exits the highway. The velocity may be negative, denoting a movement in the negative direction and vehicles may automatically pass one another on the highway with no change in velocity. The X_n are i.i.d. with distribution F and are independent of N. The pairs (V_n, τ_n) are independent of N and, they are conditionally independent given the X_n with

$$G_x(v, t) = P\{V_n \leq v, \tau_n \leq t | X_k, k \geq 1, X_n = x\},$$

a non-random distribution independent of n.
(a) Justify that $M = \sum_n \delta_{(T_n, X_n, V_n, \tau_n)}$, is a Poisson process on $\mathbb{R}_+ \times \mathbb{R}^2 \times \mathbb{R}_+$ and describe its intensity.
(b) Consider the departure process D on $\mathbb{R} \times \mathbb{R}_+$ where $D(A \times (a, b])$ is the number of departures from A in the time interval $(a, b]$. Justify that D is a Poisson process and specify its intensity. Find the expected number of departures in $(0, t]$.
(c) For a fixed t, let $N_t(A)$ denote the number of vehicles in $A \subseteq \mathbb{R}$ at time t. Justify that N_t is a Poisson process on $\mathbb{R}$ and specify its intensity.
(d) Suppose a vehicle is at the location $x \in \mathbb{R}$ at time t and let $X(t)$ denote the distance to the nearest vehicle. Specify assumptions on μ, F and $G_x(v, t)$ that would guarantee that there is at most one point at any location on the highway. Under these assumptions, find the distribution of $X(t)$.

Exercise 34. *Continuation.* In the preceding highway model, assume there are vehicles on the highway at time 0 at locations that form a Poisson process $N_0(\cdot)$ with rate λ, and this process is independent of the other vehicles.

The sojourn times of these vehicles on the highway are like those of the other vehicles vehicles, and they all operate under the distribution $G_x(v,t) = G(v)(1-e^{-\mu t})$. Solve parts (b)–(d) of the preceding exercise.

Exercise 35. In the context of Example 34, suppose N is a homogeneous Poisson process on the unit disc S in $\mathbb{R}^2$ with rate λ. For the Poisson processes N' and M related to the projection of N on the line $S' = [-1,1,]$, show that the rate function of N' is $2\lambda\sqrt{1-x^2}$, and that

$$E[M(A_u \times (0,b])] = \lambda \int_0^b (\sqrt{1-x^2} - u)\, dx,$$

for $A_u = \{(x,y) \in S : y \geq u\}$ and $b \leq \sqrt{1-u^2}$.

Next, consider the transformation of N where a point in the unit disc S is mapped to the closest point on the unit circle C. Under this map, using polar coordinates, $\overline{M}(A \times B) = \sum_n \delta_{(R_n,\Theta_n)}(A)\delta_{\Theta_n}(B)$ denotes the number of points of N in A that are mapped into B. Justify that $\overline{M}$ is a Poisson process on $S \times C$, and give an expression for $E[\overline{M}(A \times (0,b])]$, where $A = \{(r,\theta) \in S : \theta \in (0,b], r \in [\frac{1}{\sin\theta+\cos\theta}, 1]\}$ and $b \leq \pi/2$ (note that $x+y = r(\sin\theta+\cos\theta) \geq 1$ when $(x,y) \in A$).

For a fixed B, consider the process $N(r) = \overline{M}(A_r \times B)$, $r \in [0,1]$, where A_r is a disc in $\mathbb{R}^2$ with radius r. Show that N is a Poisson process and specify its rate function $\lambda(r)$.

Exercise 36. Suppose $N_1, \ldots, N_m$ are independent Poisson processes on a space S with respective intensities $\mu_1, \ldots, \mu_m$, and let $N = \sum_{i=1}^m N_i$, which is a Poisson process with intensity $\mu = \mu_1 + \cdots \mu_m$. For instance, $N_i(B)$ might be the number of crimes of type i in a region B of a city and $N(B)$ is the total number of crimes. Show that the conditional distribution of $N_1(B), \ldots, N_m(B)$ given $N(B) = n$ is a multinomial distribution.

Exercise 37. Let $N_1, \ldots, N_m$ be independent Poisson processes on $\mathbb{R}_+$ with location-dependent rates $\lambda_1(t), \ldots, \lambda_m(t)$, respectively. Let τ_i denote the time of the first occurrence in process N_i. Find the distribution of $\tau = \min_{1 \leq i \leq m} \tau_i$. Find $P\{\tau_i = \tau\}$.

Exercise 38. Messages arrive to a web page at times that form a Poisson process with rate λ. The messages are independently of high priority with probability p, and each of these high priority messages is independently of type i with probability p_i, $\leq i \leq m$. Let $N_i(t)$ denote the number of high priority messages of type i that arrive in $[0,t]$. Are $N_1, \ldots, N_m$ independent Poisson processes? If so, specify their rates.

Exercise 39. Let $X_1, X_2, \ldots$ be i.i.d. random variables with a distribution F. Show that $X_1, \ldots, X_n$ are distinct with probability one for any $n \geq 2$ if and only if F is continuous. (This statement is also true if these are random elements in a space S and "continuous F" is replaced by $F(\{x\}) = 0$, $x \in S$). Hint: Use induction and $P\{X_1 \neq X_2\} = \int_{\mathbb{R}} (1 - F(\{x\})F(dx)$.

Exercise 40. *Poisson Processes are Infinitely Divisible.* Let N be a Poisson process on S with intensity μ. Using Laplace functionals, show that for each n, there exist point processes $N_1, \ldots, N_n$ that are i.i.d. and $N \stackrel{d}{=} N_1 + \cdots + N_n$. This says that N is *infinitely divisible*.

Exercise 41. As in Example 34, suppose $N = \sum_n \delta_{(X_n, Y_n)}$ is a Poisson process on the unit disc S in $\mathbb{R}^2$ with location-dependent rate $\lambda(x, y)$. The Poisson process $M = \sum_n \delta_{((X_n, Y_n), X_n)}$ on $S \times S'$ represents the number of points of N that are projected onto the x-axis $S' = [-1, 1]$. Give an expression for the location-dependent intensity of M. Switching to polar coordinates as in Example 34, describe the process $\overline{M} = \sum_n \delta_{((R_n, \Theta_n), \Theta_n)}$ that represents the number of points of N that are mapped onto the unit circle C.

Exercise 42. Deposits to a bank account occur at times that form a Poisson process with rate λ and the amounts deposited are independent random variables with distribution F (independent of the times). Also, withdrawals occur at times that form a Poisson process with rate μ and the amounts deposited are independent random variables with distribution G (independent of the times). The deposits and withdrawals are independent. Let $X(t)$ denote the balance of the bank account at time t; the balance may be negative. Find the mean, variance and distribution of $X(t)$ when $X(0) = 0$.

Exercise 43. Suppose $X(t) = \min_{n \le N(t)} Y_n$, for $t \ge 0$, where $N = \sum_n \delta_{T_n}$ is a Poisson process on $\mathbb{R}_+$ and Y_n are i.i.d. with distribution F, independent of N. For instance, Y_n could be bids on a property and $X(t)$ is the smallest bid up to time t. Find the distribution and mean of $X(t)$. Answer this question for the more general setting in which Y_n are p-marks of T_n, where $p(t, (0, y])$ is the distribution of a typical mark at time t.

Exercise 44. *E-mail Broadcasting.* An official of an organization sends e-mail messages to various subgroups of the organization at times that form a Poisson process with intensity μ. Each message is sent to individual i with probability p_i, $i = 1, \ldots, m$, where m is the number of individuals in the organization; and the message is sent at the same time to those selected individuals. Let $N_i(t)$ denote the number of messages individual i receives from the official in the time interval $(0, t]$. Justify that N_i is a Poisson process and specify its intensity.

Consider a subset of individuals $I \subseteq \{1, \ldots, m\}$. Specify whether or not the processes N_i, $i \in I$, are independent. Is the sum $\sum_{i \in I} N_i$ Poisson? If so, specify its intensity.

Exercise 45. *Continuation.* In the setting of the preceding exercise, suppose there is a probability r_i that individual i will reply to an e-mail from the official, and then send the reply within a time that has a distribution $G_i(t)$. Let $R_i(t)$ denote the number of replies the official receives in $(0, t]$ from all

the messages sent to individual i. Is R_i Poisson? If so, specify its intensity. Is the sum $\sum_{i\in I} R_i$ Poisson? If so, specify its intensity.

Suppose individual i receives a message from the official, and is planning to reply to it. Find the probability that before the reply is sent, another message from the official will arrive?

Exercise 46. A satellite circles a body in outer space and records a special feature of the body (e.g. rocks, water, low elevations) along a path it monitors. As an idealized model, assume the feature occurs on the polar-angle space $S = [0, 2\pi]$ at angles $\Theta_1 \le \Theta_2 \ldots \le 2\pi$ that form a Poisson process with intensity μ. We will only consider one orbit of the satellite. Suppose the satellite is moving at a (deterministic) velocity of γ radians per unit time. Upon observing an occurrence at Θ_n the satellite sends a message to a station that receives it after a time τ_n. Suppose the transmission times τ_n are independent with distribution G and are independent of the positions of the occurrences. Consider the point process M on $S \times \mathbb{R}^2_+$, where $M((\alpha, \beta] \times (a, b] \times (c, d])$ is the number of occurrences in the radian set $(\alpha, \beta]$ that are observed in the time set $(a, b]$ and received at the station in the time set $(c, d]$. Describe the process M and its intensity measure in terms of the system data.

Next, let $N(t)$ denote the number of messages received at the station in $(0, t]$ whose transmission time exceeds a certain limit L, where $G(L) < 1$. Describe the process N and specify its intensity measure.

Exercise 47. *Multiclass $M/G/\infty$ System.* Consider an $M/G/\infty$ system in which items arrive at times that form a Poisson process with rate λ. There are m classes or types of items and p_i is the probability that an item is of class i. The processing time of a class i item has a distribution $G_i(\cdot)$. Assume the system is empty at time 0. Let $Q_i(t)$ denote the quantity of class i items in the system at time t. Specify its distribution. Determine whether or not $Q_1(t), \ldots, Q_m(t)$ are independent. Let $D_i(t)$ denote the number of departures of class i items in $(0, t]$. Describe these processes including their independence.

Exercise 48. *Limiting Behavior of $M/G/\infty$ System.* Consider the $M/G/\infty$ system in Section 3.12 with arrival rate λ and service distribution G, which has a mean α. Show that the limiting distribution of the quantity of items in the system $Q(t)$ is Poisson with mean $\lambda\alpha$ as $t \to \infty$. Use the fact that $\alpha = \int_0^\infty [1 - G(u)]du$. Turning to the departures, consider the point process $D_t(B)$ on $\mathbb{R}_+$ that records the numbers of departures in a time set B after time t. In particular, show that the number of departures $D_t(0, b]$ in the interval $(t, t + b]$ has a Poisson distribution with

$$E[D_t(0, b]] = \lambda\Big[\int_t^{t+b} G(u + b)du - \int_t^{t+b} G(u)du\Big].$$

Show that the limiting distribution of $D_t(0, b]$ is Poisson with mean λb. More generally, show that the finite-dimensional distributions of D_t converge to

those of a homogeneous Poisson process D with rate λ. This proves, in light of Theorem 62, that $D_t \xrightarrow{d} D$.

Exercise 49. *Spatial $M/G/\infty$ System.* Consider a system in which items enter a space S at times $T_1 \leq T_2 \leq \ldots$ that form a Poisson process with intensity measure μ. The nth item that arrives at time T_n enters S at the location X_n and remains there for a time V_n and then exits the system. Suppose $F_t(\cdot)$ is the distribution of the location in S of an item arriving at time t, and $G_{(t,x)}(\cdot)$ is the distribution of the item's sojourn time at a location x. More precisely, assume (X_n, V_n) are location-dependent marks of T_n with distribution

$$p(t, A \times (0, v]) = \int_A G_{(t,x)}(v) F_t(dx).$$

Let $N_t(B)$ denote the number of items in the set $B \in \mathcal{S}$ at a fixed time t. Show that N_t is a Poisson process on S with

$$E[N_t(B)] = \int_{(0,t]} \int_B [1 - G_{(s,x)}(t-s)] F_s(dx) \mu(ds).$$

Next, let $D((a, b] \times B)$ denote the number of departures from the set B in the time interval $(a, b]$. Show that D is a space-time Poisson process on $\mathbb{R}_+ \times S$ and specify $E[D((0, t] \times B)]$.

Exercise 50. For the network in Example 52, justify that the following processes are Poisson and specify their intensity measures.

$D(t) = \#$ of items that depart from the network in $(0, t]$.
$D_1(t) = \#$ of items that enter node 1 and
depart from the network in $(0, t]$.
$D_{(2,3,5)}(t) = \#$ of items that complete the route $(2, 3, 5)$ in $(0, t]$.

Justify that the following random variables, for a fixed t, have a Poisson distribution and specify their means.

$Q(t) = \#$ of items that are in the network at time t.
$Q_1(t) = \#$ of items that are beyond their first node at time t.
$Q_{3|2}(t) = \#$ of items in node 3 at time t that came from node 2.

Exercise 51. *Time Transformations and Cox Processes.* Suppose that $N_1(t) = \sum_n \mathbf{1}(T_n \leq t)$ is a homogeneous Poisson process on $\mathbb{R}_+$ with rate λ, and let η denote a locally finite measure on $\mathbb{R}_+$ with $\eta(\mathbb{R}_+) = \infty$. Consider the process $N(t) = N_1(\eta(t))$, $t \geq 0$. Show that N is a Poisson process on $\mathbb{R}_+$ with intensity η. Do this by showing that N is a transformation of N_1 under a map $g : \mathbb{R}_+ \to \mathbb{R}_+$; that is, find a g such that

$$N(t) = \sum_n \mathbf{1}(T_n \le \eta(t)) = \sum_n \mathbf{1}(g(T_n) \le t).$$

This fact implies that if η is a locally finite random measure on $\mathbb{R}_+$ with $\eta(\mathbb{R}_+) = \infty$ a.s., then N is a Cox process.

Exercise 52. Suppose that N is a *Cox process* on S directed by a locally-finite random measure η. Show that $E[N(B)] = \mathrm{Var}[N(B)]$, for $B \in \mathcal{S}$.

Exercise 53. *Poisson Process Directed by a Cyclic Renewal Process.* The state of a system is represented by a continuous-time cyclic renewal process $X(t)$ on states $0, 1, \ldots, K-1$ as in Example 8. The sojourn times in the states are independent, and the sojourn time in state i has a continuous distribution F_i with mean μ_i. By Exercise 47, $\lim_{t\to\infty} P\{X(t) = i\} = \mu_i / \sum_{k=0}^{K-1} \mu_k$.

Suppose the system fails occasionally such that, while it is in state i, failures occur according to a Poisson process with rate λ_i, independent of everything else. Let $N(t)$ denote the number of failures in $(0, t]$. Show that

$$t^{-1}N(t) \to \frac{\sum_{k=0}^{K-1} \lambda_k \mu_k}{\sum_{k=0}^{K-1} \mu_k}, \quad a.s. \text{ as } t \to \infty.$$

Assume the system begins in state 0 and let τ denote the first time it returns to state 0 (the time to complete a cycle). Show that

$$E[N(\tau)] = \sum_{k=0}^{K-1} \lambda_k \mu_k = \mathrm{Var}[N(\tau)].$$

Exercise 54. *Location-Dependent Compound Poisson Process.* Suppose that $Z(t) = \sum_{n=0}^{N(t)} Y_n$ is a location-dependent compound Poisson process, where $N = \sum_n \delta_{T_n}$ is a Poisson process on $\mathbb{R}_+$ with intensity measure μ, and Y_n are p-marks of T_n. Show that the process $Z(t)$ has independent increments (the increments will not be stationary in general), and

$$E[Z(t)] = \int_{(0,t]} \int_{\mathbb{R}} y p(s, dy) \mu(ds).$$

Suppose the moment generating function $\phi_{s,t}(\alpha) = \int_{\mathbb{R}} e^{\alpha y} F_{s,t}(dy)$ exists, where

$$F_{s,t}(y) = \int_{(s,t]} p(u, (0, y]) \mu(du) / \mu(s, t].$$

Show that, for $s < t$,

$$E[e^{\alpha[Z(t)-Z(s)]}] = e^{-\mu(s,t][1-\phi_{s,t}(\alpha)]}. \tag{3.56}$$

(This is the moment generating function of a compound Poisson distribution with rate $\mu(s, t]$ and distribution $F_{s,t}$.) Use the fact

$$E[e^{\alpha[Z(t)-Z(s)]}] = E[e^{\int_{\mathbb{R}} yM((s,t]\times dy)}] = E[e^{-Mh_{s,t}}],$$

where $h_{s,t}(u,y) = -\alpha y \mathbf{1}(u \in (s,t])$ and $M = \sum_n \delta_{(T_n,Y_n)}$.

Exercise 55. Suppose $Z_1(t), \ldots, Z_m(t)$, are independent compound Poisson processes with respective rates $\lambda_1, \ldots, \lambda_m$ and distributions $F_1, \ldots, F_m$. Show that $Z(t) = \sum_{i=1}^m Z_i(t)$ is a compound Poisson process with rate $\lambda = \sum_{i=1}^n \lambda_i$ and distribution $F = \lambda^{-1} \sum_{i=1}^n \lambda_i F_i$.

Exercise 56. *Partition of a Compound Poisson Process.* Suppose $Z(t) = \sum_{n=1}^{N(t)} Y_n$, for $t \geq 0$, is a location-dependent compound Poisson process with intensity measure μ and distribution $p(t,\cdot)$. Suppose the quantity Y_n at time T_n is partitioned into m pieces $\mathbf{Y}'_n = (Y'_{n1}, \ldots, Y'_{nm})$ so that $Y_n = \sum_{i=1}^m Y'_{ni}$. These pieces are assigned to m processes defined by $Z_i(t) = \sum_{n=1}^{N(t)} Y'_{ni}$. They form a partition of $Z(t)$ in that $Z(t) = \sum_{i=1}^m Z_i(t)$. Assume the $\mathbf{Y}'_n$ are p'-marks of (T_n, Y_n), where $p'((t,y), B_1 \times \cdots \times B_m)$ is the conditional distribution of a typical vector $\mathbf{Y}'_n$ given $(T_n, Y_n) = (t,y)$. Prove that $Z_i(t)$ is a compound Poisson process with intensity μ and distribution $p'(t,\cdot)$, and specify $p'(t,\cdot)$. Use the idea that the $\mathbf{Y}'_n$ are second marks of the Poisson process N as discussed in Example 39, resulting in $M' = \sum_n \delta_{(T_n,Y_n,\mathbf{Y}'_n)}$.

Exercise 57. *Origin-Dependent Cluster Processes.* The cluster process in Example 58 has the form $M(A \times B) = \sum_n N'_n(B)\delta_{X_n}(A)$, where N'_n are point processes on a space S' generated by the points X_n in S. Instead of assuming the N'_n are independent of N, consider the more general setting in which the N'_n are p-marks of X_n. Let N'_x be a point process on S' such that $p(x,C) = P\{N'_x \in C\}$. Show (by conditioning on N) that the Laplace functional of M is

$$E[e^{-Mf}] = \exp[-\int_S (1 - g(x))\mu(dx)],$$

where $g(x) = E[e^{-\int_{S'} f(x,x')N'_x(dx')}]$.

Exercise 58. The moments of a point process N on S are given by

$$E[N(A_1)^{n_1} \cdots N(A_k)^{n_k}] = (-1)^{n_1+\cdots+n_k} \frac{\partial^{n_1+\cdots+n_k}}{\partial t_1^{n_1} \cdots \partial t_k^{n_k}} E[e^{-Nf}]|_{t_1=\ldots=t_k=0},$$

where $f(x) = \sum_{i=1}^k t_i 1(x \in A_i)$. Prove this for $k = 1$ and $k = 2$. Use this fact to find expressions for the first two moments of the cluster process quantity $M(A \times B)$ in Exercise 57.

Exercise 59. *Sums of Identically Distributed Renewal Processes.* Suppose that $\tilde{N}_1, \tilde{N}_2 \ldots$ are independent renewal processes with inter-renewal distribution F. By the strong law of large numbers, we know that the sum

$\sum_{i=1}^{n} \tilde{N}_i(t)$ converges to ∞ a.s. as $n \to \infty$. (The discussion prior to Example 65 addressed the opposite case where the sum tends to 0.) To normalize this sum (analogously to that in a central limit theorem) so that it converges to a non-degenerate limit, it is natural to rescale the time axis and consider the process $N_n(t) = \sum_{i=1}^{n} \tilde{N}_i(t/n)$. This is the sum with $1/n$ as the new unit of time. Assume the derivative $\lambda = F'(0)$ exists and is positive. Show that $N_n \xrightarrow{d} N$, where N is a Poisson process with rate λ.

Exercise 60. *Poisson Limit of Thinned Processes.* Let N_n be a sequence of point processes on S. Suppose N_n is subject to a $p_n(x)$ thinning: A point of N_n at x is retained with probability $p_n(x)$ and is deleted with probability $1 - p_n(x)$. Let N_n' denote the resulting thinned process on S. Assume the thinning is uniformly null in that

$$\lim_{n\to\infty} \sup_x p_n(x) = 0, \quad B \in \hat{\mathcal{S}}.$$

Show that $N_n' \xrightarrow{d} N'$, a Poisson process on S with intensity measure μ, if

$$\int_B p_n(x) N_n(dx) \xrightarrow{d} \mu(B), \qquad B \in \hat{S}_\mu, \text{ as } n \to \infty.$$

Chapter 4
Continuous-Time Markov Chains

A continuous-time Markov chain (CTMC) is a discrete-time Markov chain with the modification that, instead of spending one time unit in a state, it remains in a state for an exponentially distributed time whose rate depends on the state. The methodology of CTMCs is based on properties of renewal and Poisson processes as well as discrete-time chains. CTMCs are natural candidates for modeling systems in real time such as production and inventory systems, computer and telecommunications networks, and miscellaneous input-output systems. Many continuous-time processes have discrete-time analogues; for instance, birth-death and Brownian motion processes are continuous-time analogues of discrete-time random walks. One's choice of a continuous- or discrete-time model for a system typically depends on how realistic it is, its ease in addressing the issues at hand, or in computing quantities of interest.

A CTMC is a continuous-time Markov process on a countable state space whose sample paths are right-continuous and piecewise-constant with finite lengths, and the number of jumps in any finite time is finite. This type of Markov process is represented by the sequence of states it visits and the sojourn times at the visits.

Our study of these processes begins with Kolmogorov's backward and forward differential equations for the transition probabilities of a CTMC, and a characterization of its infinitesimal transition rates. Next is a description of properties of sample paths of CTMCs. Included is a "uniformization" property that a CTMC with bounded transition rates can be represented as a Markov chain that takes jumps at times that form a Poisson process.

Several sections are devoted to describing the equilibrium behavior of CTMCs, including ergodic theorems for functions of CTMCs and Lévy-type expressions for expectations. Then we give detailed descriptions of reversible CTMCs, Jackson network processes, and multiclass networks.

Next, we show how Palm Probabilities are used to describe a CTMC conditioned on the occurrence of a certain type of transition. For instance, in a production system with queueing, one may be interested in the number of

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

items in the system conditioned that an arrival occurs. In this and analogous situations, we present PASTA properties that "Poisson arrivals see time averages", or "Palm actions see time averages".

The chapter ends with a description of $M/G/1$ and $G/M/1$ queueing processes, and an introduction to Markov renewal processes, which are relatives of a CTMC.

4.1 Introduction

In this section, we introduce CTMCs and describe some of their features.

A continuous-time stochastic process $\{X(t) : t \geq 0\}$ on a countable state space S is a *Markov process* if it satisfies the Markov property: for each $i, j \in S$ and $t, u \geq 0$,

$$P\{X(t+u) = j | X(t) = i, X(s), s < t\} = P\{X(u) = j | X(0) = i\}.$$

This is a *time-homogeneous* process because the last probability does not depend on t. Our interest is in such processes with nicely behaved sample paths as follows.

Definition 1. A Markov process $\{X(t) : t \geq 0\}$ on a countable state space S is a *continuous-time Markov chain* (CTMC) if its sample paths are right-continuous and piecewise constant with finite lengths a.s., and the number of transitions in any finite time interval is finite a.s. That is,

$$X(t) = X_n \qquad \text{if } t \in [T_n, T_{n+1}) \text{ for some } n, \tag{4.1}$$

where $0 = T_0 < T_1 < T_2 < \cdots$ are the jump times of the process, and X_n is the state visited at time T_n. The $T_n \to \infty$ a.s. and $X_n \neq X_{n+1}$, for each n. The *sojourn time* of the process in state X_n is $\chi_n = T_{n+1} - T_n$.[1]

Any process $X(t)$ of the form (4.1) is called a *jump process* on S with *embedded process* (X_n, ξ_n). Our first observation is a characterization of a CTMC in terms of elementary properties of its embedded process.

Theorem 2. *A jump process $X(t)$ on S with embedded process (X_n, ξ_n) is a CTMC if and only if*
(i) X_n is a discrete-time Markov chain on S with transition probabilities $\mathrm{P} = \{p_{ij}\}$, where $p_{ii} = 0$, for each i.
(ii) For nonnegative $t_0, \ldots, t_m$,

[1] This differs from the renewal process notation, where $\xi_n = T_n - T_{n-1}$ is the nth inter-renewal time.

$$P\{\xi_0 \le t_0, \ldots, \xi_m \le t_m | X_n, n \ge 0\} = \prod_{n=0}^{m} P\{\xi_n \le t_n | X_n\}, \tag{4.2}$$

and there are positive q_i, $i \in S$, *such that, for each* $n \ge 0$,

$$P\{\xi_n \le t | X_n = i\} = 1 - e^{-q_i t}, \qquad t \ge 0,\ i \in S. \tag{4.3}$$

Proof. This follows by Propositions 12 and 13 in Section 4.3.

In this theorem, the embedded Markov chain X_n may be transient, irreducible, ergodic, etc. Subsequent sections show how $X(t)$ inherits these properties of X_n. Note that the process $X(t)$ does not have absorbing states, since its sojourn times in the states it visits are finite. For particular applications, one can describe Markov processes with the jump-like behavior described above, but with absorbing states, by the techniques in Chapter 1 and this chapter. For instance, one can determine probabilities of absorbtion or times to absorbtion by the results in Section 1.7 in Chapter 1 and in Section 4.6 below. Our study will not cover such Markov processes. Similar statements apply to Markov processes that may have an infinite number of jumps in a finite time, resulting in finite lifetimes.

Condition (4.2) says that the sojourn times ξ_n are conditionally independent given the X_n, and condition (4.3) says that such a sojourn time in a state i is exponentially distributed with rate q_i.

Another way of stating Theorem 2 is as follows. This is a consequence of the Markov property of discrete-time Markov chains on general state spaces.

Remark 3. A jump process $X(t)$ with embedded process (X_n, ξ_n) is a CTMC if and only if (X_n, ξ_n) is a discrete-time Markov chain on[2] $S \times \mathbb{R}_+$ with transition probabilities

$$P\{X_{n+1} = j, \xi_{n+1} \le t | X_n = i, \xi_n\} = p_{ij}(1 - e^{-q_i t}), \qquad i, j \in S,\ t \ge 0,$$

where q_i are positive constants and $\mathrm{P} = \{p_{ij}\}$ is a Markov transition matrix with each $p_{ii} = 0$.

Hereafter, we will adopt the notation of Theorem 2 for describing CTMCs. In particular, we say that the *defining parameters* of the CTMC $X(t)$ are (α_i, p_{ij}, q_i), where $\alpha_i = P\{X(0) = i\}$ is the initial distribution. We refer to p_{ij} and q_i as the *main defining parameters*. Theorem 22 below shows that there exists a CTMC $X(t)$ for any set of defining parameters (α_i, p_{ij}, q_i). These parameters uniquely determine the distribution of a CTMC, and vice versa, as follows.

Proposition 4. *Two CTMCs have the same distribution if and only if their defining parameters are equal.*

[2] This state space is not countable.

Proof. It suffices to show that the following statements are equivalent.
(a) The two CTMCs have the same distribution.
(b) The embedded Markov chains for the CTMCs have the same distribution.
(c) The defining parameters for the CTMCs are equal.

Clearly (a) and (b) are equivalent by the construction of a CTMC. Also, (b) and (c) are equivalent by the property from Chapter 1 that two Markov chains have the same distribution if and only if their initial distributions and transition probabilities are equal.

By its definition, a CTMC satisfies the *regularity* condition that $T_n \to \infty$ a.s., which ensures that the CTMC is defined on the "entire" time axis. This condition has the following characterization in terms of the defining parameters of a CTMC.

Proposition 5. *Suppose that conditions (i) and (ii) in Theorem 2 hold. Then* $T_n \to \infty$ *a.s. if and only if the* q_i *are* P-regular *in the sense that*

$$\sum_{n=0}^{\infty} q_{X_n}^{-1} = \infty \qquad \text{a.s.} \tag{4.4}$$

In particular, $T_n \to \infty$ *a.s. if* $\sup_{i \in S} q_i < \infty$ *(which is true for finite* S*), or if the Markov chain* X_n *is recurrent.*

Proof. Clearly $T_n \to \infty$ a.s. means $Z = \sum_{n=0}^{\infty} \xi_n = \infty$ a.s. Also, by condition (ii) and $E[\xi_n | X_n] = q_{X_n}^{-1}$, we have

$$E[Z | X_n, n \geq 0] = \sum_{n=0}^{\infty} E[\xi_n | X_n] = \sum_{n=0}^{\infty} q_{X_n}^{-1}.$$

In light of these observations, the assertion to be proved is that

$$E[Z | X_n, n \geq 0] = \infty \text{ a.s.} \quad \Longleftrightarrow \quad P\{Z = \infty\} = 1. \tag{4.5}$$

Now, Exercise 7 in Chapter 3 shows that an infinite sum of independent exponential random variables is infinite a.s. if and only if the sum of their means is infinite. Applying this to the exponential sojourn times conditioned on the X_n, we have

$$P\{Z = \infty | X_n, n \geq 0\} = 1 \text{ a.s.} \quad \Longleftrightarrow \quad E[Z | X_n, n \geq 0] = \infty \text{ a.s.}$$

On the other hand, using $P\{Z = \infty\} = E[P\{Z = \infty | X_n, n \geq 0\}]$, we have

$$P\{Z = \infty | X_n, n \geq 0\} = 1 \text{ a.s.} \quad \Longleftrightarrow \quad P\{Z = \infty\} = 1.$$

These two equivalences prove (4.5).

Next, if $b = \sup_{i \in S} q_i < \infty$, then $q_{X_n}^{-1} \geq b^{-1}$, which clearly implies (4.4). Finally, if X_n is recurrent, it visits some state i infinitely often, so q_i^{-1} appears an infinite number of times in the sum in (4.4), and hence that sum is infinite.

4.2 Examples

To justify that a jump process is a CTMC, one typically verifies the conditions in Theorem 2 that the sequence of states it visits is a discrete-time Markov chain, and that its sojourn times are exponentially distributed; one must also verify the P-regularity property. Here are some standard examples.

Example 6. Poisson Process. A Poisson process $N(t)$ with rate λ is a jump process whose state increases by unit jumps, and its sojourn times are independent and exponentially distributed with rate λ. Therefore, $N(t)$ is a CTMC with one-step transition probabilities $p_{i,i+1} = 1$, and exponential sojourn rates $q_i = \lambda$, which are obviously P-regular.

Example 7. Pure Birth Process. Suppose that $X(t)$ represents the number of occurrences of a certain event in the time interval $[0, t]$. Assume the inter-occurrence times are independent, and the time between the ith and $i+1$st occurrence is exponentially distributed with rate q_i. Therefore, the state increases by unit jumps, and the sojourn times are exponentially distributed. The q_i are P-regular if and only if $\sum_{i=0}^{\infty} q_i^{-1} = \infty$. So assuming this is true, $X(t)$ is a CTMC on $\mathbb{Z}_+$ with parameters $p_{i,i+1} = 1$ and q_i. For instance, the $X(t)$ might denote the number of times a system is repaired. In this case, the q_i would typically be increasing for a machine that wears out, but it would be decreasing for a software package that is perfected as flaws are fixed.

This process is called a *pure birth process* because of the classical model in which $X(t)$ is the size of a population, and whenever the population size is i, the time to the next birth is exponentially distributed with rate q_i.

Note that if $\sum_{i=0}^{\infty} q_i^{-1} < \infty$, the jump process $X(t)$ could still be defined, but only on the time interval $[0, \sup_n T_n)$, which is finite with a positive probability. There would be an infinite number of births in this time interval.

Since a CTMC is essentially a Markov chain whose unit sojourn times in the states are replaced by exponential times, any discrete-time Markov chain has an analogous CTMC version. For instance, here is a continuous-time analogue of Exercise 49 in Chapter 1.

Example 8. Exhaustive Parallel Processing. A set of m jobs are processed in parallel until they are all completed. At the completion time, another m jobs instantaneously enter the system and are processed similarly. This is repeated indefinitely. Assume the times to process jobs are independent exponentially distributed with rate λ. Let $X(t)$ denote the number of jobs being processed at time t. This is a jump process on $S = \{1, \ldots, m\}$ and its embedded process (X_n, ξ_n) satisfies, for $2 \le i \le m$,

$$\begin{aligned} &P\{X_{n+1} = i-1, \xi_{n+1} > t | (X_m, \xi_m), m \le n, X_n = i\} \\ &\quad = P\{\min\{Y_1, \ldots, Y_i\} > t\} = e^{-i\lambda t}, \end{aligned}$$

where $Y_1, \ldots, Y_i$ are independent exponentially distributed with rate λ. Here we use the property that the minimum of independent exponential random variables is again exponential with rate that is the sum of the rates of the variables. The other transition probability is

$$\begin{aligned} &P\{X_{n+1} = m, \xi_{n+1} > t | (X_m, \xi_m), m \le n, X_n = 1\} \\ &\quad = P\{Y_1 > t\} = e^{-\lambda t}. \end{aligned}$$

Then by Remark 3, $X(t)$ is a CTMC with parameters

$$p_{i,i-1} = 1, \quad 2 \le i \le m, \quad p_{1,m} = 1,$$

and $q_i = i\lambda$. The P-regularity is due to the finite state space.

The following example describes a general framework for formulating a CTMC by clock times.

Example 9. Clock Times and Transition Rates. Suppose $X(t)$ is a jump process on S with embedded process (X_n, ξ_n) whose dynamics are as follows. Whenever the process enters a state i, a set of independent *clock times* τ_{ij}, $j \in S_i$ are started, where S_i is the subset of states in $S \backslash \{i\}$ that can be reached from state i in one step. The times τ_{ij} are exponentially distributed with rates q_{ij}. Then the sojourn time in state i is the minimum $\tau_i = \min_{j \in S_i} \tau_{ij}$, and at the end of the sojourn, the process jumps to the state $j \in S_i$ for which $\tau_{ij} = \tau_i$.

Think of τ_{ij} as the time to the next "potential" transition from i to $j \in S_i$ with *transition rate* q_{ij}, and the clock time j that is the smallest of these times "triggers" a transition from i to j.

Under these assumptions,

$$P\{X_{n+1} = j, \xi_{n+1} > t | (X_m, \xi_m), m \le n, X_n = i\} = P\{\tau_{ij} = \tau_i, \tau_i > t\}.$$

By the properties of exponential random variables in Exercise 2 of Chapter 3, the sojourn time τ_i is exponentially distributed with rate $q_i = \sum_{j \in S_i} q_{ij}$, and it is independent of the event $\tau_{ij} = \tau_i$, where $P\{\tau_{ij} = \tau_i\} = q_{ij}/q_i$. Therefore, (X_n, ξ_n) is a Markov chain with transition probabilities

$$P\{X_1 = j, \xi_1 > t | X_0 = i, \xi_0\} = \frac{q_{ij}}{q_i} e^{-q_i t}.$$

Then $X(t)$ is a CTMC with parameters $p_{ij} = q_{ij}/q_i$ and $q_i = \sum_{j \in S_i} q_{ij}$, provided the q_i are P-regular.

More insights on such transition rates are in Section 4.3. See Exercise 4 for analogous discrete-time geometric clock times, and see Exercise 5 for clocks associated with multiple sources that trigger transitions.

Example 10. Birth-Death Process. Suppose that $X(t)$ represents the number of discrete items in a population at time t, where births and deaths (or

additions and departures) in the population occur as follows. Whenever the population state is 0, the time to the next birth is exponentially distributed with rate λ_0. Also, whenever there are $i \geq 1$ items in the population, the time to the next (potential) birth is exponentially distributed with rate λ_i, and the time to the next (potential) death is exponentially distributed with rate μ_i. These times are independent and independent of the rest of the process. Assume the birth and death rates are bounded.

Under these conditions, it follows as in the preceding example that $X(t)$ is a CTMC on $\mathbb{Z}_+$ with transition rates $q_{i,i+1} = \lambda_i$ and $q_{i,i-1} = \mu_i$. The $X(t)$ is called a *birth-death process* with birth rates λ_i and death rates μ_i. Its exponential sojourn rates are $q_i = \lambda_i + \mu_i$, where $\mu_0 = 0$, and its one-step transition probabilities are

$$p_{i,i+1} = \lambda_i/(\lambda_i + \mu_i), \qquad p_{i,i-1} = \mu_i/(\lambda_i + \mu_i).$$

There are a variety of queueing processes and general input-output processes that are birth-death processes. Here is a classic example.

Example 11. $M/M/s$ Queueing Process. Consider a processing system in which items arrive according to a Poisson process with rate λ. There are s servers who process the items one at a time, where $1 \leq s \leq \infty$. The processing times are independent exponential random variables with rate μ, independent of everything else. An item that arrives when all the s servers are busy waits in a queue for processing; otherwise it goes to any available server for processing. Whenever there are i items in the system, $\min\{i, s\}$ items are being processed independently at rate μ, and so the time for a potential departure (considered as a death) has an exponential distribution with rate $\mu_i = \mu \min\{i, s\}$. This is $\mu_i = i\mu$ when $s = \infty$.

Let $X(t)$ denote the number of items in the system at time t. Clearly, $X(t)$ is a birth-death CTMC with birth rate λ in each state, and death rate $\mu_i = \mu \min\{i, s\}$ in state $i \geq 1$. This process is called an *$M/M/s$ queueing process.*

Because $X(t)$ counts the quantity of items, which are indistinguishable, several service disciplines are possible (e.g., first-come-first-served, service in random or arbitrary order, or last-come-last-served).

4.3 Markov Properties

This section proves Theorem 2 that characterizes a CTMC by its embedded process. Included are an integral equation and the Chapman-Kolmogorov equations for its transition probabilities.

For this discussion, $X(t)$ will denote a continuous-time jump process on S with embedded process (X_n, ξ_n). As in Chapter 1, we adopt the convention that $P_i\{\cdot\} = P\{\cdot \,|X(0) = i\}$, and let $E_i[\cdot]$ denote the associated conditional

expectation. We denote the *transition probabilities* of $X(t)$ by

$$p_{ij}(t) = P_i\{X(t) = j\}, \qquad i, j \in S,\ t \geq 0.$$

These probabilities will play a similar role for CTMCs that the n-step probabilities play for a discrete-time Markov chain.

We begin with two results that prove Theorem 2. The integral equation (4.6) yields the Kolmogorov differential equations in the next section that determine transition rates and probabilities for CTMCs.

Proposition 12. *If the jump process $X(t)$ on S is such that its embedded process satisfies conditions (i) and (ii) in Theorem 2, then $X(t)$ is a CTMC and its transition probabilities $p_{ij}(t)$ satisfy, for $i, j \in S$ and $t > 0$,*

$$p_{ij}(t) = e^{-q_i t}\mathbf{1}(i = j) + \int_0^t \sum_{k \neq i} p_{kj}(t - v) q_i p_{ik} e^{-q_i v}\, dv. \tag{4.6}$$

Proof. To prove $X(t)$ is a CTMC, it suffices to verify the Markov property that, for each $i, j \in S$ and $t, u \geq 0$,

$$P\{X(t + u) = j | X(t) = i, X(s), s < t\} = p_{ij}(u). \tag{4.7}$$

Consider the point process $N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t)$, $t \geq 0$, which denotes the number of transitions of the $X(t)$ up to time t. For fixed states i and j, conditioning on $N(t)$, we have

$$\begin{aligned} &P\{X(t + u) = j | X(t) = i, X(s), s < t\} \\ &\quad = \sum_{m=0}^{\infty} [a_m(t, u) + b_m(t, u)] P\{N(t) = m | X(t) = i, X(s), s < t\}, \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} a_m(t, u) &= P\{X(t + u) = j, T_{m+1} > t + u | \mathcal{F}_m(i, t)\} \\ b_m(t, u) &= P\{X(t + u) = j, T_{m+1} \leq t + u | \mathcal{F}_m(i, t)\} \\ \mathcal{F}_m(i, t) &= \{N(t) = m, X(t) = i, X(s), s < t\}. \end{aligned}$$

By condition (ii) on the exponentially distributed sojourn times and the memoryless property of the exponential sojourn time in state i, the conditional distribution of the residual sojourn time $T_{m+1} - t$ at time t is exponentially distributed with rate q_i. Then

$$a_m(t, u) = \mathbf{1}(i = j) P\{T_{m+1} - t > u | \mathcal{F}_m(i, t)\} = \mathbf{1}(i = j) e^{-q_i u}.$$

Also, conditioning on X_{m+1} and T_{m+1} and using Remark 3,

$$b_m(t,u) = \sum_{k\neq i} \int_0^u P\{X(t+u) = j | T_{m+1} = t+v, X_{m+1} = k, \mathcal{F}_m(i,t)\}$$
$$\times P\{X_{m+1} = k, T_{m+1} - t \in dv | \mathcal{F}_m(i,t)\}$$
$$= \sum_{k\neq i} \int_0^u p_{kj}(u-v) q_i e^{-q_i v} p_{ik} dv.$$

Substituting these expressions for $a_m(i,t)$ and $b_m(i,t)$ in (4.8), and noting that they are independent of m and t, we have

$$P\{X(t+u) = j | X(t) = i, X(s), s < t\} = \mathbf{1}(i=j) e^{-q_i u} + \sum_{k\neq i} \int_0^u p_{kj}(u-v) q_i e^{-q_i v} p_{ik} dv. \tag{4.9}$$

Since this expression is true for all $t \geq 0$, by setting $t = 0$ on the left-hand side, the right-hand side must equal $p_{ij}(u)$. This proves (4.7), and it proves (4.6) as well.

Proposition 13. *If $X(t)$ is a CTMC, then its embedded process (X_n, ξ_n) satisfies conditions (i) and (ii) in Theorem 2.*

Proof. First, consider the function $g_i(t) = P_i\{\xi_1 > t\}$. For any $s, t \geq 0$,

$$g_i(s+t) = P_i\{\xi_1 > s\} P_i\{\xi_1 > s+t | \xi_1 > s\} = g_i(s) g_i(t).$$

The last equality uses the memory less property of the exponential distribution. It is well known that any nonnegative, decreasing function satisfying such an equation has the form $g_i(t) = e^{-q_i t}$, for some q_i. The q_i is a positive number since $g_i(0) = 1$, and so $P_i\{\xi_1 > t\} = e^{-q_i t}$.

Now, to prove the proposition, it suffices by Theorem 2 and Remark 3 to show that, for $i, j \in S$, $s_k, t > 0$, $n \geq 0$,

$$P\{X_{n+1} = j, \xi_{n+1} > t | X_k, \xi_k = s_k,\ k \leq n,\ X_n = i\} = p_{ij} e^{-q_i t}, \tag{4.10}$$

for some $q_i > 0$ and Markov transition probabilities p_{ij}, with $p_{ii} = 0$, $i \in S$.

We will show this is true for the q_i above and the Markov probabilities $p_{ij} = P_i\{X(\xi_1) = j\}$. Using the Markov property of $X(t)$ at the time $t_n = \sum_{k=0}^n s_k$, the probability on the left-hand side of (4.10) equals

$$P\{X(t_n + \xi_{n+1}) = j, \xi_{n+1} > t | X(t_n) = i\}$$
$$= P\{X(t_n + \xi_{n+1}) = j | X(t_n) = i\}$$
$$\times P\{\xi_{n+1} > t | X(t_n) = i, X(t_n + \xi_{n+1}) = j\}.$$

Then (4.10) is true, since the last two probabilities equal $P_i\{X(\xi_1) = j\} = p_{ij}$ and

$$\begin{aligned} &P\{X(u) = i, u \in [t_n, t_n + t] | X(v) = i, v \in [t_n, t),\ X(t_n + \xi_{n+1}) = j\} \\ &\quad = P\{X(u) = i, u \in [0, t] | X(v) = i, v \in [0, t),\ X(\xi_1) = j\} \\ &\quad = P_i\{\xi_1 > t\} = e^{-q_i t}. \end{aligned}$$

Stopping times and the strong Markov property for discrete-time Markov chains have natural analogues in continuous time.

Definition 14. A random variable τ in $[0, \infty]$ is a stopping time for a continuous-time stochastic process $\{X(t) : t \in \mathbb{R}_+\}$ if, for each $t \in \mathbb{R}_+$, the event $\{T \le t\}$ is a function of the history $\{X(s) : s \le t\}$.

Remark 15. Strong Markov Property. A CTMC satisfies the Strong Markov Property, which is (4.7) with a stopping time in place of the time parameter t. A proof is in [61].

Similarly to discrete-time Markov chains, the finite-dimensional distributions of $X(t)$ are determined by the transition probabilities and the initial distribution α_i. Namely, for each $i_1, \ldots, i_n$ in S and $0 = t_0 < t_1 < \cdots < t_n$,

$$\begin{aligned} &P\{X(t_1) = i_1, \ldots, X(t_n) = i_n\} \\ &\qquad = \sum_{i_0 \in S} \alpha_{i_0} \prod_{k=1}^{n} p_{i_{k-1}, i_k}(t_{i_k} - t_{i_{k-1}}). \end{aligned} \tag{4.11}$$

This follows by induction on n using the Markov property.

Remark 16. A consequence of (4.11) is that two CTMCs are equal in distribution if and only if their initial distributions and transition probabilities are equal.

The final observation of this section is that the transition probabilities satisfy the *Chapman-Kolmogorov equations*

$$p_{ij}(s + t) = \sum_{k \in S} p_{ik}(s) p_{kj}(t), \qquad i, j \in S,\ s, t \ge 0.$$

This follows by applying (4.11) to

$$p_{ij}(s + t) = \sum_{k \in S} P_i\{X(s) = k, X(s + t) = j\}.$$

Using matrix notation $\mathrm{P}(t) = \{p_{ij}(t)\}$, the Chapman-Kolmogorov equations are $\mathrm{P}(s + t) = \mathrm{P}(s)\mathrm{P}(t)$, which means that the family of matrices $\{\mathrm{P}(t) : t \ge 0\}$ forms a semigroup. A CTMC is sometimes defined via this semigroup associated with a process that satisfies the Markov property; additional technical conditions are needed to ensure that the chains are regular and do not have "instantaneous" states. This approach, which will not be covered here, yields the same type of CTMC we are considering.

4.4 Transition Probabilities and Transition Rates

This section continues the discussion of the transition probabilities $p_{ij}(t)$ for a CTMC $X(t)$ with defining parameters (α_i, p_{ij}, q_i). We introduce the notion of transition rates for the CTMC, and show how they are related to its transition probabilities. The main result describes Kolmogorov differential equations for the transition probabilities.

We first introduce another important family of parameters for a CTMC.

Definition 17. The *transition rates* of the CTMC $X(t)$ are

$$q_{ij} = q_i p_{ij}, \qquad j \neq i.$$

Expression (4.15) below verifies that these q_{ij} are indeed "infinitesimal" transition rates in that, for $i \neq j$,

$$p_{ij}(t) = q_{ij}t + o(t) \quad \text{as } t \to 0.$$

These rates are similar to the rates q_{ij} of the exponential clock-times in Example 9, where $p_{ij} = q_{ij}/q_i$ and $q_i = \sum_{j \neq i} q_{ij}$.

We will now describe differential equations for the transition functions. Here we simplify summations by using the negative rate

$$q_{ii} = -q_i = -\sum_{j \neq i} q_{ij}.$$

Theorem 18. *For each $i, j \in S$, the derivative $p'_{ij}(t)$ exists and is continuous in t, and*

$$\lim_{t \to 0} p_{ij}(t) = 1(i = j). \tag{4.12}$$

The $p_{ij}(t)$ satisfy the Kolmogorov differential equations

$$p'_{ij}(t) = \sum_k q_{ik} p_{kj}(t), \qquad \textit{"Backward Equation"} \tag{4.13}$$

$$p'_{ij}(t) = \sum_k p_{ik}(t) q_{kj}. \qquad \textit{"Forward Equation"} \tag{4.14}$$

In particular,

$$p'_{ij}(0) = \begin{cases} q_{ij} & \text{if } j \neq i \\ -q_i & \text{if } j = i. \end{cases} \tag{4.15}$$

Proof. Consider (4.6), which is (with the change of variable $u = t - v$)

$$p_{ij}(t) = e^{-q_i t} \left[1(i = j) + \int_0^t e^{q_i u} \sum_{k \neq i} q_{ik} p_{kj}(u)\, du \right]. \tag{4.16}$$

The integral is continuous in t since its integrand is bounded on finite intervals, and so $p_{ij}(t)$ is continuous. Then the integrand is continuous, and so the derivative of the integral exists, which in turn implies that $p_{ij}(t)$ is differentiable in t. In addition, note that the limit (4.12) follows from (4.16).

Next, taking the derivative of (4.16) and using some algebra, we have

$$p'_{ij}(t) = -q_i p_{ij}(t) + \sum_{k \neq i} q_{ik} p_{kj}(t).$$

This proves that $p'_{ij}(t)$ is continuous in t, and that (4.13) holds. Letting $t \to 0$ in this equation and using (4.12) proves (4.15).

To prove (4.14), consider the Chapman-Kolmogorov equation

$$p_{ij}(t+s) = \sum_k p_{ik}(t) p_{kj}(s).$$

Taking the derivative of this with respect to s yields

$$p'_{ij}(t+s) = \sum_k p_{ik}(t) p'_{kj}(s).$$

The derivative of the sum follows by the bounded convergence theorem in the Appendix. Letting $s \to 0$, we have $p'_{ij}(t) = \sum_k p_{ik}(t) p'_{kj}(0)$. Applying (4.15) to the last term yields (4.14).

Using matrix notation, the differential equations (4.13) and (4.14) are

$$\mathrm{P}'(t) = \mathrm{Q}\mathrm{P}(t), \qquad \mathrm{P}'(t) = \mathrm{P}(t)\mathrm{Q}.$$

The matrix $\mathrm{Q} = \{q_{ij}\}$ is the *infinitesimal generator* of the semigroup $\mathrm{P}(t)$. The unique solution to either equation (with the condition $\mathrm{P}(0) = I$) is

$$\mathrm{P}(t) = e^{t\mathrm{Q}} = \sum_{n=0}^{\infty} t^n \mathrm{Q}^n / n!.$$

This is the matrix version of the well-known solution of the equations if $\mathrm{P}(t)$ were simply a real-valued function.

There are only a few CTMCs for which the preceding infinite series expression for $\mathrm{P}(t)$ simplifies to a tractable formula. An example is the $M/M/1$ queueing process, but even this case is complicated [47] and is omitted. However, using another approach, the next section shows that the large class of CTMCs with bounded sojourn rates do indeed have tractable transition probabilities that are expectations of Poisson distributions.

We end this section with several useful facts.

Remark 19. Suppose two CTMCs with the same state space have the same initial distribution. Then the CTMCs are equal in distribution if and only if

their transition rates are equal. This follows since the CTMCs are equal in distribution if and only if their transition probabilities are equal (Remark 16); and two transition probability functions are equal if their associated transition rates are equal (Theorem 18).

We will see later in Example 50 that the mean measure of the point process $N(B) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \in B)$, $B \in \mathcal{B}$, of transition times is

$$E[N(B)] = \int_B E[q_{X(t)}]\, dt, \qquad B \in \mathcal{B}.$$

This mean measure tells us something about jump times.

Remark 20. The probability is 0 that a CTMC $X(t)$ has a jump at any fixed time t (i.e., $X(t) \neq X(t-)$, or $N(\{t\}) = 1$) This follows since $E[N(\{t\})] = 0$ by the preceding expression for the mean measure.

Here is another fact about sample paths.

Remark 21. A CTMC $X(t)$ is *continuous in distribution* in that

$$\lim_{s \to t} P\{X(s) = j\} = P\{X(t) = j\}, \qquad j \in S.$$

This follows since $P\{X(s) = j\} = \sum_i P\{X(0) = i\} p_{ij}(s)$, and Theorem 18 ensures that $p_{ij}(s) \to p_{ij}(t)$ as $s \to t$.

4.5 Existence of CTMCs

We will now establish that there exists a CTMC for any set of defining parameters. We also show how to define a CTMC by a jump process that may have fictitious jumps from a state back to itself.

Theorem 22. *There exists a CTMC for any set of defining parameters.*

Proof. Consider a set of defining parameters (α_i, p_{ij}, q_i), where the q_i are necessarily P-regular by Proposition 4.4. Let U_n, V_n, for $n \geq 0$, denote independent random variables on a common probability space such that each U_n is uniformly distributed on $[0, 1]$ and each V_n is exponentially distributed with rate 1. The existence of these random variables follows from Corollary 6 in the Appendix.

Now, let f be a function as in Theorem 14 in Chapter 1 such that $X_n = f(X_{n-1}, U_n)$ is a Markov chain with transition matrix P and initial distribution α. Next, let $\xi_n = V_n / q_{X_n}$. Then define $X(t)$ to be the jump process on S with embedded process (X_n, ξ_n).

The constructed process $X(t)$ will be a CTMC with the desired properties provided that the ξ_n satisfy conditions (4.2) and (4.3) in Theorem 2. But this follows since, by the independence of the sequences X_n and V_n,

$$P\{\xi_0 \le t_0, \dots, \xi_m \le t_m | X_n, n \ge 0\} = \prod_{n=0}^{m} P\{V_n \le t_n q_{X_n} | X_n\}$$
$$= \prod_{n=0}^{m} (1 - e^{-t_n q_{X_n}}).$$

Next, we reconsider the characterization of a CTMC in Theorem 2 and show that the condition $p_{ii} = 0$, for each i, can be relaxed. Suppose (X_n, ξ_n) is a Markov chain that satisfies assumptions (i)–(ii) in Theorem 2, with the exception that $0 \le p_{ii} < 1$ is allowed. Associated with this chain, define a jump process $X(t)$ by (4.1). For any state i where $p_{ii} > 0$, the chain X_n may jump from i back to i several times in a row, but these jumps are not seen in the process $X(t)$ — they are " fictitious" jumps.

Proposition 23. *The process $X(t)$ described above is a CTMC with one-step transition probabilities $p^*_{ii} = 0$,*

$$p^*_{ij} = p_{ij}/(1 - p_{ii}), \qquad j \ne i,$$

and exponential sojourn rates $q^*_i = (1 - p_{ii})q_i$.

Proof. The sequence of "distinct" states that $X(t)$ visits is $X^*_n = X_{\nu_n}$, where $\nu_0 = 0$ and

$$\nu_{n+1} = \min\{m > \nu_n : X_m \ne X_{\nu_n}\}, \qquad n \ge 0.$$

The sojourn time in state X^*_n is $\xi^*_n = \sum_{m=\nu_n+1}^{\nu_{n+1}} \xi_m$. Then $X(t)$ is a jump process with embedded process (X^*_n, ξ^*_n). To prove the assertion, it suffices by Theorem 2 and Remark 3 to show that (X^*_n, ξ^*_n) is a Markov chain with transition probabilities

$$P\{X^*_1 = j, \xi^*_1 > t | X^*_0 = i, \xi_0\} = p^*_{ij} e^{-q^*_i t}, \tag{4.17}$$

and that the q^*_i are P*-regular. Exercise 26 shows that ν_n are stopping times of the Markov chain (X_n, ξ_n), and so by the strong Markov property at time ν_n,

$$P\{X^*_{n+1} = j, \xi^*_n > t | X^*_n = i, X^*_m, \xi^*_m, m < n\} = P_i\{X_{\nu_1} = j, \sum_{k=1}^{\nu_1} \xi_m > t\}$$
$$= P_i\{X_{\nu_1} = j\} P_i\{\sum_{k=1}^{\nu_1} \xi_m > t | X_{\nu_1} = j\}. \tag{4.18}$$

By standard Markovian reasoning

$$P_i\{X_{\nu_1} = j\} = \sum_{\ell=1}^{\infty} p_{ii}^{\ell-1} p_{ij} = p^*_{ij}.$$

Next, recall that Exercise 5 in Chapter 3 showed that a geometric sum of i.i.d. exponential random variables is again exponential, and note that ν_1 conditioned on $X(0) = i$ has a geometric distribution with mean $1/(1 - p_{ii})$. Therefore

$$P_i\{\sum_{k=1}^{\nu_1} \xi_m > t\} = e^{-q_i(1-p_{ii})t}.$$

Applying the last two displays to (4.18) and simplifying, it follows that (X_n^*, ξ_n^*) is a Markov chain with transition probabilities (4.17).

Next, using the P-regularity of the q_i and $E[\nu_{n+1}|X_n^*] = 1/(1 - p_{X_n^*, X_n^*})$, it follows, upon conditioning on X_n^*, that

$$\infty = E[\sum_{n=0}^{\infty} 1/q_{X_n}] = E[\sum_{n=0}^{\infty} \nu_{n+1}/q_{X_n^*}] = E[\sum_{n=0}^{\infty} 1/q_{X_n^*}^*].$$

Thus the q_i^* are P*-regular.

4.6 Uniformization, Travel Times and Transition Probabilities

This section gives more insight into the transient behavior of CTMCs. We begin by describing a special CTMC whose exponential sojourn rates are all the same and consequently it has tractable transition probabilities. Remarkably, any CTMC with bounded sojourn rates is equal in distribution to a CTMC with uniform rates. The rest of the section shows how to derive travel time distributions and transition probabilities based on sums of exponential random variables.

Example 24. CTMC with Identical Sojourn Rates. Suppose that X_n is a Markov chain with transition probabilities p_{ij}, with $p_{ii} < 1$. Assume that the sojourn times ξ_n in the respective states X_n are independent exponentially distributed with rate λ. Then the state of the chain at time t can be expressed as

$$X(t) = X_{N(t)}, \qquad t \geq 0,$$

where $N(t) = \sum_n \mathbf{1}(T_n \leq t)$ is a Poisson process on $\mathbb{R}_+$ with rate λ that is independent of the X_n. The jump process $X(t)$ is sometimes called the Markov chain X_n *subordinated* to the Poisson process $N(t)$ with parameters p_{ij} and λ. Furthermore, by Proposition 23, this subordinated Markov chain $X(t)$ is a CTMC, and its main defining parameters are

$$p_{ij}^* = p_{ij}/(1 - p_{ii}), \quad j \neq i, \qquad q_i^* = \lambda(1 - p_{ii}).$$

A striking feature of the process $X(t)$ is that its transition probabilities have a tractable form. Indeed, conditioning on $N(t)$,

$$p_{ij}(t) = E[P_i\{X_{N(t)} = j | N(t)\}] = E[p_{ij}^{N(t)}].$$

Therefore,

$$p_{ij}(t) = \sum_{n=0}^{\infty} p_{ij}^n e^{-\lambda t}(\lambda t)^n / n!. \tag{4.19}$$

One can compute these probabilities by truncating the series.

Next, we establish the important result that any CTMC with bounded exponential sojourn rates can be represented as a subordinated Markov chain.

Proposition 25. (Uniformization of a CTMC) *Suppose $X(t)$ is a CTMC on S with defining parameters (α_i, p_{ij}, q_i), where the q_i are bounded. Then the process $X(t)$ is equal in distribution to a subordinated Markov chain $\hat{X}(t)$ with parameters $\hat{\alpha}_i = \alpha_i$,*

$$\hat{p}_{ij} = q_{ij}/\lambda, \quad j \neq i, \qquad \hat{p}_{ii} = 1 - q_i/\lambda, \qquad \hat{q}_i = \lambda,$$

for any fixed $\lambda \geq \sup_i q_i$.

Proof. By Proposition 23, the subordinated chain $\hat{X}(t)$ is also a CTMC with parameters

$$(\alpha_i, \frac{\hat{p}_{ij}}{1 - \hat{p}_{ii}}, \hat{q}_i(1 - \hat{p}_{ii})) = (\alpha_i, p_{ij}, q_i).$$

Since the latter are the parameters for $X(t)$, it follows by Proposition 4 that the processes $X(t)$ and $\hat{X}(t)$ have the same distribution.

One consequence of this uniformization principle is that any CTMC with bounded sojourn rates has transition probabilities of the form (4.19). This is true for any birth-death process with bounded rates, such as an $M/M/s$ system with finite s. In addition, uniformization is a key tool in Markov decision theory (e.g., [90, 54, 109]), where a continuous-time Markov decision process is usually formulated by a simpler discrete-time Markov decision process.

We now turn to another approach for deriving transition probabilities that involves evaluating sums of independent exponential sojourn times associated with a path of states. A tool for this is the following proposition, which is of general interest; Exercise 9 in Chapter 3 is a related result.

Proposition 26. *If Y_k, $k \in I$, are independent exponentially distributed random variables with rates q_k that are distinct, then the distribution of $Z = \sum_{k \in I} Y_k$ is the mixture of exponential distributions*

$$F_Z(t) = \sum_{k\in I}(1 - e^{-q_k t}) \prod_{\ell\in I,\ell\neq k} \frac{q_\ell}{q_\ell - q_k}, \qquad t \geq 0. \tag{4.20}$$

Proof. Consider the Laplace transform

$$L(\alpha) = E[e^{-\alpha Z}] = \prod_{k\in I} E[e^{-\alpha Y_k}] = \prod_{k\in I} \frac{q_k}{\alpha + q_k}.$$

This is a ratio of polynomials in α, in which the denominator has distinct roots $-q_k$, $k \in I$. Therefore, its partial-sum expansion is

$$L(\alpha) = \sum_{k\in I} c_k \frac{q_k}{\alpha + q_k}, \tag{4.21}$$

where

$$c_k = (\alpha + q_k)L(\alpha)\Big|_{\alpha=-q_k} = \prod_{\ell\in I,\ell\neq k} \frac{q_\ell}{q_\ell - q_k}.$$

Then the distribution with Laplace transform (4.21) is the mixture[3] of exponential distributions given by (4.20).

Example 27. Travel Times on Paths. Let Z_I denote the travel time of a CTMC $X(t)$ on a path of states $I = (i_0, i_1, \ldots, i_m)$ whose exponential sojourn rates are distinct. Then the distribution of Z_I is given by (4.20). Moreover, the travel time Z in a set $\mathcal{P}$ of such paths has the distribution

$$P\{Z \leq t\} = \sum_{I\in\mathcal{P}} p_I F_{Z_I}(t),$$

where $p_I = P\{X_0 = i_0\}p_{i_0,i_1} \cdots p_{i_{m-1},i_m}$ is the probability of traversing I.

Example 28. Machine Deterioration Model. Consider a CTMC $X(t)$ that represents the state of deterioration of a machine at time t, where the set of states is $S = \{0, 1, \ldots, \ell\}$. Assume that its deterioration is nondecreasing ($p_{ij} = 0$, $j \leq i \leq \ell$), and that $p_{\ell,0} = 1$, so after its stay in state ℓ, it is replaced by a new machine. Then the lifetime of a machine is its travel time Z in the set $\mathcal{P}$ of all nondecreasing paths from 0 to ℓ. Thus the machine life-time distribution is as in Example 27.

Example 29. If $X(t)$ is a pure birth process as in Example 7 with distinct sojourn rates q_i, then its transition probabilities are

$$p_{ij}(t) = \sum_{k=i}^{j-1} [e^{-q_j t} - e^{-q_k t}] \frac{q_k}{q_k - q_j} \prod_{\substack{\ell=i \\ \ell\neq k}}^{j-1} \frac{q_\ell}{q_\ell - q_k}.$$

[3] Note that the coefficients c_k may be negative and do not necessarily sum to 1.

To see this, consider the travel time $Z = \sum_{k \in I} Y_k$ through the states $I = (i, i+1, \ldots, j-1)$, where Y_i are the independent exponential sojourn times in the states. Then

$$p_{ij}(t) = P\{Z \le t < Z + Y_j\} = \int_0^t P\{Y_j > t - s\} F_Z(ds).$$

Substituting the exponential distribution for Y_j and the distribution (4.20) in this integral and integrating proves the assertion.

Sometimes transition probabilities can be determined by exploiting properties of Poisson processes as follows.

Example 30. $M/M/\infty$ System. Suppose $X(t)$ is an $M/M/\infty$ process with Poisson arrival rate λ and exponential service rate μ. To derive its transition probabilities at time t, we will consider two independent populations of items: those present at time 0, and those arrivals in $(0, t)$ that are still in the system at time t. Now, each item present at time 0 is still in the system at time t with the probability $e^{-\mu t}$ that the residual service time exceeds t (the exponential distribution is memoryless). Then if there are i items present at time 0, the number of these still present at time t has the Binomial distribution with parameters i and $1 - e^{-\mu t}$.

Next, note that since this process is a special case of the $M/G/\infty$ process described in Chapter 3, we know that the number of the Poisson arrivals in $(0, t)$ that are still present at time t has a Poisson distribution with mean

$$\eta(t) = \int_0^t e^{-\mu(t-s)} \lambda ds = (\lambda/\mu)(1 - e^{-\mu t}).$$

Then the number of items in the system at time t is the sum of these independent binomial and Poisson random variables. Consequently, the transition probabilities are

$$p_{ij}(t) = \sum_{k=0}^{i} \binom{i}{k} e^{-k\mu t} (1 - e^{-\mu t})^{i-k} \frac{\eta(t)^{j-k}}{(j-k)!} e^{-\eta(t)}.$$

4.7 Stationary and Limiting Distributions

The classification of states of a CTMC and its equilibrium behavior are closely related to those properties of its embedded Markov chain. We will now characterize stationary and limiting distributions for CTMCs by applying results for discrete-time Markov chains in Chapter 1.

For this discussion $X(t)$ will denote a CTMC on S with embedded chain (X_n, ξ_n) and defining parameters (α_i, p_{ij}, q_i). Also, q_{ij} and $p_{ij}(t)$ will denote its transition rates and transition probabilities.

Our first concern is a classification of states. This mirrors the classification for the embedded chain X_n. A state $i \in S$ is *recurrent* (or *transient*) for $X(t)$ if i is recurrent (or transient) for X_n. The process $X(t)$ is *irreducible* if X_n is. Recall the convention that X_n does not have any absorbing states; consequently, neither does $X(t)$. CTMCs do not have periodic states, because their sojourn times in states are continuous random variables. However, their embedded chains may be periodic in discrete time.

To describe positive recurrence, we will use the discrete- and continuous-time first passage times $\nu_i = \min\{n \geq 1 : X_n = i\}$ and

$$\tau_i = \inf\{t > \xi_0 | X(t) = i\} = \sum_{n=0}^{\nu_i - 1} \xi_n.$$

A recurrent state i for $X(t)$ is *positive recurrent* or *null recurrent* according as the mean $E_i[\tau_i]$ is finite or infinite. Finally, $X(t)$ is *ergodic* if it is irreducible and all of its states are positive recurrent.

In relating $X(t)$ to its embedded chain X_n, keep in mind that $X(t)$ is irreducible and recurrent if and only if X_n is irreducible and recurrent. On the other hand, Exercise 28 shows that $X(t)$ may be positive recurrent while X_n is not, and vice versa. Of course, if the transition rates are bounded away from 0 and ∞, then $X(t)$ is ergodic if and only if X_n is.

Our last preliminary is a formula for relating mean cycle values of $X(t)$ to those of X_n.

Lemma 31. *If $X(t)$ is irreducible and recurrent, then, for $f : S \to \mathbb{R}$,*

$$E_i[\int_0^{\tau_i} f(X(t))\, dt] = E_i[\sum_{n=0}^{\nu_i - 1} f(X_n)/q_{X_n}], \tag{4.22}$$

provided these means exist.

Proof. Using the pull-through property, the left side of (4.22) is

$$E_i[\sum_{n=0}^{\nu_i - 1} f(X_n)\xi_n] = E_i\Big[\sum_{n=0}^{\infty} \mathbf{1}(\nu_i > n) f(X_n) E_i[\xi_n | X_k,\, k \leq n]\Big]$$

and the last expression equals the right side of (4.22).

We will now describe stationary and invariant measures for CTMCs.

Definition 32. A probability measure p on S is a *stationary distribution* for $X(t)$ if

$$p_{ij}(t) = \sum_{k=0}^{i} \binom{i}{k} (1 - e^{-\mu t})^k e^{-(i-k)\mu t} \frac{\eta(t)^{j-k}}{(j-k)!} e^{-\eta(t)}.$$

This equation in matrix notation is $p = p\mathrm{P}(t)$, where $p = (p_i : i \in S)$ is a row vector. More generally, any measure γ on S that satisfies $\gamma = \gamma \mathrm{P}(t)$, $t \geq 0$, is an *invariant measure* for $X(t)$.

As in discrete time, a stationary distribution is sometimes called an *equilibrium distribution*, and it is related to the notion of a stationary process. The proof of the following result parallels that of Proposition 52 in Chapter 1. Recall that a continuous-time process $\{X(t) : t \geq 0\}$ on a space S is *stationary* if, for any $s_1 < \cdots < s_m$,

$$(X(s_1 + t), \ldots, X(s_m + t)) \stackrel{d}{=} (X(s_1), \ldots, X(s_m)), \qquad t \geq 0.$$

Proposition 33. *For the CTMC $X(t)$, which need not be irreducible or recurrent, the following statements are equivalent.*
(a) *$X(t)$ is a stationary process.*
(b) *$X(t) \stackrel{d}{=} X(0)$, $t > 0$.*
(c) *The distribution of $X(0)$ is a stationary distribution.*

We will now establish the existence of an invariant measure for $X(t)$. Recall the transition-rate notation $Q = \{q_{ij}\}$, where $q_{ij} = q_i p_{ij}$ $(j \neq i)$, and

$$q_{ii} = -q_i = \sum_{j \neq i} q_{ij}.$$

Theorem 34. *For an irreducible, recurrent CTMC $X(t)$ and a nonnegative measure γ on S, the following statements are equivalent.*
(a) *γ is an invariant measure for $X(t)$.*
(b) *$\gamma_i q_i$ is an invariant measure for X_n.*
(c) *γ satisfies the balance equations $\gamma Q = 0$, or equivalently,*

$$\gamma_i \sum_{j \neq i} q_{ij} = \sum_{j \neq i} \gamma_j q_{ji}, \qquad i \in S. \tag{4.23}$$

Furthermore, for a fixed i, the measure γ defined by

$$\gamma_j = E_i\Big[\int_0^{\tau_i} \mathbf{1}(X(t) = j)\, dt\Big], \qquad j \in S, \tag{4.24}$$

is a positive invariant measure for $X(t)$, and this measure is unique up to a multiplication by a constant.

Proof. (a) $\Leftrightarrow$ (c). For any measure γ on S, using the representation $\mathrm{P}(t) = e^{tQ}$, we have

$$\gamma \mathrm{P}(t) = \gamma e^{tQ} = \gamma + \sum_{k=1}^{\infty} \frac{t^k}{k!} \gamma Q^k.$$

Thus γ is an invariant measure for $X(t)$ (i.e., $\gamma \mathrm{P}(t) = \gamma$) if and only if $\gamma Q^k = 0$ for each k. But the latter (by induction) is equivalent to $\gamma Q = 0$.

(b) $\Leftrightarrow$ (c). Clearly (b) implies (c) since using $p_{ij} = q_{ij}/q_i$,

$$\gamma_i q_i = \sum_{j \neq i} \gamma_j q_j p_{ji} = \sum_{j \neq i} \gamma_j q_{ji}.$$

Reversing the last two sums proves that (c) implies (b).

Theorem 35. *For an irreducible, recurrent CTMC $X(t)$ and a positive distribution measure p on S, the following statements are equivalent.*
(a) *$X(t)$ is ergodic with stationary distribution p.*
(b) *p satisfies the equations $pQ = 0$.*
(c) *For a fixed i, $E_i[\tau_i]$ is finite, and*

$$p_j = \frac{1}{E_i[\tau_i]} E_i\Big[\int_0^{\tau_i} \mathbf{1}(X(t) = j)\, dt \Big], \qquad j \in S. \tag{4.25}$$

Proof. The equivalence of (a) and (b) follows by Theorem 34 (statements (a) and (c)).

Next, suppose $X(t)$ is ergodic with stationary distribution p. Then each $E_i[\tau_i]$ is finite and p is a positive distribution. Also, by Theorem 34, p must be a multiple of γ in (4.24), and so, for a fixed i, we have $p_j = c_i \gamma_j$ (c_i as well as γ_j depends on i). Then summing this on j yields $1 = c_i \sum_j \gamma_j = c_i E_i[\tau_i]$. These observations imply $p_j = \gamma_j / E_i[\tau_i]$, which is equivalent to (4.25). This proves that (a) implies (c).

Finally, suppose (c) holds. The p given by (4.25) is positive and it is a multiple of the invariant measure γ in (4.24). In addition, it clearly sums to 1. Therefore it is a stationary distribution for $X(t)$. This proves that (c) implies (a).

Remark 36. Another representation of the distribution in (4.25) is

$$p_j = \frac{1}{q_j E_j[\tau_j]} \qquad j \in S.$$

This follows from (4.25) by setting $i = j$.

Recall that the preceding results do not require that X_n be ergodic. If it is, however, the stationary distributions of $X(t)$ and X_n are related as follows.

Proposition 37. *If $X(t)$ and X_n are ergodic with respective stationary distribution p and π, then*

$$p_j = \frac{\pi_j/q_j}{\sum_k \pi_k/q_k}, \qquad j \in S.$$

Proof. Remark 36 tell us that $p_j = 1/(q_j E_j[\tau_j])$. Also, by Lemma 31 with $f(\cdot) = 1$ and Proposition 69 in Chapter 1,

$$E_j[\tau_j] = E_j\Big[\sum_{n=0}^{\nu_j - 1} q_{X_n}^{-1}\Big] = \pi_j^{-1} \sum_k \pi_k/q_k.$$

Combining these two observations proves the assertion.

Before getting into applications of stationary distributions, we will relate them to limiting distributions. We begin by describing the regenerative property of CTMCs, which is comparable to Proposition 67 of Chapter 1 for discrete-time chains. The proof, like that in discrete time, is by an application of the strong Markov property in continuous time.

Proposition 38. *An irreducible, recurrent CTMC $X(t)$ is a delayed regenerative process over the times at which the process enters a fixed state.*

We end this section by establishing that the stationary distribution of an ergodic CTMC is also its limiting distribution. The *limiting distribution* of the CTMC $X(t)$ is defined by

$$p_j = \lim_{t\to\infty} p_{ij}(t), \qquad i \in S,$$

provided the limits exist and do not depend on i.

Theorem 39. (Limiting Distribution) *If the CTMC $X(t)$ is ergodic, then its stationary distribution given by (4.25) is its limiting distribution.*

Proof. By Proposition 38, $X(t)$ is a delayed regenerative process over the hitting times of a state i. Also, the i.i.d. times between hitting i have a nonarithmetic distribution due to the exponential sojourn times in the states. Then by the characterization of limiting distributions for regenerative processes in Corollary 46 in Chapter 2, it follows that the limiting distribution of $X(t)$ is the same as the stationary distribution given by (4.25).

4.8 Regenerative Property and Cycle Costs

One use of stationary distributions is in evaluating means of functionals of a CTMC in between successive visits to a fixed state. This section contains formulas for two such "cycle" means. The formulas are of interest by themselves and are used to evaluate limiting averages of functionals, as we will see in the next section. An analogous result for Markov chains is Proposition 69 in Chapter 1.

The first formula is for a cycle cost or utility that is an integral of a rate function.

Proposition 40. *If $X(t)$ is an ergodic CTMC with stationary distribution p, then, for $f : S \to \mathbb{R}$,*

$$E_i\Big[\int_0^{\tau_i} f(X(t))\,dt\Big] = \frac{1}{p_i q_i}\sum_j f(j)p_j. \tag{4.26}$$

provided the sum is absolutely convergent.

Proof. The left-hand side of (4.26) equals $\sum_j f(j)E_i\Big[\int_0^{\tau_i} \mathbf{1}(X(t) = j)\,dt\Big]$, and this equals the right-hand side of (4.26) by (4.25) and Remark 36.

The next formula is for a cycle cost for a process with regenerative increments; see Section 2.10.

Proposition 41. *Associated with as ergodic CTMC $X(t)$ with stationary distribution p, suppose that $\{Z(t) : t \geq 0\}$ is a real-valued stochastic process with delayed regenerative increments over the times at which $X(t)$ enters a fixed state i, and $Z(0) = 0$. Assume that*

$$E_i\Big[Z(T_{n+1}) - Z(T_n)\Big|X_m, m \leq n\Big] = h_i(X_n), \qquad i \in S,\ n \geq 0,$$

for some $h_i : S \to \mathbb{R}$. Then

$$E_i[Z(\tau_i)] = \frac{1}{p_i q_i}\sum_j h_i(j)p_j q_j, \tag{4.27}$$

provided the sum is absolutely convergent.

Proof. Since $\{\nu_i > n\}$ is a function of $X_0, \ldots, X_n$, using the pull-through property for conditional probabilities and the hypotheses,

$$\begin{aligned} E_i[Z(\tau_i)] &= E_i\Big[\sum_{n=0}^{\nu_i-1}[Z(T_{n+1}) - Z(T_n)]\Big] \\ &= E_i\Big[\sum_{n=0}^{\infty}\mathbf{1}(\nu_i > n)E_i\Big[[Z(T_{n+1}) - Z(T_n)]\Big|X_0,\ldots,X_n\Big]\Big] \\ &= E_i\Big[\sum_{n=0}^{\nu_i-1}h_i(X_n)\Big]. \end{aligned}$$

Applying Lemma 31 to the last term and then using Proposition 40, we obtain

$$E_i[Z(\tau_i)] = E_i\Big[\int_0^{\tau_i} h_i(X(t))q_{X(t)}\,dt\Big] = \frac{1}{p_i q_i}\sum_j h_i(j)p_j q_j.$$

4.9 Ergodic Theorems

In this section, we present several SLLNs for CTMCs that apply to a variety of functionals associated with cost and performance parameters. These results follow from the general SLLN for processes with regenerative increments covered in Chapter 2. Included are insights on the rate of convergence of the SLLNs, based on the refined SLLN in Chapter 2.

For this development, $X(t)$ will be an ergodic CTMC with stationary distribution p, and τ_i is the hitting time of state i.

Theorem 42. *Suppose that $\{Z(t) : t \geq 0\}$ is a real-valued stochastic process that has delayed regenerative increments over the times at which $X(t)$ enters a fixed state i, and $Z(0) = 0$. Assume $E_i[\sup_{t\leq\tau_i}|Z(t)|]$ is finite, and*

$$E_i\Big[Z(T_{n+1}) - Z(T_n)\Big|X_m, m \leq n\Big] = h(X_n), \qquad n \geq 0,$$

for some $h : S \to \mathbb{R}$. Then assuming the sum is absolutely convergent,

$$\lim_{t\to\infty} t^{-1}Z(t) = \sum_j h(j)q_j p_j \qquad \textit{a.s.}$$

Proof. By the SLLN for processes with regenerative increments (Theorem 54 in Chapter 2), we have

$$\lim_{t\to\infty} t^{-1}Z(t) = E_i[Z(\tau_i)]/E_i[\tau_i] \qquad \text{a.s.}$$

But this limit is finite and equals $\sum_j h(j)q_j p_j$ by (4.27) and Remark 36.

The process $Z(t)$ in the preceding theorem is a general model for various functionals of the CTMC $X(t)$. Here are some illustrations.

Example 43. Jump Times. Consider the process $N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \le t)$, which records the number of jump times T_n in $(0, t]$. This point process is of interest by itself and it is used in functionals based on jump times.

The average number of jumps per unit time (or rate of $N(t)$) is

$$\lim_{t\to\infty} t^{-1} N(t) = \sum_j p_j q_j \quad \text{a.s.},$$

provided the sum is finite. This follows by Theorem 42 with $Z(t) = N(t)$ and $Z(T_{n+1}) - Z(T_n) = 1$.

In addition, knowing the rate of $N(t)$, Theorem 10 in Chapter 2 tells us that the average sojourn time over all the states determined by the X_n is

$$\lim_{n\to\infty} n^{-1} \sum_{m=0}^{n-1} \xi_m = 1/\sum_j p_j q_j \quad \text{a.s.}$$

Example 44. Integral Functionals. Suppose $Z(t) = \int_0^t V(s)\, ds$, where $V(t)$ is a delayed regenerative process over the times at which $X(t)$ enters a fixed state. Then the integral process $Z(t)$ has delayed regenerative increments over these times. The $V(t)$ might be a cost or utility rate at time t associated with a CTMC and auxiliary environmental information, such as $V(t) = g(X(t), Y(t))$, for $g : S \times S' \to \mathbb{R}$. The limit statement in Theorem 42 will be true for a particular application upon verifying the other assumptions in the theorem.

The preceding example sets the stage for the following classical result.

Theorem 45. (SLLN for CTMCs). *For the ergodic CTMC $X(t)$ with stationary distribution p, and a function $f : S \to \mathbb{R}$,*

$$\lim_{t\to\infty} t^{-1} \int_0^t f(X(s))\, ds = \sum_j f(j) p_j \quad \textit{a.s.}, \tag{4.28}$$

provided the sum is absolutely convergent.

Proof. First note that, for $n \ge 0$,

$$E_i\Big[\int_{T_n}^{T_{n+1}} f(X(t))\, dt \Big| X_m, m \le n \Big] = E_i[f(X_n)\xi_n | X_n] = f(X_n) q_{X_n}^{-1}. \tag{4.29}$$

This also holds with $|f(j)|$ in place of $f(j)$. Then using (4.27),

$$E_i[\sup_{t\le\tau_i} |Z(t)|\,] \le E_i\Big[\int_0^{\tau_i} |f(X(t))|\, dt \Big] = \frac{1}{p_i q_i} \sum_j |f(j)| p_j. \tag{4.30}$$

This quantity is finite by assumption. Thus (4.28) follows by Theorem 42.

The preceding SLLN has the following rate of convergence.

Example 46. Refined SLLN. In the context of the preceding theorem, assume the time between entrances to a fixed state i has a finite variance σ_i^2, and let $\mu_i = E_i[\tau_i] = 1/q_i p_i$. Then

$$E\Big[\int_0^t f(X(s))\,ds\Big] = [t + (\sigma_i^2 + \mu_i^2)/2\mu_i] \sum_j f(j)p_j - \sum_j f(j)p_j/q_j + o(1), \quad \text{as } t \to \infty. \tag{4.31}$$

In particular, the expected time spent in state i has the asymptotic behavior

$$E\Big[\int_0^t \mathbf{1}(X(s) = i)\,ds\Big] = tp_i + p_i[(\sigma^2 + \mu_i^2)/2\mu_i - 1/q_i] + o(1), \quad \text{as } t \to \infty.$$

Applying Theorem 85 in Chapter 2 to $Z(t) = \int_0^t f(X(s))\,ds$ yields (4.31); the required constants are (see Corollary 40)

$$a = E_i[Z(\tau_i)] = \mu_i \sum_j f(j)p_j,$$

$$c = -\frac{1}{\mu_i} E_i\Big[\int_0^{\tau_i} f(X(t))\,dt\Big] = -\sum_j f(j)p_j/q_j.$$

Example 44 and Theorem 45 concerned integrals of cost or utility "rates". Here is another type of functional associated with transition times.

Example 47. Functionals of Embedded Processes. Suppose $Z(t) = \sum_{n=0}^{N(t)} V_n$, where V_n is a cost or parameter associated with the jump time T_n. As an example, suppose that $g(i, j, t, y)$ is a cost incurred at the beginning of a sojourn time of length t in state i that ends with a jump to state j, and y is an auxiliary variable. Then

$$V_n = g(X_n, X_{n+1}, \xi_n, Y_n), \qquad n \geq 0,$$

is the cost incurred at time T_n, where Y_n are auxiliary random variables and $T_0 = 0$. Assume the Y_n are i.i.d. and independent of the CTMC. By the exponential sojourn time and jump probabilities p_{jk}, the expected cost incurred at time T_n in state $X_n = j$ is

$$\begin{aligned} h(j) &= E[V_n | X_m, m \leq n-1, X_n = j] \\ &= \sum_{k \neq j} p_{jk} \int_0^\infty E[g(j, k, t, Y_1)] q_j e^{-q_j t}\,dt. \end{aligned}$$

Assume this exists and is finite when g is replaced by $|g|$.

Under these assumptions, the average cost is

$$\lim_{t\to\infty} t^{-1}\sum_{n=0}^{N(t)} V_n = \sum_j h(j)q_j p_j \qquad \text{a.s.}$$
$$= \sum_j p_j \Big[\sum_{k\neq j} q_{jk}\int_0^\infty E[g(j,k,t,Y_1)]q_j e^{-q_j t}\,dt\Big], \quad (4.32)$$

provided the sum is absolutely convergent. In particular,

$$\lim_{t\to\infty} t^{-1}\sum_{n=0}^{N(t)} g(X_n, X_{n+1}) = \sum_j p_j \sum_{k\neq j} q_{jk} g(j,k) \quad \text{a.s.}, \qquad (4.33)$$

Expression (4.32) will follow by Theorem 42 upon verifying its assumptions. Clearly,

$$E_i[Z(T_{n+1}) - Z(T_n)|X_m, m \le n] = E_i[V_n|X_n] = h(X_n),$$

Also,

$$E_i[\sup_{t\le\tau_i}|Z(t)|\,] \le E_i\Big[\sum_{n=0}^{N(\tau_i)}|V_n|\,\Big],$$

and, similarly to (4.30), the last term is finite when the right-hand side of (4.32) is absolutely convergent. Thus Theorem 42 yields (4.32).

A typical application of the SLLN (4.32) involves defining V_n by an appropriate function g, evaluating the mean function $h(j)$, and verifying the absolute convergence of the sum $\sum_j h(j)q_j p_j$. Here is an illustration.

Example 48. Consulting Company. Potential projects arrive to a consulting company according to a Poisson process with rate λ. The times to do the projects are independent exponentially distributed with rate μ. The company is small and can only handle one project at a time. Therefore, when it is working on a project, any potential projects that arrive are rejected.[4] Suppose the revenues from completing the projects are i.i.d. with mean α and the revenue from the rejected projects are i.i.d. with mean $\bar{\alpha}$ (which may be higher than α due to a tarnished image of the company). Of interest are the average revenue from completing the projects and the average revenue lost from the rejected projects.

To derive these averages, let $X(t)$ denote the state of the company at time t, where the states are 1 if a job is in service and 0 otherwise. Clearly $X(t)$ is a CTMC with transition rates $q_{10} = \mu$ (the service rate) and $q_{01} = \lambda$ (the rate of the Poisson arrival process, which has stationary independent increments).

[4] This model, which is sometimes called an $M/M/1/1$ system, might also be appropriate for any service station or computer that can only work on one job at a time.

It is easily seen that the stationary distribution of this two-state CTMC is[5]

$$p_0 = \lambda^{-1}/(\lambda^{-1} + \mu^{-1}) = \mu/(\lambda + \mu), \qquad p_1 = 1 - p_0 = \lambda/(\lambda + \mu).$$

The revenue received at time T_n for completing a project when the company is in state X_n is $V_n = Y_n \mathbf{1}(X_n = 0)$, where $Y_1, Y_2, \ldots$ are i.i.d. project revenues, independent of the CTMC, with mean α. Although this revenue may be received any time during the engagement of a project, we assume, with no loss in generality, that it is received at the beginning of a project. Then it follows from (4.33) that the average revenue from completing the projects is

$$\lim_{t\to\infty} t^{-1} \sum_{n=0}^{N(t)} V_n = \alpha p_0 \quad \text{a.s.}$$

To formulate the lost revenue from rejected projects, note that when the company is busy on a project, the lost revenue can be represented by a compound Poisson process $M(t)$ with rate λ and the distribution of an increment is that of a single lost-project with mean $\bar{\alpha}$. Then the lost-project revenue at time T_n is $V'_n = M_n(\xi_n)\mathbf{1}(X_n = 1)$, where M_n are i.i.d. compound Poisson processes with $M_n \stackrel{d}{=} M$ that are independent of the (X_n, ξ_n). Although the arrival times of rejected projects occur during a sojourn time in state 1, we assume, with no loss in generality, that their lost-project revenues are incurred at time T_n. Note that V'_n has the form $V'_n = g(X_n, \xi_n, M_n)$ as in (4.32) and

$$h(j) = E[V'_n | X_n = j] = E[M_n(\xi_n)|X_n = j]\bar{\alpha}\mathbf{1}(j = 1) = \lambda\mu^{-1}\bar{\alpha}\mathbf{1}(j = 1).$$

Therefore, by (4.32), the average lost-revenue is

$$\lim_{t\to\infty} t^{-1} \sum_{n=0}^{N(t)} V'_n = \lambda\mu^{-1}\bar{\alpha}p_1 \quad \text{a.s.}$$

One could use the preceding information to maximize profit. For instance, suppose the service rate μ could be varied by changing the number of workers or amount of overtime, and the cost per unit time of having a rate μ is $C(\mu)$. Then the average revenue would be $\alpha p_0 - \lambda\mu^{-1}\bar{\alpha}p_1 - C(\mu)$. One could then select μ to maximize this average revenue.

Another quantity of interest is the average number of projects that take longer than d days to complete. As a slightly different application of (4.32), the average number of projects that take longer than d days to complete is

$$\lim_{t\to\infty} t^{-1} \sum_{n=0}^{N(t)} \mathbf{1}(X_n = 1, \xi_n > d) = p_1 e^{-d\mu} \quad \text{a.s.}$$

[5] The two-state Markov chain was covered in Exercise 10 in Chapter 1.

Our next example concerns rates of several types of transitions associated with the ergodic CTMC $X(t)$ with stationary distribution p. Analogous results for discrete-time are in Section 1.13.

Example 49. Movement Between Sets. The average number of transitions the CTMC makes from a set of states A to a set of states B per unit time is

$$\lambda(A,B) = \lim_{t\to\infty} t^{-1} \sum_{n=0}^{N(t)} \mathbf{1}(X_n \in A, X_{n+1} \in B).$$

By (4.33) this limit exists and equals

$$\lambda(A,B) = \sum_{i\in A} p_i \sum_{j\in B} q_{ij}.$$

In particular, the rate at which the CTMC moves from i to j is $\lambda(i,j) = p_i q_{ij}$.

Recall that the total balance equations for the CTMC are

$$p_i \sum_{j\neq i} q_{ij} = \sum_j p_j q_{ji}, \qquad i \in S.$$

This is equivalent to $\lambda(i,S) = \lambda(S,i)$. That is, the rate at which the CTMC moves out of state i is equal to the rate at which it moves into i. More generally, it follows as in Chapter 1 that $\lambda(A,A^c) = \lambda(A^c,A)$, the rate at which the chain moves out of a set A is the rate at which it moves into A.

The number of entrances the CTMC makes into a set $A \subset S$ up to time t is defined by

$$N_A(t) = \sum_{n=0}^{N(t)} \mathbf{1}(X_n \notin A, X_{n+1} \in A).$$

Then as above, the average number of entrances per unit time into A is

$$\lambda(A) = \lim_{t\to\infty} t^{-1} N_A(t) = \sum_{i\in A^c, j\in A} p_i q_{ij} \quad \text{a.s.}$$

This is the same as the rate $\lambda(A^c,A)$ at which the chain enters A.

A related quantity is the nth time the chain enters the set A, which is $\tau_A(n) = \min\{t : N_A(t) = n\}$. By Theorem 10 in Chapter 3, the limiting average of these times is $\lim_{t\to\infty} n^{-1}\tau_A(n) = 1/\lambda(A)$ a.s.

4.10 Expectations of Cost and Utility Functions

We have been discussing limiting averages for functionals of CTMCs. We now switch back to finite time and present an important formula for means

of functionals of CTMCs. In addition to being useful for cost functions, the formula is a key tool for our analysis of Palm probabilities later in Sections 4.15 and 4.16.

Here $X(t)$ will denote a CTMC with the usual notation, but with no other assumptions about it being ergodic, etc. We begin with two examples of the upcoming main result.

Example 50. Mean Measure of Transition Times. Consider the point process $N(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \le t)$ of transition times of $X(t)$. Its mean by (4.35) is

$$E[N(t)] = \int_0^t E[q_{X(s)}]\, ds, \qquad t \ge 0,$$

provided the integral exists. This mean is finite if $\sum_i q_i < \infty$. In particular, if $X(t)$ is stationary with distribution p, then

$$E[N(t)] = tE[N(1)] = t \sum_i p_i q_i.$$

Example 51. Discounted Rewards. Suppose $X(t)$ is in equilibrium with stationary measure p, and a discounted reward $r(i)e^{-\alpha t}$ is received whenever it enters state i at time t. Then the expected total discounted reward over the infinite horizon is

$$\begin{aligned} E\Big[\sum_{n=1}^{\infty} r(X_n)e^{-\alpha T_n}\Big] &= \int_{\mathbb{R}_+} E[q_{X(t)} r(X(t))]e^{-\alpha t}\, dt \\ &= \alpha^{-1} \sum_i p_i q_i r(i). \end{aligned}$$

The preceding are examples of the following main result, which is formula (4.35) for the mean of various functions of $X(t)$. This result is a generalization of the Lévy formula in Example 53 below.

Theorem 52. *Suppose $\{V_n : n \ge 1\}$ are real-valued random variables associated with the CTMC $X(t)$. Assume that the conditional expectations*

$$h(\zeta_n) = E[V_n | \zeta_n], \qquad n \ge 1, \tag{4.34}$$

exist, where $\zeta_n = (T_n, X_{n-1}, X_n)$ and $h : \mathbb{R}_+ \times S^2 \to \mathbb{R}$ is independent of n. Then

$$E\Big[\sum_{n=1}^{\infty} V_n\Big] = \int_{\mathbb{R}_+} E\Big[\sum_{j \ne X(t)} q_{X(t),j} h(t, X(t), j)\Big]\, dt, \tag{4.35}$$

provided the last integral is finite with $|h|$ in place of h.

Proof. By Exercise 8 in Chapter 3, we know that if ξ is an exponential random variable with rate λ, then $E[h(\xi)] = \lambda E[\int_0^{\xi} h(u)\, du]$. Using an analogous

expression for conditional expectations applied to ξ_{n-1}, along with (4.34) and Markov properties of (X_n, ξ_n), we have

$$\begin{aligned} E[V_n] &= E\Big[E[V_n|\zeta_n]\Big] \;=\; E[h(\zeta_n)] \\ &= E\Big[E[h(T_{n-1}+\xi_{n-1}, X_{n-1}, X_n)|T_{n-1}, X_{n-1}, X_n]\Big] \\ &= E\Big[q_{X_{n-1}} \int_0^{\xi_{n-1}} h(T_{n-1}+u, X_{n-1}, X_n)\, du\Big] \\ &= E\Big[q_{X_{n-1}} \int_{T_{n-1}}^{T_n} h(t, X_{n-1}, X_n)\, dt\Big]. \end{aligned}$$

The change-of-variable $t = T_{n-1} + u$ is used for the last line. Under further conditioning,

$$\begin{aligned} E[V_n] &= E\Big[\sum_{j\neq X_{n-1}} E\Big[q_{X_{n-1}} \int_{T_{n-1}}^{T_n} h(t, X_{n-1}, j)\, dt \Big| X_{n-1}, T_{n-1}, X_n = j\Big] \\ &\qquad\qquad \times P\{X_n = j|T_{n-1}, X_{n-1}\}\Big] \\ &= E\Big[\sum_{j\neq X_{n-1}} q_{X_{n-1},j} \int_{T_{n-1}}^{T_n} h(t, X_{n-1}, j)\, dt\Big]. \end{aligned}$$

The last line uses the fact that the conditional probability in the second line is $q_{X_{n-1},j}/q_{X_{n-1}}$. Finally, using $X_{n-1} = X(t)$ for $t \in [T_{n-1}, T_n)$

$$\begin{aligned} E\Big[\sum_{n=1}^{\infty} V_n\Big] &= E\Big[\sum_{n=1}^{\infty} \sum_{j\neq X_{n-1}} q_{X_{n-1},j} \int_{T_{n-1}}^{T_n} h(t, X_{n-1}, j)\, dt\Big] \\ &= \int_{\mathbb{R}_+} E\Big[\sum_{j\neq X(t)} q_{X(t),j} h(t, X(t), j)\Big]\, dt. \end{aligned}$$

Here are two examples.

Example 53. Lévy Formula. For $f : S^2 \to \mathbb{R}$ and $t \geq 0$,

$$E\Big[\sum_{n=1}^{N(t)} f(X_{n-1}, X_n)\Big] = \int_0^t E\Big[\sum_{j\neq X(s)} q_{X(s),j} f(X(s), j)\Big] ds,$$

provided the last integral is finite with $|f|$ in place of f. This formula is a special case of Theorem 52, since

$$\sum_{n=1}^{N(t)} f(X_{n-1}, X_n) = \sum_{n=1}^{\infty} f(X_{n-1}, X_n)\mathbf{1}(T_n \leq t).$$

Example 54. For $f : \mathbb{R}_+ \times S \to \mathbb{R}$,

$$E\Big[\sum_{n=1}^{\infty} f(T_n, X_n)\Big] = \int_{\mathbb{R}_+} E[q_{X(t)} f(t, X(t))]\, dt,$$

provided integral is finite with $|f|$ in place of f.

Remark 55. The results above also apply with obvious modifications to the process $\{X(t) : t \in \mathbb{R}\}$ defined on the entire time axis $\mathbb{R}$, which is natural when considering stationary processes. In particular, the transition times would be labeled $\cdots < T_{-2} < T_{-1} < T_0 \le 0 < T_1 < T_2 \cdots$, and $\sum_{n=1}^{\infty}$ and $\int_{\mathbb{R}_+}$ would be replaced by $\sum_{n\in\mathbb{Z}}$ and $\int_{\mathbb{R}}$.

4.11 Reversibility

The focus in this and the next few sections will be on describing reversible CTMCs. Recall that Chapter 1 introduced the notion of reversibility for discrete-time Markov chains. We now describe an analogous reversibility for CTMCs The results includes a canonical formula for the stationary distribution of reversible processes and a characterization of a CTMC in reverse time. The next section describes several tools for formulating reversible CTMCs and further examples.

As usual, $X(t)$ will denote a CTMC with embedded process (X_n, ξ_n) and the standard notation. We begin with terminology on reversibility.

Definition 56. A CTMC $X(t)$ (or its transition rate matrix q_{ij}) is *reversible* with respect to a measure γ on S if γ satisfies the *detailed balance* equations

$$\gamma_i q_{ij} = \gamma_j q_{ji} \qquad i \neq j \in S. \tag{4.36}$$

Our first observation is that if $X(t)$ is reversible with respect to γ, then γ is an invariant measure for the chain. This follows since summing the detailed balance equations on j yields the total balance equations $\gamma Q = 0$ or

$$\gamma_i \sum_{j\neq i} q_{ij} = \sum_{j\neq i} \gamma_j q_{ji}.$$

Of course, when γ is finite, it can be normalized to be the stationary distribution of the process.

Another observation is that the reversibility of $X(t)$ is equivalent to the reversibility of its embedded chain X_n. In other words, $X(t)$ is reversible with respect to γ if and only if X_n is reversible with respect to $\pi_i = \gamma_i q_i$, $i \in S$. This follows by the definition of reversibility and $q_{ij} = q_i p_{ij}$.

The definition of reversibility for CTMCs is essentially the same as that for discrete-time Markov chains in Chapter 1; the only difference is that

transition rates are used instead of transition probabilities. However, reversibility in continuous time has more interesting applications as we will see.

The discrete-time results in Chapter 1 also apply (with transition rates in place of transition probabilities) to CTMCs, and so we can exploit these results here. For instance, if $X(t)$ is reversible, then it has the *two-way communication property*: for each $i \neq j$ in S, the transition rates q_{ij} and q_{ji} are both positive or both equal to 0. Also, $X(t)$ is reversible with respect to γ if and only if

$$\sum_{i\in A, j\neq i\in B} \gamma_i q_{ij} = \sum_{j\in B, i\neq j\in A,} \gamma_j q_{ji}, \qquad A, B \subset S.$$

This says that the rate of transitions between any two sets is equal to the rate of the reverse transitions, when the chain is ergodic.

The important characterization of invariant measures in Theorem 95 in Chapter 1 has the following continuous-time analogue.

Theorem 57. *If a CTMC $X(t)$ is reversible, then an invariant measure for it is $\gamma_{i_0} = 1$ and*

$$\gamma_i = \prod_{k=1}^{\ell} \frac{q_{i_{k-1},i_k}}{q_{i_k,i_{k-1}}}, \qquad i \in S\backslash\{i_0\},$$

where i_0 is a fixed state and $i_0, i_1, \ldots, i_\ell = i$ is any path from i_0 to i.

Here is the companion result.

Kolmogorov Criterion. A CTMC $X(t)$ with the two-way communication property is reversible if and only if, for each $n \geq 3$ and $i_1, \ldots, i_n$ in S with $i_n = i_1$, and $i_k \neq i_{k+1}$, $1 \leq k \leq n-1$,

$$\prod_{k=1}^{n-1} q_{i_k,i_{k+1}} = \prod_{i=1}^{n-1} q_{i_{k+1},i_k}.$$

Quintessential examples of reversible processes are birth-death processes, including $M/M/s$ queueing processes. These are special cases of the following general models.

Example 58. Up-Down Process. Let $X(t) = (X_1(t), \ldots, X_m(t))$ be a CTMC on $S = \mathbb{Z}_+^m$, where a typical state is denoted by $x = (x_1, \ldots, x_m)$. The CTMC might represent quantities of jobs at m processing stations, stock levels of m products in a warehouse, outstanding orders at m part suppliers, or simply locations of items moving in the plane. Consider the general case in which the only movements may be up or down according to the transition rates

$$q_{xy} = \frac{u(y)}{u(x)}\mathbf{1}(x < y) + \frac{v(x)}{v(y)}\mathbf{1}(x > y),$$

where $u(\cdot)$ and $v(\cdot)$ are positive functions on S.

Viewing $\tilde{u}(x) = -\log u(x)$ as a *potential* function, $u(y)/u(x) = e^{-(\tilde{u}(y)-\tilde{u}(x))}$ reflects the change in potential by component "increases" from x to y. Similarly $v(x)/v(y)$ reflects a "decrease" in potential. An easy check shows that $X(t)$ is reversible with respect to $\gamma_x = u(x)/v(x)$. This model has an obvious extension to a partially-ordered state space S (e.g., one can model excursions on partially-ordered graphs).

Example 59. Batch Birth-Death Processes. A special case of the preceding process is a CTMC $X(t) = (X_1(t), \ldots, X_m(t))$ that describes the numbers of items in m populations, where $X_i(t)$ is the number of items in the ith population. The population state x may increase to $x+a$ or decrease to $x-a$, where a is a positive m-vector in a fixed subset A of S of allowable increments. For simplicity, assume the set A contains the unit vectors $e_1, \ldots, e_m$, where e_i is the vector with 1 in position i and 0 elsewhere.

Assume that $\lambda_i(x_i)$ is a single-unit birth rate in population i when x_i are present, and that the birth rate of a batch of size a_i is

$$\lambda_i(x_i)\lambda_i(x_i+1)\cdots\lambda_i(x_i+a_i-1).$$

This is like a compound transition $x_i \to x_i+1 \to \cdots \to x_i+a_i$ occurring instantaneously under the single birth rate function $\lambda_i(\cdot)$. Single death rates $\mu_i(x_i)$ and batch death rates are defined similarly. In other words, the transition rates of the CTMC are

$$q_{xy} = \sum_{a\in A}\left[\prod_{i=1}^{m}\prod_{k=1}^{a_i}\lambda_i(x_i+k-1)1(y=x+a) + \prod_{i=1}^{m}\prod_{k=1}^{a_i}\mu_i(x_i-k+1)1(y=x-a)\right].$$

Note that these transition rates are the same as those in the preceding example with

$$u(x) = \prod_{i=1}^{m}\prod_{k=1}^{x_i}\lambda_i(k-1), \qquad v(x) = \prod_{i=1}^{m}\prod_{k=1}^{x_i}\mu_i(k).$$

Therefore, an invariant measure for this process is

$$\gamma_x = \prod_{i=1}^{m}\prod_{k=1}^{x_i}\lambda_i(k-1)/\mu_i(k).$$

Interestingly, because of the multiplicative nature of the batch rates, this measure does not depend on the form of the set A of batch sizes as long as it contains the m unit vectors. In particular, any process has the same invariant

measure as the birth-death process with "single" births and deaths in which $A = \{e_1, \ldots, e_m\}$.

Example 60. Classical Birth-Death Processes and Queues. From the preceding example, it follows that a classical birth-death process $X(t)$ as in Example 10 with birth and death rates λ_i and μ_i, respectively, is reversible with respect to the measure $\gamma_0 = 1$ and

$$\gamma_i = \frac{\lambda_0 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}, \qquad i \geq 1.$$

Then the process is ergodic if and only if the sum of these γ_i is finite. In that case, its stationary distribution is $p_i = c\gamma_i$, $i \geq 0$, where $c^{-1} = \sum_{i=0}^{\infty} \gamma_i$.

In particular, the classical $M/M/s$ queueing process in Example 11 with Poisson arrival rate λ and service rate μ is reversible. The process with $s = \infty$ servers is automatically ergodic and the process with $1 \leq s < \infty$ servers is ergodic if and only if $\lambda < s\mu$. Their stationary distributions are as follows.

$M/M/1$ system with $\lambda < \mu$: $\qquad p_i = (1 - \lambda/\mu)(\lambda/\mu)^i, \qquad i \geq 0.$

$M/M/s$ system with $\lambda < s\mu$ and $s < \infty$:

$$p_i = \begin{cases} c(\lambda/\mu)^i/i! & 0 \leq i \leq s \\ p_s(\lambda/s\mu)^{i-s} & i > s. \end{cases}$$

$M/M/\infty$ system: $\qquad p_i = e^{-\lambda/\mu}(\lambda/\mu)^i/i! \qquad i \geq 0.$

We end this section by showing how the reversibility of the CTMC $X(t)$ via detailed balance equations is related to $X(t)$ viewed backward in time. For a fixed $\tau > 0$, consider the process

$$X^\tau(t) = X(\tau - t), \quad 0 \leq t \leq \tau.$$

Each sample path of this process corresponds to a sample path of $X(t)$ in reverse time starting at τ (like viewing a video tape in reverse). The process $\{X^\tau(t) : 0 \leq t \leq \tau\}$ is called the *time reversal of* $X(t)$ *at* τ.[6]

Lemma 61. *The process $X^\tau(t)$ is a non-time-homogeneous CTMC with time-dependent transition probabilities, for $0 \leq s \leq t \leq \tau$,*

$$P\{X^\tau(t) = j | X^\tau(s) = i\} = \frac{P\{X(\tau - t) = j\}}{P\{X(\tau - s) = i\}} p_{ji}(t - s). \tag{4.37}$$

[6] Note that $X^\tau(t)$ has "left-continuous" paths instead of the conventional right-continuous paths. The right-continuous version $X((\tau - t)-)$ is a little cumbersome.

If in addition $X(t)$ is stationary with distribution p, then $X^\tau(t)$ is a time-homogeneous CTMC that is stationary, and its transition probabilities (4.37) reduce to

$$p_i^{-1} p_j p_{ji}(t), \qquad t \in [0, \tau]. \tag{4.38}$$

Hence its transition rates are $p_i^{-1} p_j q_{ji}$.

Proof. Consider the probability

$$P\{X^\tau(t) = j | X^\tau(s) = i, A\} = \frac{P\{X(\tau - t) = j, X(\tau - s) = i, A\}}{P\{X(\tau - s) = i, A\}},$$

for any event A that is a function of $\{X^\tau(u) : 0 \le u < s\}$, and $0 < s \le t < \tau$. To prove the first assertion, it suffices to show that this fraction equals the right side of (4.37). But this equality follows since the denominator equals

$$P\{X(\tau - s) = i\} P(A | X(\tau - s) = i)$$

and the numerator equals

$$P\{X(\tau - t) = j\} P\{X(\tau - s) = i | X(\tau - t) = j\} P(A | X(\tau - s) = i),$$

because A is a function of $\{X(u) : \tau < u < \tau + s\}$ and $X(t)$ is Markovian.

When $X(t)$ is stationary with distribution p, then clearly the transition probabilities (4.37) (for $s = 0$) reduce to (4.38), and the associated transition rates, by Theorem 18, are $p_i^{-1} p_j q_{ji}$. Therefore, $X(t)$ is a time-homogeneous CTMC. In addition, $X^\tau(t) = X(\tau - t)$ has the distribution p for each $t \in [0, \tau]$, and hence $X^\tau(t)$ is stationary with distribution p by Proposition 33.

The notion of a time-reversal process on a finite time interval has a natural extension to the entire time axis.

Definition 62. The *time-reversal* of an ergodic CTMC $X(t)$, with transition rates q_{ij} and stationary distribution p, is a CTMC $\bar{X}(t)$ with initial distribution p and transition rates

$$\bar{q}_{ij} = p_i^{-1} p_j q_{ji}.$$

The finite-dimensional distributions of $\bar{X}(t)$ are the reverse-time of those for $X(t)$ as depicted by property (4.39) below.

Proposition 63. *If the CTMC $X(t)$ is stationary with distribution p, then its time-reversal $\bar{X}(t)$ is stationary with distribution p and*

$$(\bar{X}(t_1), \ldots, \bar{X}(t_n)) \stackrel{d}{=} (X(\tau - t_1), \ldots, X(\tau - t_n)), \quad t_1 < \ldots < t_n \le \tau. \tag{4.39}$$

Proof. By Remark 19, two CTMCs are equal in distribution if and only if their initial distributions and transition rates are equal. Then by the description of $X^\tau(t)$ in Lemma 61 when $X(t)$ is stationary, it follows that $\bar{X}(t)$ and

$X^\tau(t)$ on $[0,\tau]$ are equal in distribution. This proves (4.39). Also, $\bar{X}(t)$ being stationary on $[0,\tau]$, for each τ, implies that it is stationary on $[0,\infty)$.

The CTMC $X(t)$ is defined to be *reversible in time* if it is equal in distribution to its time-reversal process $\bar{X}(t)$; that is (in light of (4.39)),

$$(X(t_1),\ldots,X(t_n)) \stackrel{d}{=} (X(\tau-t_1),\ldots,X(\tau-t_n)), \quad t_1<\ldots<t_n\le\tau. \tag{4.40}$$

How is this time-reversibility related to the (algebraic) reversibility we have been discussing?

Proposition 64. *The CTMC $X(t)$ is reversible in time if and only if it is stationary and reversible.*

Proof. We will use the property that two CTMCs are equal in distribution if and only if their initial distributions and transition rates are equal. Suppose $X(t)$ is reversible in time. Then the transition rates $\bar{q}_{ij}=p_i^{-1}p_jq_{ji}$ of $\bar{X}(t)$ are equal to the rates q_{ij} of $X(t)$, which implies $p_iq_{ij}=p_jq_{ji}$. Thus, $X(t)$ is reversible. In addition, (4.40) implies $X(0)\stackrel{d}{=}X(\tau)$ for each τ, and so $X(t)$ is stationary by Proposition 33.

Conversely, suppose $X(t)$ is stationary and reversible and its distribution is p. Then by Proposition 63, $\bar{X}(t)$ is a stationary CTMC with distribution p. Also, the reversibility supposition implies

$$\bar{q}_{ij}=p_i^{-1}p_jq_{ji}=q_{ij}.$$

Thus, $\bar{X}(t)$ and $X(t)$ are equal in distribution, and hence $X(t)$ is reversible in time.

Remark 65. For a CTMC $\{X(t):t\in\mathbb{R}\}$ on the entire time axis $\mathbb{R}$, the time-reversibility definition (4.40) is equivalent to

$$(X(t_1),\ldots,X(t_n)) \stackrel{d}{=} (X(t_n),\ldots,X(t_1)), \quad t_1<\ldots<t_n \text{ in } \mathbb{R}.$$

4.12 Modeling of Reversible Phenomena

There are a surprising number of complex systems that can be modeled by functions of reversible processes or by reversible transition rates. This section illustrates this with a few key principles of reversibility that help one identify or construct reversible processes.

For this discussion, $X(t)$ will denote a CTMC on S. Since many interesting models are in multi-dimensional spaces, we adopt slightly different notation and let $q(x,y)$, for $x,y\in S$, denote its transition rates, and let $p(x,y)$, $q(x)$ and $\gamma(x)$ denote the usual transition probabilities, sojourn rates and invariant

measure. For simplicity, assume that all the transition rates in this section are irreducible and recurrent, and hence they have positive invariant measures.

The examples in this section are based on elementary properties of reversibility that follow directly from the definition. We begin with two properties that are often applied together.

Proposition 66. (State Space Truncation) *Let $q(x, y)$ be a transition rate function on $\tilde{S}$ that is reversible with respect to γ. If $X(t)$ is a CTMC on a subset $S \subset \tilde{S}$ with transition rates $q(x, y)$ for $x, y \in S$, then $X(t)$ is reversible with respect to γ restricted to S.*

In other words, a reversible process restricted to a subset of its state space is also reversible. The next observation is that a vector-valued CTMC of independent reversible components is reversible. The first assertion uses the fact that two independent CTMCs cannot have a common jump time; this is due in part to Remark 20 that says the probability is 0 that a CTMC has a jump at a fixed time.

Proposition 67. (Juxtaposition of Processes) *If $X_1(t), \ldots, X_m(t)$ are independent CTMCs on spaces $S_1, \ldots, S_m$, then $X(t) = (X_1(t), \ldots, X_m(t))$ on $S = S_1 \times \cdots \times S_m$ is a CTMC with transition rates*

$$q(x, y) = q_i(x_i, y_i) \qquad \text{if } x_j = y_j,\ j \neq i, \text{ for some } i.$$

where $x = (x_1, \ldots, x_m)$ and $q_i(x_i, y_i)$ are the transition rates for $X_i(t)$. If in addition each $X_i(t)$ is reversible with respect to γ_i, then $X(t)$ is reversible with respect to $\gamma(x) = \gamma_1(x_1) \cdots \gamma_m(x_m)$.

Reversible processes arise in many contexts where several independent reversible processes are linked together by certain interactions or constraints. This is exemplified by the following application of the preceding results.

Example 68. Birth-Death and Queueing Processes: Dynamic Population Constraints. Consider a birth-death process (e.g., an $M/M/s$ queueing process) with invariant measure $\gamma(x)$ as in Example 60. We will discuss two variations of this process in which the quantity of items allowed in the system is bounded. These apply naturally to $M/M/s$ queues in which the waiting space for items is limited.

First, assume the allowable number of items in the population is a fixed constant m, so that there are no births when the system is full (e.g., arrivals to a queue are turned away). Let $X(t)$ denote the quantity of items in the population at time t. By Proposition 66, it follows that $X(t)$ is a reversible CTMC on $S = \{0, 1, \ldots, m\}$ with respect to the invariant measure γ of the unconstrained process restricted to S.

In particular, if $m = s$, there is no queueing, and the limiting probability of a full system (the *Erlang Loss Probability*) is

$$p_s = \frac{(\lambda/\mu)^s/s!}{\sum_{i=0}^{s}(\lambda/\mu)^i/i!}.$$

Next, consider the more complicated situation in which the number of items in the system at time t, denoted by $X(t)$, cannot exceed a value $Y(t)$. Assume that $Y(t)$ operates as a reversible Markov process on a subset of $\mathbb{Z}_+$ with transition rates $q_Y(y, y')$ and stationary distribution $p_Y(y)$, but it is constrained by the inequality $X(t) \le Y(t)$. That is, whenever $X(t) = Y(t)$, there are no births; also, transitions of $Y(t)$ below $X(t)$ are not allowed. More precisely, assume that $(X(t), Y(t))$ is an irreducible CTMC on the space $S = \{(x, y) \in \mathbb{Z}_+^2 : x \le y\}$, and its transition rates are

$$q((x,y),(x',y')) = q_X(x,x')\mathbf{1}(y' = y, x' \le y) + q_Y(y,y')\mathbf{1}(x' = x, y' \ge x). \tag{4.41}$$

Here $q_X(x, x')$ is the transition rate function for the unrestricted birth-death process on the nonnegative integers.

Since $q_X(x, x')$ and $q_Y(y, y')$ are reversible, the transition rates (4.41) without the inequality constraints $x' \le y$ are as in Proposition 67. Consequently, (4.41) without the inequality constraints is reversible with respect to the product measure $\gamma(x)p_Y(y)$. Thus, by Proposition 66, the $(X(t), Y(t))$ is reversible with respect to $p(x, y) = \gamma(x)p_Y(y)$.

The next observation is that a transition rate function is reversible if it is a multiplication or compounding of reversible transition functions. This also follows directly from the definition of reversibility.

Proposition 69. (Compound Transition Rates) *Suppose the transition rates of a CTMC $X(t)$ are of the form*

$$q(x, y) = q_1(x, y) \cdots q_m(x, y), \qquad x \ne y \in S,$$

where $q_i(x, y)$ is a transition rate function on S, for $1 \le i \le m$. If each $q_i(x, y)$ is reversible with invariant measure γ_i, then $X(t)$ is invariant with respect to $\gamma(x) = \gamma_1(x_1) \cdots \gamma_m(x_m)$.

Using this property, the truncations of states in Proposition 66 has the following extension to modifications as well as truncations of transitions.

Example 70. Transition Modifications. Consider a CTMC on $\tilde{S}$ whose transition rates $\tilde{q}(x, y)$ are modified by multiplying it by a transition rate function $r(x, y)$ on a subset $S \subseteq \tilde{S}$. The resulting process $X(t)$ is a CTMC on S with transition rates

$$q(x, y) = r(x, y)\tilde{q}(x, y), \qquad x, y \in S.$$

Suppose $\tilde{q}(x, y)$ is reversible on $\tilde{S}$ with respect to a positive $\tilde{\gamma}$. Then $X(t)$ is reversible if and only if $r(x, y)$ is reversible. In this case, an invariant measure for $X(t)$ is $\gamma(x) = \rho(x)\tilde{\gamma}(x)$, where ρ is an invariant measure for $r(x, y)$. This follows from Proposition 69 and the definition of reversibility. For instance, the function $r(x, y)$ might simply be symmetric ($r(x, y) = r(y, x)$, $x, y \in S$).

There are many rate-modification functions $r(x, y)$ on a space $\tilde{S} \subseteq \mathbb{R}_+^m$, such as

$$r(x, y) = \mathbf{1}(x \leq y) + \mathbf{1}(x \geq y) \quad \text{(transitions are either up or down)},$$
$$r(x, y) = \mathbf{1}(\max_j\{|y_j - x_j|\} \leq b) \quad \text{(component differences are} \leq b).$$

Here are three more illustrations.

Example 71. Jumps Affected by a Random Environment. Suppose the transition rate function $\tilde{q}(x, y)$ represents a CTMC on $\tilde{S}$ that is subject to a random environment that affects its jumps as follows. Whenever the process is in state x, a jump to a new state y is allowed with probability $r(x, y)$ and is not allowed with positive probability $1 - r(x, y)$. This jump modification is independent of everything else. The resulting system process $X(t)$ is a CTMC as in Example 70.

Example 72. Resource Constraints. Consider Example 70, where $\tilde{q}(x, y)$ represents a system on $\tilde{S} \subseteq \mathbb{R}_+^m$ that requires certain resources to sustain it. In particular, assume that whenever it is in a state $x = (x_1, \ldots, x_m)$, it requires a quantity $a_{ij}x_j$ of a resource $i \in I$ for each component j, and b_i is the maximum of the resource i that is available. Then setting $r(x, y) = \mathbf{1}(\sum_j a_{ij}y_j \leq b_i, i \in I)$ constrains the system to new states y that do not exceed the resources.

Example 73. Communication Network with Capacity Constraints and Blocking. Consider a communication network, as introduced in [21], that services m types of items. The items arrive to the network according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_m$. For its communication across the network, each type j unit requires the simultaneous use of a_{ij} channels on link i for each i in the set I of links of the network. Some of the a_{ij} may be 0. If these quantities of channels are available, they are assigned to the item, and the item holds the channels for a time that is exponentially distributed with rate μ_j. At the end of this time, the item releases the channels and exits the network. The total number of channels available on link $i \in I$ is b_i. If an item arrives and its required channel quantities are not available, then it cannot enter the network (it is blocked from entering or lost).

Let $X(t) = (X_1(t), \ldots, X_m(t))$ denote the numbers of the m types of items in the network at time t. When $X(t)$ is in state $x = (x_1, \ldots, x_m)$, the number of channels in use on link i is $\sum_j a_{ij}x_j$. Then the state space of $X(t)$ is

$$S = \{x : 0 \leq \sum_j a_{ij}x_j \leq b_i, i \in I\}.$$

Note that if the state of the process is x, then a type j item cannot enter the network when

$$x \in B_j = \{x : \sum_k a_{ik} x_k > b_i - a_{ij}, \text{ for some } i \in I\}.$$

Without these constraints, one can view the $X_i(t)$ as independent birth-death processes (Example 59) with respective birth and death rates λ_i and μ_i, for $1 \leq i \leq m$. Under these constraints, by Propositions 66 and 67, $X(t)$ is a multivariate birth–death process with single-unit movements that is constrained to be in S, and its stationary distribution is

$$p(x) = c \prod_{j=1}^{m} (\lambda_{j-1}/\mu_j)^{x_j}, \quad x \in S,$$

where c is the normalization constant.

Many network performance parameters can be expressed in terms of this distribution. Of prime importance is the probability that an item is blocked from entering the network. The probability that a type j arrival is blocked in equilibrium is $p(B_j) = \sum_{x \in B_j} p(x)$. Ideally, the channel capacities b_i would be sized such that this blocking probability would be less than some small amount such as .01.

One may also be interested in which links cause the blocking. For instance, the probability that a type j item is blocked because the load on link i is full is $\sum_{x \in B_j} p(x) \mathbf{1}(\sum_k a_{ik} x_k > b_i)$.

What is the average number of type j items blocked per unit time? To determine this, consider $\tau_j(t) = \int_0^t \mathbf{1}(X(s) \in B_j)\, ds$, which is the amount of time in $[0, t]$ that type j items are blocked. Now, the number of type j items blocked in $[0, t]$ can be expressed as $N_j(\tau_j(t))$, where $N_j(t)$ is a Poisson process with rate λ_j that is independent of $X(t)$. Thus, by the strong law of large numbers for $N_j(t)$ and for $\tau_j(t)$, the number of type j items blocked per unit time is

$$\lim_{t \to \infty} t^{-1} N_j(\tau_j(t)) = \lim_{t \to \infty} \tau_j(t)^{-1} N_j(\tau_j(t)) \tau_j(t)/t = \lambda_j p(B_j) \quad \text{a.s..}$$

A process for assessing loads on the network is $Y(t) = (Y^i(t) : i \in I)$, where $Y^i(t) = \sum_j a_{ij} X^j(t)$ is the number of channels on link i that are in use at time t. Although this process $Y(t)$ is not Markovian, its stationary distribution, as a function of p, is

$$p_Y(y) = \sum_{x \in S} p(x) 1(\sum_j a_{ij} x_j = y_i, i \in I).$$

This distribution can be used to determine various performance parameters such as the percent of time that link i is idle or the stationary probability that link i has more channels in use than link k. Another parameter of interest is the average number of channels in use on link i, which is $\sum_j a_{ij} \sum_{x \in S} x_j p(x)$.

4.13 Jackson Network Processes

In Chapter 1, we introduced Jackson network processes in discrete time as examples of Markov chains. These models are appropriate when arrivals to the network and service completions occur exactly at discrete times. On the other hand, in many communications and industrial processing networks, arrivals typically occur at times that form Poisson processes or point processes in continuous time, such as those that can be approximated by Poisson processes as in Chapter 3. In addition, the processing times may be more realistic as continuous random variables. In these cases, it is appropriate to use continuous-time models. An important family of such models are Jackson network processes that are CTMCs. This section describes the equilibrium behavior of these Jackson network processes, and the next section describes a related family of network processes.

The network terminology here will be somewhat like that in Section 1.15. Consider an m-node network in which discrete items move among the nodes. The state of the network at time t will be represented by a CTMC $X(t) = (X_1(t), \ldots, X_m(t))$, where $X_i(t)$ denotes the number of items at node i. The state space S is a set of vectors $x = (x_1, \ldots, x_m)$ with nonnegative integer entries, and $q(x, y)$ denotes the transition rates. We also use $|x| = \sum_{i=1}^{m} x_i$. Typical nodes will be labeled $i, j, k, \ldots$

The network may be any one of the following types.

- Closed network with ν items and $S = \{x : |x| = \nu\}$.
- Open network with unlimited capacity and $S = \{x : |x| < \infty\}$.
- Open network with finite capacity ν and $S = \{x : |x| \leq \nu\}$.

Think of the items as moving in the *node set*

$$\mathbb{M} = \begin{cases} \{1, \ldots, m\} & \text{if the network is closed} \\ \{0, 1, \ldots, m\} & \text{if the network is open.} \end{cases}$$

A typical transition is triggered by the movement of a single item. When the network is in state x and one item moves from some i to j in $\mathbb{M}$, the network has a transition

$$x \to T_{ij}x = x - e_i + e_j.$$

Here e_i is the m-dimensional vector with 1 in position i and 0 elsewhere, and e_0 is the zero vector. Most of the following development applies to any of the preceding types of networks; we designate the network type when the distinction is needed.

The movement of items and the resulting transition rates of $X(t)$ are conveniently described by clock times as in Example 9. Specifically, assume that whenever the network is in state x, the time to the next movement of a single item from node i to node j, resulting in a transition $x \to T_{ij}x$, is exponentially distributed with rate $\lambda_{ij}\phi_i(x_i)$, where $\phi_0(\cdot) = 1$. The λ_{ij} are

nonnegative with each $\lambda_{ii} = 0$ (this is for convenience — Exercise 39 shows how it can be relaxed). The $\phi_i(x_i)$ is positive except that $\phi_i(x_i) = 0$ if $x_i = 0$ and $i \neq 0$.

Under this assumption of independent exponential times to item movements, the network process $X(t)$ is a CTMC with transition rates

$$q(x,y) = \begin{cases} \lambda_{ij}\phi_i(x_i) & \text{if } y = T_{ij}x \text{ for some } j \neq i \text{ in } \mathbb{M} \\ 0 & \text{otherwise.} \end{cases}$$

We call $X(t)$ a *Jackson network process*.

Before describing its behavior, a few comments are in order. The exponential sojourn time of $X(t)$ in state x has the rate

$$q(x) = \sum_i \phi_i(x_i) \sum_j \lambda_{ij},$$

and the one-step transition probability is $p(x,y) = q(x,y)/q(x)$. Whenever the process is in state x and an item moves out of node i, the probability that the item moves to node j is

$$p_{ij} = p(x, T_{ij}x) / \sum_k p(x, T_{ik}x) = \lambda_{ij} / \sum_k \lambda_{ik}, \qquad i,j \in \mathbb{M}.$$

Since this probability is independent of x, one can view the items as being independently routed at each transition via the Markov chain *routing probabilities* p_{ij}.

We follow the standard convention that λ_{ij} may either be a routing probability (with $\sum_{k\in\mathbb{M}} \lambda_{ik} = 1$) or a nonnegative intensity of selecting the nodes i and j, and call it the i-to-j *routing rate*. For an open, unlimited-capacity network, the routing assumption implies that items enter the network at the nodes according to independent Poisson processes with respective rates λ_{0j} (this is proved in Example 95); and $\lambda_{0j} = 0$ means that node j does not have arrivals from outside. Think of λ_{ij} as the transition rates of a CTMC that determine the movement of a single item in the node set $\mathbb{M}$ (which includes 0 when the network is open), and p_{ij} are its one-step transition probabilities.

The $\phi_i(x_i)$ is the *service rate* or *departure intensity* at node i when the network state is x. Since this rate does not depend on x_k for $k \neq i$, the nodes are said to "operate" independently. For instance, if node i operates like an $M/M/s$ system with s independent servers that serve at the rate μ_i, then $\phi_i(x_i) = \mu_i \min\{x_i, s\}\mathbf{1}(x_i \geq 1)$.

Another interpretation is that $\phi_j(x_i)$ represents an *egalitarian processor-sharing* scheme in which at any instant, the time to the next "potential" departure of any particular item at the node is exponentially distributed with rate $\phi_i(x_i)/x_i$, independent of the other items. That is, node i works on the each item with this rate, and all the items receive service simultaneously.

These service and routing rates may have other interpretations for general movements of items or particles.

In a transition $x \to T_{ij}x$, we refer to a "single item" moving from i to j. However, more than one item may actually move in the transition, as long as the node populations before and after the transition are x and $T_{ij}x$, respectively. For instance, in a manufacturing network, a part exiting a certain node i may be considered as a completed part that actually exits the network and triggers another item outside the network to take its place and enter node j.

We are now ready to describe the equilibrium behavior of the three types of Jackson processes we have been discussing. With no loss in generality, we assume the routing rates λ_{ij} are irreducible. This is equivalent to $X(t)$ being irreducible; see Exercise 37. Let w_i, $i \in \mathbb{M}$, denote a positive invariant measure that satisfies the *routing balance equations* or *traffic equations*

$$w_i \sum_{j \in \mathbb{M}} \lambda_{ij} = \sum_{j \in \mathbb{M}} w_j \lambda_{ji}, \qquad i \in \mathbb{M}. \tag{4.42}$$

To simplify some expressions, we assume $w_0 = 1$ when the network is open.[7]

Theorem 74. *If $X(t)$ is a closed Jackson process with ν items, then it is ergodic and its stationary distribution is*

$$p(x) = c f_1(x_1) \cdots f_m(x_m), \qquad x \in S = \{x : |x| = \nu\}, \tag{4.43}$$

where $f_i(x_i) = w_i^{x_i} \prod_{k=1}^{x_i} \phi_i(k)^{-1}$ and the w_i satisfy (4.42). The normalization constant is

$$c = \Big(\sum_{x \in S} f_1(x_1) \cdots f_m(x_m) \Big)^{-1}.$$

Theorem 75. *If $X(t)$ is an open Jackson process with finite capacity ν, then the assertions of Theorem 74 apply to this process with $S = \{x : |x| \le \nu\}$.*

Theorem 76. *If $X(t)$ is an open Jackson process with unlimited capacity, then it has an invariant measure of the form (4.43) with $S = \{x : |x| < \infty\}$. Hence, the process is ergodic if and only if*

$$c_i^{-1} = \sum_{k=0}^{\infty} f_i(k) < \infty, \qquad 1 \le i \le m.$$

In this case, its stationary distribution is

$$p(x) = p_1(x_1) \cdots p_m(x_m), \qquad x \in S,$$

where $p_i(x_i) = c_i f_i(x_i)$.

[7] Recall the convention that $\prod_{k=1}^{0}(\cdot) = 1$.

Proof. To prove these theorems, it suffices to show that p given by (4.43) satisfies the balance equations, which in this case are

$$p(x) \sum_{i,j \in \mathbb{M}} q(x, T_{ij}x) = \sum_{i,j \in \mathbb{M}} p(T_{ji}x) q(T_{ji}x, x), \qquad x \in S. \tag{4.44}$$

This proof by substitution is comparable to the proof of Theorem 86 in Chapter 1 for the closed network in discrete time. The substitution is the same for each type of state space. The other details in Theorem 76 are obvious.

Remark 77. Product-Form Distributions. The stationary distributions in the three preceding results have a product form, but only the last one for an open unlimited-capacity network represents a product of probabilities for independent random variables. In this case, if one were to take a snapshot of the node quantities $X_1(t), \ldots, X_m(t)$ at a fixed time t, they would be independent random variables with distributions as in Theorem 76. Of course, these node quantities at different times are not independent.

Remark 78. Partial Balance. The stationary distribution p in the preceding results, which satisfies the total balance equations (4.44), also satisfies the *partial-balance equations*

$$p(x) \sum_{j \in \mathbb{M}} q(x, T_{ij}x) = \sum_{j \in \mathbb{M}} p(T_{ji}x) q(T_{ji}x, x), \qquad i \in S,\ x \in S.$$

Summing these equations on i yields the total balance equations. From the law of large numbers for CTMCs, the partial balance equations say that the average number of items departing from node i per unit time when $X(t)$ is in state x equals the average number of items entering node i per unit time that land $X(t)$ in state x. Or, loosely speaking, the equilibrium flow of items out of node i, for any state x, equals the flow into i. Because of this, the equations are also called *station balance equations.*

Remark 79. Traffic Equations. Although the traffic equations (4.42) precede the theorems and are ostensively an assumption, they are also a consequence of the results. Namely, upon substituting the measure p given by (4.43) in the preceding partial balance equations, the service-rate functions cancel and the traffic equations are what is left. In other words, the traffic equations are a necessary and sufficient condition for p to satisfy the partial balance equations.

The modeling of a network by a Jackson process typically involves specifying the nodes and service actions, verifying the assumption of an exponential time to a transition and identifying the routing and service rates. The next step is to solve the routing equations for the parameters w_i, the only unknown parameters for the stationary distribution. With the stationary distribution in hand, one can then derive various network performance parameters.

The following model of a maintenance network is similar to those in industrial and military settings for maintaining expensive equipment to produce goods or services or to perform a mission.

Example 80. Production–Maintenance Network. Consider a system shown in Figure 4.1 in which ν machines (subsystems, trucks or electronic equipment) are available for use at a facility or location called node 1. At most s_1 machines can be in use at node 1 at any time for producing goods or services. Therefore, if x_1 machines are present then $\min\{x_1, s_1\}$ of these will be in use. After a machine is put into use, it operates continuously until it fails or degrades to a point that it requires a repair. The total operating time is exponentially distributed with rate μ_1. At the end of this time, the machine is transported to a repair facility. The transportation system (which may involve initial processing and rail or air travel) is called node 2, and the machine's time at this node is exponentially distributed with rate μ_2; there is no queueing for the transportation.

The repair facility consists of nodes $3, 4, 5$, which are single-server nodes with respective rates μ_3, μ_4, μ_5. Depending on the nature of the repair, the machine goes to one of these nodes with respective probabilities p_{23}, p_{24}, p_{25}. After its repair, the machine goes to another transportation system, called node 6, for an exponentially distributed time with rate μ_6. And then it enters node 1 to begin another production/repair cycle.

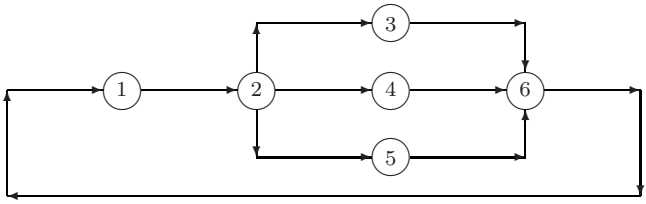

Fig. 4.1 Production–Maintenance Network

Let $X(t)$ denote the process representing the numbers of machines at the six nodes at time t. Under the preceding assumptions, $X(t)$ is a closed Jackson process in which each node i is an s_i-server node, where

$$s_2 = s_6 = \infty, \qquad s_3 = s_4 = s_5 = 1.$$

The service rate of each server at node i is μ_i. The routing intensities are the routing probabilities

$$\lambda_{ij} = 1, \quad (i,j) \in \{(1,2),(3,6),(4,6),(5,6),(6,1)\}, \quad \lambda_{2j} = p_{2j}, \quad j = 3,4,5.$$

the other λ_{ij}'s are 0. A solution of the traffic equations (4.42) for these routing probabilities is $w_i = 1$ for $i = 1, 2, 6$ and $w_i = p_{2i}$ for $i = 3, 4, 5$. Then by Theorem 74, the stationary distribution of $X(t)$ is

$$p(x) = c\frac{1}{x_2!x_6!}\prod_{k=1}^{x_1}\frac{1}{\min\{k, s_1\}}\prod_{i=1}^{6}(w_i/\mu_i)^{x_i}, \quad x \in S,$$

where c is the normalization constant.

The quality of this maintenance system is measured by the number of machines in productive use at node 1. Suppose the aim is to find the number of machines ν^* to provision for the network such that the probability of having less than $\bar{x}_1$ machines in use at node 1 in equilibrium is below β (e.g., .10). From the stationary distribution above, it follows that the equilibrium probability of having less than $\bar{x}_1$ machines at node 1 (as a function of ν) is

$$\begin{aligned}\alpha(\nu) &= \sum_x p(x)1(x_i < \bar{x}_1) \\ &= c\sum_{n=0}^{\bar{x}_1-1}\frac{1}{\mu_1^n n!}\sum_{x_2,\ldots,x_6}\mathbf{1}(\sum_{j=2}^{6}x_j = \nu - n)\frac{1}{x_2!x_6!}\prod_{j=2}^{6}(w_j/\mu_j)^{x_j}.\end{aligned}$$

Then the provisioning quantity $\nu^* = \min\{\nu : h(\nu) \le \beta\}$ is obtained by computing $\alpha(\nu)$ for $\nu = \bar{x}_1, \bar{x}_1 + 1 \ldots$ until it falls below β.

4.14 Multiclass Networks

This section shows how the Jackson network models with homogeneous items extend to networks with multiple types of items, where the routing and services may depend on an item's type. The difference is that we now keep track of the number of items of each type at a node and envision each item as moving within a set of class-node indices. We will describe the equilibrium behavior of two types of multiclass networks.

The first model is for an open network in which each item chooses a particular route through the network, and the item's class label at any instant is designated by the route it is traversing and its stage on the route. Consider an open m-node network that processes items that travel through it as follows. A typical route of an item is a finite sequence $r = (r_1, \ldots, r_\ell)$ of nodes inside the network, where r_s is the node the item visits at *stage* s of its route, $1 \le s \le \ell$; the length of the route $\ell = \ell(r)$ depends on the route. Upon leaving the last node r_ℓ, the item exits the network. A node may appear more than once on a route, and the set of all relevant routes, for simplicity, is finite. Items that traverse a route r arrive to the network according to a Poisson process with rate $\lambda(r)$, and these arrival processes are independent for all the routes. Then the total arrival stream to the network is a Poisson process with rate $\sum_r \lambda(r)$.

The preceding description applies to several scenarios. One is that a deterministic route r is an attribute of an item and all items that traverse

a given route are in the same class. A second scenario is that each item carries a permanent class label that determines its route. A third possibility is that deterministic routes are obtained by random routes as follows. The items arrive to the network by a Poisson process with rate λ, and each item independently selects or is assigned a route r with probability $p(r)$. In this case, $\lambda(r) = p(r)\lambda$. For instance, a route may be selected by Markov probabilities p_{jk} such that $p_{0r_1}p_{r_1r_2}\cdots p_{r_{\ell-1}r_\ell}$ is the probability of the route $r = (r_1, \ldots, r_\ell)$. Combinations of the preceding scenarios yield further possibilities.

To represent the network by a multiclass network process, we assign a class label to each item to denote its routing status at any time in the network. Namely, if an item is traversing route r and is at stage s in this route, we call it a rs-item (which of course resides at node r_s). Let $\mathbb{M}$ denote the set of all route-stage labels rs, including the outside node 0 as well. Then the possible states of the network are given by the state space

$$S = \{x = (x_{rs} : rs \in \mathbb{M}\backslash\{0\})\},$$

where x_{rs} denotes the number of rs-items in the network at node r_s.

Like a Jackson network, assume that whenever the network is in state x, the time to the next departure of an rs-item from its current node location r_s is exponentially distributed with rate $\phi_{rs}(x)$, independent of everything else. The departing item goes immediately to its next node r_{s+1} and becomes an $r(s+1)$-item. In case $s = \ell$, the $r_{\ell+1} = 0$, which means that the route is complete and the item exits the network. It is informative to think of the items as moving in the route-stage set $\mathbb{M}$ as well as among the nodes $0, 1, \ldots, m$.

Let $\{X(t) : t \geq 0\}$ denote the stochastic process representing the state of the network in the space S. Under the preceding assumptions, $X(t)$ is a CTMC with transition rates

$$q(x,y) = \begin{cases} \lambda(r) & \text{if } y = x + e_{r1} \text{ for some } r \\ \phi_{rs}(x) & \text{if } y = x - e_{rs} + e_{r(s+1)} \text{ for some } rs \in \mathbb{M} \\ 0 & \text{otherwise.} \end{cases}$$

Here e_{rs} is the unit vector with 1 in component rs and 0 elsewhere, and $e_0 = 0$. Note that $x_i = \sum_{rs} x_{rs}\mathbf{1}(r_s = i)$ is the number of items at node i

Assume that each node is a processor-sharing node with service rates

$$\phi_{rs}(x) = \frac{x_{rs}}{x_i}\mu_i(x_i),$$

where $i = r_s$. Here $\mu_i(x_i) > 0$ is a load-dependent service rate for the x_i items at node i that is apportioned equally among the x_{rs} rs-items at the node. The CTMC $X(t)$ is a multiclass *Kelly network process* [67] with processor-sharing nodes [67].

Theorem 81. *The Kelly network process has an invariant measure*

$$p(x) = \prod_{rs \in \mathbb{M}} \lambda(r)^{\nu_r} f_{rs}(x), \quad x \in S,$$

where $\nu_r = \sum_{s=1}^{\ell} x_{rs}$ *(the number of items on route* r*) and*

$$f_{rs}(x) = \frac{1}{x_{rs}!} \prod_{i=1}^{m} \frac{x_i!}{\mu_i(1) \cdots \mu_i(x_i)}.$$

Proof. The rates at which the items move in the set $\mathbb{M}$ are

$$\lambda_{0,r1} = \lambda(r), \qquad \lambda_{rs,r(s+1)} = 1, \quad rs \neq 0.$$

The traffic equations for these routing rates are simply $w_0 = 1$ and, for each route r,

$$w_{r1} = \lambda(r), \quad w_{rs} = w_{r(s-1)}, \quad s = 2, \ldots, \ell.$$

A solution to these equations is $w_{rs} = \lambda(r)$ for each $rs \neq 0$. Then like an open Jackson network with unlimited capacity, $X(t)$ is irreducible and one can show by substitution that $p(x)$ satisfies the balance equations and hence it is an invariant measure for the process.

We will now consider a network model in which each item carries an attribute or class label from a finite set. An item's class is a distinguishing characteristic that determines its routing or service rates; the route-stage class label in the preceding model is an example. The class label may be permanent, or temporary and subject to change as the item moves. Examples of permanent labels are:

- The size of an item when it is a batch of subunits such as data packets, orders to be filled, or capacity of a circuit.
- The type of part or tool in a manufacturing network.
- The origin or destination of a item.
- The general direction in which an item moves through the network (e.g., north to south).

Examples of temporary labels are:

- The status of a part as it is being produced.
- The number of nodes an item has visited.
- The number of times an item has been fed back to the node where it resides.
- The phase of service that an item is undergoing, when it has a phase-type distribution.

Our focus will be on a network in which an item's service rate at a node is a compounding of two intensities — one intensity is a function of the total number of items at the node, and the other intensity is a function of the number of items in the same class as the one being served. The other operating rules of the network are similar to those for Jackson networks.

In particular, consider an open m-node network that processes discrete items whose class at any instant is designated by an abstract label α, and $x_{\alpha i}$ denotes the number of α-items at node i. Here αi is in a set $\mathbb{M}$ of class node pairs, including $\alpha 0$ for the outside node 0. Envision the items moving within the set $\mathbb{M}$. The evolution of the network is represented by a CTMC $\{X(t) : t \geq 0\}$ with state space

$$S = \{x = (x_{\alpha i} : \alpha i \in \mathbb{M},\, i \neq 0)\}.$$

The number of items at node i is $x_i = \sum_\alpha x_{\alpha i}$.

Whenever the process is in a state x, a typical transition consists of an α-item departing from node i and moving instantaneously into a node j and entering there as a β-item. We denote this transition by $x \to x - e_{\alpha i} + e_{\beta j}$, where $e_{\alpha i}$ denotes the unit vector with a 1 in component αi and 0 elsewhere, and $e_{\alpha 0} = 0$. It is allowable that $i = j$ or $\alpha = \beta$, provided $\alpha i \neq \beta j$.

Assume the transition rates of the process $X(t)$ are of the form

$$q(x, y) = \begin{cases} \lambda_{\alpha i, \beta j} \phi_{\alpha i}(x) & \text{if } y = x - e_{\alpha i} + e_{\beta j} \text{ for some } \alpha i \neq \beta j \text{ in } \mathbb{M} \\ 0 & \text{otherwise.} \end{cases}$$

The $\phi_{\alpha i}(\cdot)$ are *service rate functions* or intensities, and $\lambda_{\alpha i, \beta j}$ are *routing rates* or intensities. We will consider service rates of the form $\phi_{\alpha 0}(x) = 1$ and

$$\phi_{\alpha i}(x) = g_{\alpha i}(x_i) h_{\alpha i}(x_{\alpha j}), \qquad i \neq 0, \tag{4.45}$$

where $g_{\alpha i}(x_i)$ is the node intensity and $h_{\alpha i}(x_{\alpha i})$ is the class intensity. The routing rates $\lambda_{\alpha j, \beta k}$ may be reducible, but they do not contain transient states.

It is clear that the structure for this multiclass network $X(t)$ is basically the same as that of a Jackson network. Consequently, the theory of Jackson networks also applies to this multiclass analogue —one just replaces the single-node subscripts i with a double subscript αi. In particular, an invariant measure for it is as follows.

Theorem 82. *For the multiclass network process $X(t)$ described above, an invariant measure is*

$$p(x) = \prod_{\alpha i \in \mathbb{M}} w_{\alpha i}^{x_{\alpha i}} f_{\alpha i}(x), \quad x \in S,$$

where $f_{\alpha 0}(x) = 1$,

$$f_{\alpha i}(x) = \prod_{k=1}^{x_i} g_{\alpha i}(k)^{-1} \prod_{k'=1}^{x_{\alpha i}} h_{\alpha i}(k')^{-1}, \qquad i \neq 0,$$

and $w_{\alpha i}$ are positive numbers, with $w_{\alpha 0} = 1$, that satisfy the traffic equations

$$w_{\alpha i} \sum_{\beta j \in \mathbb{M}} \lambda_{\alpha i, \beta j} = \sum_{\beta j \in \mathbb{M}} w_{\beta j} \lambda_{\beta j, \alpha i}, \quad \alpha i \in \mathbb{M}. \tag{4.46}$$

Note that a Kelly network process (with rs replaced by αi) is a special case of this multiclass network. Several other examples are discussed in [101, 111]; here is another one.

Example 83. The multiclass network described above is a *BCMP network*, introduced by Baskett, Chandy, Muntz, and Palacios in [9], if each of its nodes is one of the following four types.

- *First-Come, First-Served* node with service rates $\phi_{\alpha i}(x) = \mu_i(x_i)$. Each item (as in a Jackson network) has an exponential service time with the same load-dependent service rate $\mu_i(x_i)$.
- *Processor-Sharing* node with service rates $\phi_{\alpha i}(x) = x_{\alpha i} x_i^{-1} \mu_{\alpha i}(x_i)$. The $\mu_{\alpha i}(x_{\alpha i})$ is a customer-load-dependent service rate, which is apportioned equally among the $x_{\alpha i}$ α-items at the node.
- *Last-Come, First-Served with Preemption* node with service rates as in the preceding PS case.
- *Infinite-Server* node with service rates $\phi_{\alpha i}(x) = x_{\alpha i} \mu_{\alpha i}(x_{\alpha i})$.

An invariant measure for this BCMP network is $p(x) = \prod_{\alpha i \in \mathbb{M}} w_{\alpha i}^{x_{\alpha i}} f_{\alpha i}(x)$ as in Theorem 82, where $f_{\alpha 0}(x) = 1$ and the other functions are

$$f_{\alpha i}(x) = \frac{1}{\mu_i(1) \cdots \mu_i(x_i)} \qquad \text{FCFS}$$

$$f_{\alpha i}(x) = \frac{1}{x_{\alpha i}!} \frac{x_i!}{\mu_{\alpha i}(1) \cdots \mu_{\alpha i}(x_i)} \qquad \text{PS or LCFSPR}$$

$$f_{\alpha i}(x) = \frac{1}{x_{\alpha i}!} \frac{1}{\mu_{\alpha i}(1) \cdots \mu_{\alpha i}(x_{\alpha i})} \qquad \text{IS.}$$

This completes our discussion of networks.

4.15 Poisson Transition Times

In the next five sections, we will discuss the behavior of ergodic CTMCs at their transition times. This section presents criteria under which the times of a certain type of transition of a CTMC form a Poisson process. For instance, in a stationary birth-death process, the times at which deaths occur form a Poisson process, and in a Jackson network process, the times of departures from certain nodes form independent Poisson processes. Probabilities of a CTMC at these special transition times that form a Poisson process will be the focus of the following four sections.

For this discussion,[8] $X = \{X(t) : t \in \mathbb{R}\}$ will denote an ergodic CTMC with transition rates q_{ij} and stationary distribution p. We consider X on the entire time axis since some of the results here are more natural for stationary processes. Its transition times are depicted by the point process

$$N(B) = \sum_{n\in\mathbb{Z}} \mathbf{1}(T_n \in B), \qquad B \in \mathcal{B},$$

where the transition times T_n are labeled such that

$$\ldots < T_{-2} < T_{-1} < T_0 \le 0 < T_1 < T_2 \ldots.$$

By Example 50, $E[N(B)] = \int_B E[q_{X(t)}]\, dt$. This is finite for bounded B if $\sum_i q_i < \infty$.

The point process N of all the transition times is usually not a Poisson process, although it is for a Markov chain subordinated to a Poisson process. However, an important feature of a CTMC is that certain types of its transition times may be Poisson processes. The results that follow give necessary and sufficient conditions for such subsets of transition times of N to be Poisson processes.

We begin by defining what is a "special" transition of the CTMC X. It is natural to describe these transitions in terms of its sample paths. Let $\mathbb{D}$ denote the set of all functions $x : \mathbb{R} \to S$ that are right-continuous with left-hand limits, and are piece-wise constant with a finite number of jumps in any finite time interval.[9] Recall that a typical sample path of X is a function in $\mathbb{D}$. In other words, the process X is a $\mathbb{D}$-valued random variable (or a random element in $\mathbb{D}$). Accordingly, we assume the σ-field associated with $\mathbb{D}$ is the smallest σ-field $\mathcal{D}$ under which, for each t, the projection map $x \to x(t)$ is measurable: this ensures that each $X(t)$ is a well-defined S-valued random variable.

We will also refer to

$$\mathrm{S}^t X = \{X(t+u) : u \in \mathbb{R}\}, \qquad t \in \mathbb{R},$$

which is the process X with its time parameter shifted[10] by the amount t. The $\mathrm{S}^t X$ is what an observer sees of the path at time t. It is natural to describe a certain "transition" of X at a time t when $\mathrm{S}^t X$ is in a set $\mathcal{T}$ in $\mathcal{D}$.

Definition 84. Suppose $\mathcal{T} \in \mathcal{D}$ is such that

[8] We will now use the shorthand notation X for the CTMC instead of $X(t)$ as in the preceding sections.

[9] In Chapter 5, $D = D[0,1]$ denotes the space of functions on $[0,1]$ that are right-continuous with left-hand limits, without the restrictions that they be piece-wise constant or take only a finite number of jumps in a finite interval.

[10] The S^t is the time-shift operator on $\mathbb{D}$. The shift notation S and the state-space notation S are different.

$$x(0-) \neq x(0), \qquad x \in \mathcal{T}. \tag{4.47}$$

We say that a $\mathcal{T}$*-transition* of X occurs at time t if $\mathrm{S}^t X \in \mathcal{T}$. The times at which these $\mathcal{T}$-transitions occur are depicted by the point process $N_{\mathcal{T}}$ on $\mathbb{R}$ defined by

$$N_{\mathcal{T}}(B) = \sum_{t \in B} \mathbf{1}(\mathrm{S}^t X \in \mathcal{T}) = \sum_{n \in \mathbb{Z}} \mathbf{1}(\mathrm{S}^{T_n} X \in \mathcal{T}, T_n \in B), \qquad B \in \mathcal{B}.$$

These quantities are finite on bounded time sets B, since X can only take a finite number of jumps in such sets.

Note that if a $\mathcal{T}$-transition occurs at time t ($\mathrm{S}^t X \in \mathcal{T}$), then condition (4.47) ensures that $X(t-) \neq X(t)$, which implies that t is a jump time of X.

Example 85. Suppose $X(t)$ denotes the value of a stock at time t. Then the times at which the value increases are $\mathcal{T}$-transitions, where

$$\mathcal{T} = \{x \in \mathbb{D} : x(0) > x(0-)\}.$$

The times between these transitions are the times between increases in the stock value. This type of transition only involves values of X before and after the transition; the past and future are not involved.

An example of a transition time that involves more sample-path information is a time at which the stock value jumps into an interval (a, b) and thereafter the stock value reaches b before it reaches the value a. In this case,

$$\mathcal{T} = \{x \in \mathbb{D} : x(0-) \notin (a, b), x(0) \in (a, b), h_b(x) < h_a(x)\},$$

where $h_a(x) = \inf\{t > 0 : x(t) \leq a\}$ is the time at which the path x hits $(-\infty, a]$ starting from time 0, and $h_b(x)$ is defined similarly. The time between such transitions is the time for the stock value X to reach $[b, \infty)$ given that it does not reach $(-\infty, a]$.

We begin with some elementary observations about the point process $N_{\mathcal{T}}$ of $\mathcal{T}$-transitions. Its mean measure, which follows by the generalized Lévy formula in Theorem 52 and Remark 55, is

$$E[N_{\mathcal{T}}(B)] = \int_B E[\alpha_{\mathcal{T}}(X(t))]\, dt, \qquad B \in \mathcal{B},$$

where

$$\alpha_{\mathcal{T}}(i) = \sum_{j \neq i} q_{ij} P\{X \in \mathcal{T} | X_{-1} = i, X_0 = j\}, \qquad i \in S. \tag{4.48}$$

The quantity $\alpha_{\mathcal{T}}(i)$ is the infinitesimal *rate at which a* $\mathcal{T}$*-transition is initiated* from state i. Clearly, $N_{\mathcal{T}}(B) \leq N(B)$, and so $E[N_{\mathcal{T}}(B)] \leq E[N(B)]$.

The next result establishes that $N_{\mathcal{T}}$ is stationary when X is. However, $N_{\mathcal{T}}$ may be stationary when X is not, which is the case in some examples below.

Proposition 86. *If X is stationary, then the point process $N_{\mathcal{T}}$ of $\mathcal{T}$-transitions of X is stationary with rate*

$$\lambda_{\mathcal{T}} = E[N_{\mathcal{T}}(0,1]] = \sum_i p_i \alpha_{\mathcal{T}}(i),$$

which is finite when $\sum_i p_i q_i < \infty$.

Proof. From Section 2.15, $N_{\mathcal{T}}$ is stationary if, for any $B_1, \ldots, B_m \in \mathcal{B}$ and $n_1, \ldots, n_m$,

$$P\{N_{\mathcal{T}}(B_1 + t) = n_1, \ldots, N_{\mathcal{T}}(B_m + t) = n_m\} \text{ is the same for each } t. \quad (4.49)$$

It suffices for the B_k to be bounded sets, since the distribution of $N_{\mathcal{T}}$ is determined by its finite-dimensional distributions on bounded sets. For a bounded B, using the change-of-variable $v = u - t$,

$$N_{\mathcal{T}}(B + t) = \sum_{u \in B+t} \mathbf{1}(\mathrm{S}^u X \in \mathcal{T}) = \sum_{v \in B} \mathbf{1}(\mathrm{S}^v(\mathrm{S}^t X) \in \mathcal{T}).$$

Then the assumption $\mathrm{S}^t X \stackrel{d}{=} X$ implies (4.49) for $m = 1$. A similar argument justifies (4.49) for any m, and hence $N_{\mathcal{T}}$ is stationary.

Our criteria for Poisson $\mathcal{T}$-transitions involve the following concepts.

Definition 87. A $\mathcal{T}$-transition has a *uniform initiation rate* $\lambda_{\mathcal{T}}$ if

$$\alpha_{\mathcal{T}}(i) = \lambda_{\mathcal{T}}, \qquad \text{for each } i \in S,$$

where $\alpha_{\mathcal{T}}(i) = \sum_{j \neq i} q_{ij} P\{X \in \mathcal{T} | X_{-1} = i, X_0 = j\}$ is the infinitesimal initiation rate of a $\mathcal{T}$-transition from state i.

Definition 88. *The future of $N_{\mathcal{T}}$ is independent of the past of X* if, for each $t \in \mathbb{R}$, the future quantities $\{N_{\mathcal{T}}(B) : B \subset (t, \infty)\}$ are independent of $\{X(s) : s \leq t\}$.

Here is our first criterion for $N_{\mathcal{T}}$ to be Poisson.

Theorem 89. *The following statements are equivalent.*
(a) *$N_{\mathcal{T}}$ is a Poisson process with rate $\lambda_{\mathcal{T}}$ and the future of $N_{\mathcal{T}}$ is independent of the past of X.*
(b) *$\mathcal{T}$ has a uniform initiation rate $\lambda_{\mathcal{T}}$.*

Proof. First, suppose (b) is true. We will prove (a) for the case in which the sojourn rates q_i are bounded. A standard proof in [101] for unbounded rates involves a martingale characterization of Poisson processes, which we do not cover.

By the uniformization principle in Proposition 25, the CTMC X is equal in distribution to a CTMC $\hat{X} = \{\hat{X}(t) : t \in \mathbb{R}\}$ — a subordinated Markov chain — with transition probabilities

$$\hat{p}_{ij} = q_{ij}/\lambda, \qquad \hat{p}_{ii} = 1 - q_i/\lambda,$$

where $\lambda = \sup_i q_i$. The point process $\hat{N}$ of its transition times is a Poisson process with rate λ independent of the embedded states $\hat{X}_n$, $n \in \mathbb{Z}$. Let $\hat{N}_{\mathcal{T}}$ denote the point process of $\mathcal{T}$-transitions for $\hat{X}$.

Since X and $\hat{X}$ are equal in distribution, any deterministic function of one of them is equal in distribution to the function of the other one. In particular, $N_{\mathcal{T}}$ and $\hat{N}_{\mathcal{T}}$ are equal in distribution. Then to prove (a), it suffices to prove (a) for $\hat{N}_{\mathcal{T}}$ and $\hat{X}$.

To this end, note that by the independence of the Markov chain $\hat{X}_n$ and the Poisson process $\hat{N}$, it follows that $\hat{N}_{\mathcal{T}}$ is a thinning of $\hat{N}$, where the probability of a point of $\hat{N}$ being retained is

$$\begin{aligned} P\{\mathrm{S}^{\bar{T}_n}\hat{X} \in \mathcal{T}\} &= \sum_i P\{\hat{X}_{n-1} = i\} \sum_{j \neq i} \hat{p}_{ij} P\{\hat{X} \in \mathcal{T} | X_{-1} = i, X_0 = j\} \\ &= \sum_i P\{\hat{X}_{n-1} = i\} \alpha_{\mathcal{T}}(i)/\lambda = \ \lambda_{\mathcal{T}}/\lambda. \end{aligned}$$

Here we use $\hat{p}_{ij} = q_{ij}/\lambda$ and the uniform initiation rate assumption. This assumption also implies that a $\mathcal{T}$-transition can be triggered from any state. Thus by the Poisson thinning principle (Example 40 in Chapter 3), it follows that $\hat{N}_{\mathcal{T}}$ is a Poisson process with rate $\lambda_{\mathcal{T}}$.

In addition, note that since the Poisson process $\hat{N}$ is independent of the Markov chain $\hat{X}_n$, and $\hat{N}$ has independent increments, it follows that the future of $\hat{N}$ is independent of the past of $\hat{X}$. The preceding arguments prove that $\hat{X}$ satisfies (a), which implies that X satisfies (a).

To prove the converse, suppose (a) is true. Then, for any i and $t \geq 0$,

$$\begin{aligned} \lambda_{\mathcal{T}} t = E[N_{\mathcal{T}}(0,t]] &= E[N_{\mathcal{T}}(0,t] | X(0) = i] \\ &= \int_0^t E[\alpha_{\mathcal{T}}(X(s)) | X(0) = i]\, ds. \end{aligned}$$

The integrand is continuous in s since X is a CTMC. Taking the derivative of this equation with respect to t yields

$$\lambda_{\mathcal{T}} = E[\alpha_{\mathcal{T}}(X(t)) | X(0) = i], \qquad t \geq 0.$$

Then, using the first jump time $T_1 = \inf\{t > 0 : X(t) \neq X(0)\}$, we can write

$$\lambda_{\mathcal{T}} = \alpha_{\mathcal{T}}(i) P\{T_1 > t | X(0) = i\} + E[\alpha_{\mathcal{T}}(X(t)) \mathbf{1}(T_1 \leq t) | X(0) = i].$$

Letting $t \downarrow 0$ proves (b).

Example 90. Births in a Birth-Death Process. Suppose X denotes the population size in an ergodic birth-death process with birth and death rates λ_i and μ_i. Its transition rates are $q_{ij} = \lambda_i \mathbf{1}(j = i+1) + \mu_i \mathbf{1}(j = i-1)$.

Consider the point process $N_{\mathcal{T}}$ of the times at which births occur, where $\mathcal{T} = \{x \in \mathbb{D} : x(0) = x(0-) + 1\}$. The initiation rate of births is clearly

$$\alpha_{\mathcal{T}}(i) = q_{i,i+1} = \lambda_i, \qquad \text{for each } i \in S.$$

Then Theorem 89 yields the result that $\lambda_i = \lambda$, for each i, is a necessary and sufficient condition for the birth process $N_{\mathcal{T}}$ to be Poisson with rate λ and future births are independent of the past of the process X. Note that this conclusion is true for $M/M/s$ queueing processes. This proves that the arrival process is Poisson with rate λ, which is usually a preliminary assumption in applications.

The Poisson criterion for $\mathcal{T}$-transition times in the preceding result applies to non-stationary as well as stationary CTMCs. Our next criterion for Poisson $\mathcal{T}$-transitions is only for stationary CTMCs. This involves considering a CTMC in reverse time and applying Theorem 89.

Recall that when X is stationary with distribution p, its time-reversal, according to Proposition 63, is a stationary CTMC $\bar{X} = \{\bar{X}(t) : t \in \mathbb{R}\}$ with transition rates $\bar{q}_{ij} = p_i^{-1} p_j q_{ji}$. Note that the time reversal of $\mathcal{T}$ is the set $\bar{\mathcal{T}}$ of all paths $\bar{x}(t)$ in $\mathbb{D}$ such that $\bar{x}(t) = x(-t)$ at each continuity point of x, for $x \in \mathcal{T}$. Then the initiation rate of $\bar{\mathcal{T}}$-transitions of $\bar{X}$ is

$$\begin{aligned} \bar{\alpha}_{\bar{\mathcal{T}}}(i) &= \sum_{j \neq i} \bar{q}_{ij} P\{\bar{X} \in \bar{\mathcal{T}} | \bar{X}_{-1} = i, \bar{X}_0 = j\} \qquad (4.50) \\ &= p_i^{-1} \sum_{j \neq i} p_j q_{ji} P\{X \in \mathcal{T} | X_{-1} = j, X_0 = i\}. \end{aligned}$$

We say that the *time-reversal of $\mathcal{T}$ has a uniform initiation rate $\lambda_{\mathcal{T}}$* if

$$\bar{\alpha}_{\bar{\mathcal{T}}}(i) = \lambda_{\mathcal{T}}, \qquad \text{for each } i \in S.$$

Theorem 91. *If X is stationary, the following statements are equivalent.*
(a) *$N_{\mathcal{T}}$ is a Poisson process with rate $\lambda_{\mathcal{T}}$ and the future of X is independent of the past of $N_{\mathcal{T}}$.*
(b) *The time-reversal of $\mathcal{T}$ has a uniform initiation rate $\lambda_{\mathcal{T}}$.*

Proof. Because of the Markov property, a $\mathcal{T}$-transition of X has the same probability as a $\bar{\mathcal{T}}$-transition of $\bar{X}$. Furthermore, $N_{\mathcal{T}}$ is equal in distribution to the point process $\bar{N}_{\bar{\mathcal{T}}}$ of $\bar{\mathcal{T}}$-transitions of $\bar{X}$. Now, by Theorem 89, (b) is equivalent to the statement that $\bar{N}_{\bar{\mathcal{T}}}$ is a Poisson process with rate $\lambda_{\mathcal{T}}$ and the future of $\bar{N}_{\bar{\mathcal{T}}}$ is independent of the past of $\bar{X}$, which is equivalent to (a).

Example 92. Deaths in a Birth-Death Process. Suppose that X denotes the birth-death process in Example 90, but now assume it is stationary. Consider the point process $N_{\mathcal{T}}$ of the times at which deaths occur, which means that $\mathcal{T} = \{x \in \mathbb{D} : x(0) = x(0-) - 1\}$. Note that the rate at which deaths are initiated from state i is

$$\alpha_{\mathcal{T}}(i) = \sum_{j \neq i} q_{ij} \mathbf{1}(j = i - 1) = \mu_i \mathbf{1}(i \geq 1).$$

This is not the same for each i, even if the μ_i are all equal. Therefore Theorem 89 does not apply to establish that $N_{\mathcal{T}}$ is a Poisson process.

To see if Theorem 91 is applicable, we have to determine the initiation rate of the reverse-time of death transitions. We know from Example 60 that the stationary distribution of $X(t)$ is $p_i = c\lambda_0 \cdots \lambda_{i-1}/(\mu_1 \cdots \mu_i)$. Then the initiation rate (4.50) of the reverse-time of death transitions (which are birth transitions for the time reversal of X) is

$$p_i^{-1} \sum_{j \neq i} p_j q_{ji} \mathbf{1}(j = i + 1) = p_i^{-1} p_{i+1} q_{(i+1),i} = \lambda_i, \qquad \text{for each } i \in S.$$

Thus, by Theorem 91, $\lambda_i = \lambda$, for each i, is a necessary and sufficient condition for $N_{\mathcal{T}}$ to be a Poisson process with rate λ and the past departures are independent of future births or deaths.

Of course, if the birth rate is the constant λ, then the death rate should equal this birth rate since X is stationary. However, just because the point process of birth times is Poisson, it is not immediately clear that the process of death-times would be Poisson; but now we know it is.

A special case of this example says that the times of departures from a stationary $M/M/s$ queue is a Poisson process with the same rate as the arrivals.

For multi-dimensional CTMCs such as Jackson network processes, one may be interested in whether several point processes of transition times are Poisson. We now show that the preceding criteria for Poisson $\mathcal{T}$-transition times extend to multiple transition processes. The key idea is to formulate multiple processes as a partition of a Poisson process as in Section 3.10.

Theorem 93. *For disjoint transition sets $\mathcal{T}_1, \ldots, \mathcal{T}_\ell$, the following statements are equivalent.*
(a) *The point processes $N_{\mathcal{T}_1}, \ldots, N_{\mathcal{T}_\ell}$ are independent Poisson processes with respective rates $\lambda_{\mathcal{T}_1}, \ldots, \lambda_{\mathcal{T}_\ell}$, and their futures are independent of the past of X.*
(b) *Each $\mathcal{T}_k$ has a uniform initiation rate $\lambda_{\mathcal{T}_k}$, for $1 \leq k \leq \ell$.*

Proof. Suppose (b) is satisfied. Then the initiation rate of $\mathcal{T} = \cup_{k=1}^{\ell} \mathcal{T}_k$ is

$$\alpha_{\mathcal{T}}(i) = \sum_{k=1}^{\ell} \alpha_{\mathcal{T}_k}(i) = \lambda_{\mathcal{T}}, \qquad i \in S,$$

where $\lambda_{\mathcal{T}} = \sum_{k=1}^{\ell} \lambda_{\mathcal{T}_k}$. Then by Theorem 89, $N_{\mathcal{T}}$ is a Poisson process with rate $\lambda_{\mathcal{T}}$ and the future of $N_{\mathcal{T}}$ is independent of the past of X.

Now, each of the $\mathcal{T}$-transitions of X is independently triggered by a $\mathcal{T}_k$-transition ($1 \le k \le \ell$) (where $\mathcal{T}_k \subseteq \mathcal{T}$) with probability

$$\begin{aligned} &P\{\mathrm{S}^{T_n} X \in \mathcal{T}_k | \mathrm{S}^{T_n} X \in \mathcal{T}\} \\ &\quad = P\{X \in \mathcal{T}_k\}/P\{X \in \mathcal{T}\} = \lambda_{\mathcal{T}_k}/\lambda_{\mathcal{T}}. \end{aligned}$$

That is, the processes $N_{\mathcal{T}_1}, \ldots, N_{\mathcal{T}_\ell}$ form a partition of $N_{\mathcal{T}}$ with the preceding probabilities. Therefore, by Corollary 41 in Chapter 3, these processes are independent Poisson processes with respective rates $\lambda_{\mathcal{T}_1}, \ldots, \lambda_{\mathcal{T}_\ell}$, and, being parts of $N_{\mathcal{T}}$, their futures are independent of the past of X.

If (a) holds, then (b) follows by Theorem 89.

Here is a companion result for multiple Poisson processes for a stationary CTMC; it follows by an obvious extension of the proof of Theorem 91.

Theorem 94. *If the CTMC X is stationary, then for disjoint transition sets $\mathcal{T}_1, \ldots, \mathcal{T}_\ell$, the following statements are equivalent.*
(a) *$N_{\mathcal{T}_1}, \ldots, N_{\mathcal{T}_\ell}$ are independent Poisson processes with respective rates $\lambda_{\mathcal{T}_1}, \ldots, \lambda_{\mathcal{T}_\ell}$ and their pasts are independent of the future of X.*
(b) *The time-reversal $\bar{\mathcal{T}}_k$ of $\mathcal{T}_k$ has a uniform initiation rate $\bar{\lambda}_{\bar{\mathcal{T}}_k}$, $1 \le k \le \ell$.*

Example 95. Jackson Networks. Suppose that X is an ergodic, infinite-capacity open Jackson network process with routing rates λ_{ij} and service rates $\phi_i(x_i)$, $1 \le i \le m$. Let $I \subset \{1, \ldots, m\}$ denote the set of nodes at which arrivals can enter the network from outside. Consider the point process

$$N_{0i}(B) = \sum_{t \in B} \mathbf{1}(X(t) = X(t-) + e_i), \qquad B \in \mathcal{B},\ i \in I,$$

of arrival times of items from outside the network into node i. The processes N_{0i}, $i \in I$, are independent Poisson processes with rates λ_{0i}, $i \in I$; and their futures are independent of the past of X. This follows by Theorem 93 since, for each $i \in I$,

$$\sum_{y \ne x} q(x,y)\mathbf{1}(y = x + e_i) = \lambda_{0i}, \qquad x \in S.$$

Now suppose X is stationary. Consider the point process N_{j0} of departure times (to outside) from node j in the set of nodes J that have departures. Then N_{j0}, $j \in J$, are independent Poisson processes with rates $w_j \lambda_{j0}$, $j \in J$, and their pasts are independent of the future of X. This follows by Theorem 94 since, for each $j \in J$,

$$p(x)^{-1}\sum_{y\neq x} p(y)q(y,x)\mathbf{1}(y=x+e_j) = p(x)^{-1}p(x+e_j)q(x+e_j,x)$$
$$= w_i\lambda_{j0}, \qquad x\in S.$$

In some cases, the point process of times at which items move from one node to another is a Poisson processes, provided that a unit may make this move only once during its stay in the network; this is true of acyclic networks as in Exercise 30.

4.16 Palm Probabilities

This section continues the study of $\mathcal{T}$-transitions by characterizing the past and future of a CTMC at such a transition. The issue is how to evaluate any probability for a CTMC conditioned that a $\mathcal{T}$-transition occurs. Since the occurrence of a $\mathcal{T}$-transition at any time has probability 0, we cannot use conventional conditional probabilities. Instead, we formulate these conditional probabilities as Palm probabilities. Further properties of Palm probabilities are described in the next three sections.

As an example of what lies ahead, consider an $M/M/1$ queueing process in equilibrium. Items arrive according to a Poisson process, and so the probability of an arrival at any time is 0. However, when an item does arrive to the system, it is of interest to know the distribution of the number of items it encounters in the system and the distribution of the arrival's sojourn time in the system. Viewing the arrival times as $\mathcal{T}$-transitions, we will prove, using Palm probabilities, that the distribution of the number of items an arrival encounters is the stationary distribution of the process, and that the arrival's sojourn time has an exponential distribution.

For the following discussion, $X = \{X(t) : t \in \mathbb{R}\}$ will denote an ergodic CTMC with transition rates q_{ij} and stationary distribution p. Unless specified otherwise, we will assume that X is stationary. Then as noted in Proposition 86, the rate at which a $\mathcal{T}$-transition is initiated from state i is

$$\alpha_{\mathcal{T}}(i) = \sum_{j\neq i} q_{ij}P\{X\in\mathcal{T}|X_{-1}=i, X_0=j\},$$

and the point process $N_{\mathcal{T}}$ of $\mathcal{T}$-transitions of X is stationary with rate

$$\lambda_{\mathcal{T}} = E[N_{\mathcal{T}}(0,1]] = \sum_i p_i\alpha_{\mathcal{T}}(i).$$

The aim is to formulate conditional probabilities given that a $\mathcal{T}$-transition occurs. Recall that a $\mathcal{T}$-transition of X occurs at time t if $\mathrm{S}^t X \in \mathcal{T}$. Given that this event occurs, we want to find the conditional distribution of X, or

$N_{\mathcal{T}}$, or other functions of X. These probabilities are not standard conditional probabilities, since the probability that a transition occurs at any time is 0 (recall Remark 20). However, we can formulate these conditional probabilities as Palm probabilities. A Palm probability is defined in more general contexts as a Radon-Nikodým derivative, but for a CTMC, the definition reduces simply to a ratio of rates.

Definition 96. The *Palm probability* of a stationary CTMC X conditioned that a $\mathcal{T}$-transition occurs at any time t is the probability measure $P_{\mathcal{T}}$ defined, for any event A generated[11] by the process X, as

$$P_{\mathcal{T}}(A) = \frac{1}{\lambda_{\mathcal{T}}} \sum_i p_i \sum_{j \neq i} q_{ij} P\{A,\, X \in \mathcal{T} | X_{-1} = i, X_0 = j\}.$$

This probability for a transition at time t is independent of t, and so it is often associated with a transition at time 0.

The meaning of a Palm probability as a limit of standard conditional probabilities, like that of a Radon-Nykodym derivative, is given below in Proposition 97. Before getting into examples, we will comment on the definition and present a few properties of Palm probabilities.

The Palm probability $P_{\mathcal{T}}$ is defined as a function of the underlying probability P, and the two probabilities are different. In particular, while the P-probability of a $\mathcal{T}$-transition is 0, the $P_{\mathcal{T}}$-probability is 1 because

$$P_{\mathcal{T}}\{N_{\mathcal{T}}(\{0\}) = 1\} = P_{\mathcal{T}}\{X \in \mathcal{T}\} = 1.$$

This property is consistent with saying that $P_{\mathcal{T}}$ is a conditional probability "given that a $\mathcal{T}$-transition occurs" at time 0.

As another formulation, note that any event A generated by X is a function of the sample paths of X, and so $P_{\mathcal{T}}(A) = P_{\mathcal{T}}\{X \in \mathcal{T}'_A\}$, where

$$\mathcal{T}'_A = \{x \in \mathcal{T} : A \text{ occurs for the path } x\}.$$

Consequently,

$$P_{\mathcal{T}}(A) = \frac{\lambda_{\mathcal{T}'_A}}{\lambda_{\mathcal{T}}} = \frac{E[N_{\mathcal{T}'_A}(a,b]]}{E[N_{\mathcal{T}}(a,b]]}, \qquad a < b. \tag{4.51}$$

This ratio is the expected number of $\mathcal{T}$-transitions at which a $\mathcal{T}'_A$-event occurs in a fixed time interval divided by the expected number of all $\mathcal{T}$-transitions in the interval (or the portion of $\mathcal{T}$-transitions at which a $\mathcal{T}'_A$-event occurs). The second ratio is due to $E[N_{\mathcal{T}}(a,b]] = (b-a)\lambda_{\mathcal{T}}$ since $N_{\mathcal{T}}$ is stationary.

Another useful expression is, for any $\mathcal{T}' \subset \mathcal{T}$,

[11] The A is in the σ-field of events generated by the random variables $\{X(t) : t \in \mathbb{R}\}$.

$$P_{\mathcal{T}}\{X \in \mathcal{T}'\} = \frac{1}{(b-a)\lambda_{\mathcal{T}}} E\Big[\sum_{n\in\mathbb{Z}} \mathbf{1}(\mathrm{S}^{\tau_n} X \in \mathcal{T}', \tau_n \in (a,b])\Big], \tag{4.52}$$

where $\cdots < \tau_{-1} < \tau_0 \le 0 < \tau_1 < \cdots$ are the occurrence times of the $\mathcal{T}$-transitions.

A Palm probability measure, like any probability measure, has an associated expectation, conditional probabilities, etc. For instance, for any $g : \mathbb{D} \to \mathbb{R}$, the function $g(X)$ of the CTMC X under the probability $P_{\mathcal{T}}$ has the expectation

$$E_{\mathcal{T}}[g(X)] = \frac{1}{\lambda_{\mathcal{T}}} \sum_i p_i \sum_{j\neq i} q_{ij} E[g(X)|X_{-1} = i, X_0 = j],$$

provided the double sum is finite for $|g|$ in place of g.

Now, let us see how a typical Palm probability we are studying is related to a standard conditional probability. The following result shows that the Palm probability is a limit of conditional probabilities. The conditioning says that there is at least one $\mathcal{T}$-transition in $(t, 0]$ and a $\mathcal{T}$-transition is guaranteed to occur at time 0 as $t \uparrow 0$. This limit representation is analogous to that for standard conditional probabilities conditioned on a continuous random variable.

Proposition 97. *For any event A generated by X,*

$$P_{\mathcal{T}}(A) = \lim_{t\uparrow 0} P(A|X \in \mathcal{T},\, X(0) \neq X(t)).$$

Proof. Letting $\mathcal{T}'_A \subset \mathcal{T}$ be as in (4.51),

$$P(A|X \in \mathcal{T},\, X(0) \neq X(t)) = \frac{P\{X \in \mathcal{T}'_A,\, X(0) \neq X(t)\}}{P\{X \in \mathcal{T},\, X(0) \neq X(t)\}}.$$

Conditioning on $X(t)$, $X(0)$ and using the stationarity of X along with limits of transition probabilities in Theorem 18, we have, for $t < 0$,

$$\begin{aligned}(-t)^{-1} &P\{X \in \mathcal{T},\, X(0) \neq X(t)\} \\ &= \sum_i p_i \sum_{j\neq i} (-t)^{-1} p_{ij}(-t) P\{X \in \mathcal{T} | X(t) = i,\, X(0) = j\} \\ &\to \lambda_{\mathcal{T}} \qquad \text{as } t \uparrow 0.\end{aligned}$$

This limit statement also holds with $\mathcal{T}$ replaced by $\mathcal{T}'_A$. Combining these observations yields

$$\lim_{t\uparrow 0} P(A|X \in \mathcal{T},\, X(0) \neq X(t)) = \frac{\lambda_{\mathcal{T}'_A}}{\lambda_{\mathcal{T}}} = P_{\mathcal{T}}(A).$$

Palm probabilities for stationary processes and time-dependent Palm probabilities for any non-stationary process are discussed in [60, 100]. Our study is only an introduction that covers Palm probabilities for a CTMC viewed at its transition times. We will not cover extensions (requiring more technical material) that consider other random variables associated with the CTMC or times that are not transition times. For instance, in analyzing an $M/M/s$ system, one may be interested in the times at which a service time exceeds a certain high level, and the state of the system at those times. Or if a random reward is received for processing each item, one may be interested in times at which this reward is 0 due to a defective service.

The theory of Palm probabilities also applies to any stochastic process (that need not be stationary or a CTMC), where one considers conditioning on a point occurring at a time t and the Palm probability is a function of t. The definition of time-dependent Palm probabilities for (non-stationary) CTMCs is as follows. Here, we denote the expected infinitesimal rate at which an $\mathcal{T}$-transition is initiated at time t by

$$\begin{aligned}\lambda_{\mathcal{T}}(t) &= E[\alpha_{\mathcal{T}}(X(t))] \\ &= \sum_i P\{X(t)=i\} \sum_{j\neq i} q_{ij} P\{X \in \mathcal{T} | X_{-1}=i, X_0=j\}.\end{aligned}$$

Definition 98. The *time-dependent Palm probability* of a CTMC X conditioned that a $\mathcal{T}$-transition occurs at time $t \in \mathbb{R}$ is the probability measure $P^t_{\mathcal{T}}$ defined, for any event A generated by X, by

$$P^t_{\mathcal{T}}(A) = \frac{\lambda_{\mathcal{T}^t_A}(t)}{\lambda_{\mathcal{T}}(t)},$$

where $\mathcal{T}^t_A = \{x \in \mathcal{T} : \text{the event } A \text{ occurs for the path } S^t x\}$.

It is natural to consider events based on $\mathrm{S}^t X$ (what an observer sees of X from location t). So in particular,

$$P^t_{\mathcal{T}}\{\mathrm{S}^t X \in \mathcal{T}'\} = \frac{\lambda_{\mathcal{T}'}(t)}{\lambda_{\mathcal{T}}(t)}, \qquad \mathcal{T}' \subset \mathcal{T}.$$

Is this definition consistent with Definition 96 for a stationary X? It is because if X is stationary, then $P\{X(t)=i\} = p_i$, and so $P^t_{\mathcal{T}} = P_{\mathcal{T}}$, $t \in \mathbb{R}$. Another consistency condition is that for large t the time-dependent Palm probability, for any non-stationary ergodic CTMC X, is approximately equal to the stationary Palm probability.

Remark 99. Time-dependent Palm probabilities defined above converge to stationary Palm probabilities in that

$$P^t_{\mathcal{T}}\{\mathrm{S}^t X \in \cdot\} \xrightarrow{w} P_{\mathcal{T}}\{X \in \cdot\} \qquad \text{as } t \to \infty.$$

This follows from the definition of $P^t_{\mathcal{T}}$ since $P\{X(t)=i\} \to p_i$, $i \in S$.

The preceding remark also holds for more general time-dependent Palm probabilities of a stochastic process X that is asymptotically stationary (which is a property of an ergodic CTMC; see Exercise 22).

Many properties of Palm probabilities for stationary properties that do not require the process X to be in equilibrium (like those in the next section) extend to time-dependent Palm probabilities. In these cases, probabilities such as $P^t_{\mathcal{T}}\{S^t \in \mathcal{T}'\}$ are used in place of $P_{\mathcal{T}}\{X \in \mathcal{T}'\}$.

4.17 PASTA at Poisson Transitions

Using the material in the last two sections, we will now explore Palm probabilities of the stationary CTMC X at its Poisson transition times. At such a transition, the initiation rate of transitions is independent of the state, and so it appears that the Palm probability of the state of the chain should be the usual probability. For instance, in a stationary $M/M/s$ system at Poisson arrival times, the distribution of the number of items an arrival encounters should be the same as the stationary distribution. This property is an example of a general theorem we now present.

The main result is as follows. It gives necessary and sufficient conditions under which the Palm probability of the state of the stationary CTMC X before (or after) a $\mathcal{T}$-transition is equal to the ordinary probability of being in that state. These equalities are in the same spirit as the classical PASTA[12] property that "Poisson arrivals see time averages".

Proposition 100. (PASTA Before a $\mathcal{T}$-Transition). *A necessary and sufficient condition for*

$$P_{\mathcal{T}}\{X(0-) = i\} = p_i, \qquad i \in S,$$

is that $\mathcal{T}$ has a uniform initiation rate.
(PASTA After a $\mathcal{T}$-Transition). *A necessary and sufficient condition for*

$$P_{\mathcal{T}}\{X(0) = i\} = p_i, \qquad i \in S,$$

is that the reverse-time $\mathcal{T}$ has a uniform initiation rate.

This result coupled with Theorems 89 and 91 imply that if X satisfies either one of these PASTA properties, then $N_{\mathcal{T}}$ is a Poisson process.

Proof. Recall that $\mathcal{T}$ has a uniform initiation rate if $\alpha_{\mathcal{T}}(i)$ is the same for each i. By the definition of $P_{\mathcal{T}}$,

[12] Since PASTA is a mind-arousing acronym associated with a popular comfort food, we will also use it here to describe what one might call *Palm "actions" see time averages.*

$$P_{\mathcal{T}}\{X(0-) = i\} = \frac{p_i \alpha_{\mathcal{T}}(i)}{\sum_k p_k \alpha_{\mathcal{T}}(k)}.$$

This ratio equals p_i if and only if $\alpha_{\mathcal{T}}(i) = \sum_k p_k \alpha(k)$, for each i, which is equivalent to $\mathcal{T}$ having a uniform initiation rate. This proves the first assertion.

The second assertion follows by a similar argument since by an interchange of sums

$$\begin{aligned} P_{\mathcal{T}}\{X(0) = i\} &= \frac{\sum_j p_j q_{ji} P\{X \in \mathcal{T} | X_{-1} = j, X_0 = i\})}{\sum_\ell p_\ell \sum_k q_{\ell k} P\{X \in \mathcal{T} | X_{-1} = \ell, X_0 = k\}} \\ &= \frac{p_i \bar{\alpha}_{\mathcal{T}}(i)}{\sum_k p_k \bar{\alpha}_{\mathcal{T}}(k)}, \end{aligned}$$

where $\bar{\alpha}_{\mathcal{T}}(i) = p_i^{-1} \sum_j p_j q_{ji} P\{X \in \mathcal{T} | X_{-1} = j, X_0 = i\}$.

Here is a classic illustration of PASTA that leads to the characterization of sojourn times in queues.

Example 101. Waits in an $M/M/s$ system. Suppose $X(t)$ denotes the number of items in an $M/M/s$ system at time t, where $s < \infty$. Assume this CTMC is stationary with distribution p; we know its stationary probabilities for $i \geq s$ are $p_i = p_s(\lambda/s\mu)^{i-s}$, where λ and μ are the arrival and service rates. We saw in Examples 90 and 92 that the initiation rate of arrivals has a uniform rate λ, and the time-reversal of departures also has a uniform initiation rate of λ. Hence the arrival and departure processes are both Poisson processes with rate λ.

Furthermore, because of these two uniform initiation rates, X satisfies the before and after PASTA properties in Proposition 100, namely

$$P_{\mathcal{T}}\{X(0-) = i\} = p_i = P_{\mathcal{T}}\{X(0) = i\}.$$

We will now explore other features of the queueing process at arrivals. First, consider the waiting time W in the queue (prior to its service) of an item that arrives at time 0. This is also the time $W = \min\{t \geq 0 : X(t) < s\}$ until a server is available at or after time 0. Viewing arrival times as $\mathcal{T}$-transitions with $\lambda_{\mathcal{T}} = \sum_i p_i q_{i,i+1} = \lambda$ and using Definition 96,

$$\begin{aligned} P_{\mathcal{T}}\{W > 0\} &= P_{\mathcal{T}}\{X(0) \geq s\} \\ &= \frac{1}{\lambda} \sum_{i=s}^{\infty} p_i q_{i,i+1} \;=\; p_s/(1 - \lambda/s\mu). \end{aligned}$$

As an aside, note that the preceding probability is

$$P_{\mathcal{T}}\{X(0) \geq s\} = P\{X(0) \geq s, A\},$$

where A is the event that following time 0 an arrival occurs before a service completion. This is another illustration of the difference between $P_{\mathcal{T}}$ and P.

Now, the rest of the distribution of W is

$$\begin{aligned} P_{\mathcal{T}}\{W > t\} &= \frac{1}{\lambda}\sum_{i=s}^{\infty} p_i q_{i,i+1} P_{i+1}\{W > t\} \\ &= p_s \sum_{i=s}^{\infty} (\lambda/s\mu)^{i-s} P\{M(t) \le i - s\}. \end{aligned}$$

Here M is a Poisson process with rate $s\mu$ representing the departure process from the s busy servers. Then $M(t)$ is the number of items that can enter service by time t, and so $M(t) \le i - s$ is the event that no more than the $i - s$ items the arrival encountered waiting in the queue at time 0 have entered service by time t. Substituting the Poisson probabilities in the last display, we obtain

$$P_{\mathcal{T}}\{W > t\} = P_{\mathcal{T}}\{W > 0\} P\{W^* > t\},$$

where W^* has an exponential distribution with rate $s\mu - \lambda$.

Another way of expressing this result is

$$P_{\mathcal{T}}\{W \le t | W > 0\} = P\{W^* \le t\},$$

so the exponential time W^* is the duration of an arrival's wait in the queue given that there is a wait. Exercise 19 explores a slightly different "virtual" waiting time process without Palm probabilities.

Next, note that the sojourn time in the system of an arrival at time 0 is $\hat{W} = W + \xi$, where ξ is a service time that is independent of W. Using this independence, we have

$$P_{\mathcal{T}}\{W + \xi \le t | W > 0\} = P\{W^* + \xi \le t\},$$

where W^* and ξ are independent waiting and service times as above. Therefore,

$$\begin{aligned} P_{\mathcal{T}}\{\hat{W} \le t\} = P_{\mathcal{T}}\{W = 0\} P\{\xi \le t\} \\ + P_{\mathcal{T}}\{W > 0\} P\{W^* + \xi \le t\}. \end{aligned}$$

Also,

$$E_{\mathcal{T}}[\hat{W}] = E_{\mathcal{T}}[\xi + W] = 1/\mu + s\mu p_s/(s\mu - \lambda)^2.$$

For the $M/M/1$ system, the preceding shows that $P_{\mathcal{T}}\{W > 0\} = \lambda/\mu$, and Exercise 43 shows that the distribution of $\hat{W}$ reduces to an exponential distribution with rate $\mu - \lambda$.

Let us see an example of when PASTA does not hold.

Example 102. Times of Jumps From One Set to Another. Consider the times at which the CTMC X jumps from A to B, which are two disjoint, nonempty subsets of S. These are $\mathcal{T}$-transition times where

$$\mathcal{T} = \{x \in \mathbb{D} : x(0-) \in A, x(0) \in B\}.$$

Also, when $A = B^c$, these transition times are hitting times of B.

By the comments at the beginning of this chapter, the point process $N_{\mathcal{T}}$ of these transition times is stationary; the initiation rate of jumps from A to B is $\alpha_{\mathcal{T}}(i) = \sum_{j\in B} q_{ij}\mathbf{1}(i \in A)$, and and the rate of $N_{\mathcal{T}}$ is

$$\lambda_{\mathcal{T}} = E[N_{\mathcal{T}}(0,1]] = \sum_{i\in A} p_i \sum_{j\in B} q_{ij}.$$

Now, these transition times from A to B do not enjoy the PASTA properties in Proposition 100. Indeed, a $\mathcal{T}$-transition does not have a uniform initiation rate since $\alpha_{\mathcal{T}}(i) = 0$ on A^c and it is positive elsewhere. Similarly, the time-reversal of $\mathcal{T}$ has the initiation rate

$$\bar{\alpha}_{\bar{\mathcal{T}}}(i) = p_i^{-1} \sum_{j\in A} p_j q_{ji} \mathbf{1}(i \in B), \qquad i \in S,$$

and this is 0 on B^c and it is positive elsewhere.

4.18 Relating Palm and Ordinary Probabilities

This section presents formulas that relate expectations and probabilities under $P_{\mathcal{T}}$ to their counterparts under P, and vice versa. These basic formulas are often used when studying stationary processes that are functions of a stationary CTMC. Examples are given in the next section.

We begin by reviewing a few more properties of stationary processes that need not be Markov processes. Suppose that $X = \{X(t) : t \in \mathbb{R}\}$ is a stationary process on a space S with sample paths in $\mathbb{D}$. The following results also apply to a stationary process X' with the time set $\mathbb{R}_+$, because there exists a stationary process X on $\mathbb{R}$ such that $X \stackrel{d}{=} X'$ on $\mathbb{R}_+$.

Definition 103. A sample-path event A for the stationary process X is *shift invariant* if

$$\{X \in A\} = \{\mathrm{S}^t X \in A\}, \qquad t \in \mathbb{R}.$$

The process X is *ergodic* if $P\{X \in A\} = 0$ or 1 for each shift invariant event.

A stationary "ergodic CTMC" is ergodic in this sense (so our earlier use of ergodic is consistent with this one). This and other stationarity properties, including the ergodic theorem below, are proved in [37, 61].

The next result describes a large class of functions of a stationary process X that are also stationary processes. As an elementary example, if $Y(t) = f(X(t))$, where $f : S \to S'$, then Y is stationary. In contrast, the Markov property does not have a similar heredity property — if X is a Markov processes, then Y is typically not a Markov process.

Proposition 104. *Associated with the stationary process X with sample paths in $\mathbb{D}$, suppose*

$$Y(t) = f(\mathrm{S}^t X), \quad t \in \mathbb{R},$$

where f is a function on $\mathbb{D}$ to some space S', which need not be countable. Then Y is stationary. If, in addition, X is ergodic, then so is Y.

Proof. The process Y is stationary, since by its definition and the stationarity of X,

$$\begin{aligned} \mathrm{S}^t Y &= \{Y(u+t) : u \in \mathbb{R}\} = \{f(\mathrm{S}^u(\mathrm{S}^t X)) : u \in \mathbb{R}\} \\ &\stackrel{d}{=} \{f(\mathrm{S}^u X) : u \in \mathbb{R}\} = Y. \end{aligned}$$

Next, assume X is ergodic. To prove Y is ergodic, it suffices to show that, for any shift-invariant event A for Y, there is a corresponding shift-invariant event for X. In this case, the corresponding event is

$$B = \{x \in \mathbb{D} : \{f(\mathrm{S}^u x) : u \in \mathbb{R}\} \in A\},$$

because clearly $\{\mathrm{S}^t Y \in A\} = \{\mathrm{S}^t X \in B\}$ for each t.

Here is an important *ergodic theorem* for stationary processes.

Theorem 105. *Suppose the stationary process X is real-valued with sample paths in $\mathbb{D}$ that are Lebesgue-integrable on bounded sets. If X is ergodic and $E|X(0)|$ is finite, then*

$$\lim_{t\to\infty} t^{-1} \int_0^t X(s)\, ds = E[X(0)] \qquad a.s.$$

The preceding properties of stationary processes also apply to discrete-time processes — just replace the parameter $t \in \mathbb{R}$ by an integer $n \in \mathbb{Z}$. For instance, Theorem 105 would read: If $\{X_n : n \in \mathbb{Z}\}$ is a real-valued stationary ergodic process with $E|X(0)| < \infty$, then

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} X_m = E[X_0] \qquad a.s. \tag{4.53}$$

This is an extension of the classical SLLN for i.i.d. sequences in Chapter 2.

We will now return to studying the stationary CTMC X and its Palm probability measure $P_{\mathcal{T}}$ for the point process $N_{\mathcal{T}}$ of $\mathcal{T}$-transition times

$$\ldots < \tau_{-1} < \tau_0 \le 0 < \tau_1 < \ldots$$

Applications often involve relating means and probabilities under $P_{\mathcal{T}}$ to those under P, or vice versa. The next three results are key tools for these relations.

Theorem 106. (Campbell Formula)[13] *For $f : \mathbb{R} \times \mathbb{D} \to \mathbb{R}$,*

$$E\Big[\int_{\mathbb{R}} f(t, \mathrm{S}^t X) N_{\mathcal{T}}(dt)\Big] = \lambda_{\mathcal{T}} \int_{\mathbb{R}} E_{\mathcal{T}}[f(t, X)]\, dt,$$

provided the expressions are finite with $|f|$ in place of f.

Proof. We will apply the extended Lévy formula in Theorem 52 to evaluate the expectation of

$$\int_{\mathbb{R}} f(t, \mathrm{S}^t X) N_{\mathcal{T}}(dt) = \sum_{n \in \mathbb{Z}} V_n,$$

where $V_n = f(T_n, \mathrm{S}^{T_n} X)\mathbf{1}(\mathrm{S}^{T_n} X \in \mathcal{T})$. Using the Markov property,

$$\begin{aligned} E[V_n | T_n = t, X_{n-1} = i, X_n = j] \\ &= E\Big[f(t, \mathrm{S}^{T_n} X)\mathbf{1}(\mathrm{S}^{T_n} X \in \mathcal{T}) | T_n = t, X_{n-1} = i, X_n = j\Big] \\ &= E\Big[f(t, X)\mathbf{1}(X \in \mathcal{T}) | X_{-1} = i, X_0 = j\Big]. \end{aligned}$$

Then by Theorem 52 and the definition of $P_{\mathcal{T}}$,

$$\begin{aligned} E\Big[\sum_{n \in \mathbb{Z}} V_n\Big] &= \int_{\mathbb{R}} \sum_i p_i \sum_{j \ne i} q_{ij} E\Big[f(t, X)\mathbf{1}(X \in \mathcal{T}) | X_{-1} = i, X_0 = j\Big] dt \\ &= \lambda_{\mathcal{T}} \int_{\mathbb{R}} E_{\mathcal{T}}[f(t, X)] dt. \end{aligned}$$

This proves the assertion.

A variety of stationary processes associated with X are as follows.

Proposition 107. *For $f : \mathbb{R} \times \mathbb{D} \to \mathbb{R}$, the process*

$$Y(t) = \int_{\mathbb{R}} f(t - u, \mathrm{S}^u X) N_{\mathcal{T}}(du), \qquad t \in \mathbb{R},$$

is stationary and ergodic, and

[13] An analogous Campbell formula for a more general process Y and a Palm probability P_N for a point process N on $\mathbb{R}$ (that need not be stationary) with mean measure $\mu(B) = E[N(B)]$, is

$$E\Big[\int_{\mathbb{R}} f(t, Y(t)) N(dt)\Big] = \int_{\mathbb{R}} E_N[f(t, Y(0))]\mu(dt).$$

$$E[Y(0)] = \lambda_{\mathcal{T}} E_{\mathcal{T}}\Big[\int_{\mathbb{R}} f(u, X)\, du\Big],$$

provided this integral is finite with $|f|$ in place of f.

Proof. Since there are a countable number of $\mathcal{T}$-transitions of X, using the change-of-variable $v = u - t$, we can write

$$\begin{aligned} Y(t) &= \sum_{u\in\mathbb{R}} f(t-u, \mathrm{S}^u X)\mathbf{1}(\mathrm{S}^u X \in \mathcal{T}) \\ &= \sum_{v\in\mathbb{R}} f(-v, \mathrm{S}^v(\mathrm{S}^t X))\mathbf{1}(\mathrm{S}^v(\mathrm{S}^t X) \in \mathcal{T}). \end{aligned}$$

This has the form $Y(t) = \phi(\mathrm{S}^t X)$, where $\phi : \mathbb{D} \to \mathbb{R}$, and so Y is stationary and ergodic by Proposition 104. Also, using Theorem 106 and $v = -u$,

$$\begin{aligned} E[Y(0)] &= E\Big[\int_{\mathbb{R}} f(-u, \mathrm{S}^u X) N_{\mathcal{T}}(du)\Big] \\ &= \lambda_{\mathcal{T}} \int_{\mathbb{R}} E_{\mathcal{T}}[f(-u, X)]\, du = \lambda_{\mathcal{T}} \int_{\mathbb{R}} E_{\mathcal{T}}[f(v, X)]\, dv. \end{aligned}$$

The following is an explicit formula for P in terms of $P_{\mathcal{T}}$.

Corollary 108. (Inversion Formula) *For $g : \mathbb{D} \to \mathbb{R}$,*

$$E[g(X)] = \lambda_{\mathcal{T}} E_{\mathcal{T}}\Big[\int_0^{\tau_1} g(\mathrm{S}^t X)\, dt\Big], \tag{4.54}$$

provided these expressions are finite with $|g|$ in place of g. Hence

$$P\{X \in A\} = \lambda_{\mathcal{T}} \int_0^\infty P_{\mathcal{T}}\{\mathrm{S}^t X \in A, \tau_1 \geq t\}\, dt.$$

Proof. We can write

$$\begin{aligned} g(\mathrm{S}^t X) &= g(\mathrm{S}^t X) \sum_{n\in\mathbb{Z}} \mathbf{1}(\tau_n \leq t < \tau_{n+1}) \\ &= \int_{\mathbb{R}} g(\mathrm{S}^{t-u}(\mathrm{S}^u X))\mathbf{1}(t-u \in [0, h(\mathrm{S}^u X)) N_{\mathcal{T}}(du), \end{aligned}$$

where $h(\mathrm{S}^u X) = \inf\{v > 0 : \mathrm{S}^v(\mathrm{S}^u X) \in \mathcal{T}\}$. Then Proposition 107 yields

$$E[g(X)] = \lambda_{\mathcal{T}} E_{\mathcal{T}}\Big[\int_{\mathbb{R}} g(\mathrm{S}^u X)\mathbf{1}(u \in [0, h(X))\, du\Big].$$

This equality is the same as (4.54), since $h(X) = \tau_1$, when $\tau_0 = 0$. The second assertion is a special case of (4.54).

4.19 Stationarity Under Palm Probabilities

As in the preceding sections, suppose that X is a stationary CTMC and $\ldots < \tau_{-1} < \tau_0 \le 0 < \tau_1 < \ldots$ are the occurrence times of $\mathcal{T}$-transitions of X. Consider the sequence $\{\tau_{n+1} - \tau_n : n \in \mathbb{Z}\}$ of times between $\mathcal{T}$-transitions. This sequence is not stationary under P, but it is stationary under the Palm probability $P_{\mathcal{T}}$. This property is an example of the main result here that the sequence of sample paths of X observed at $\mathcal{T}$-transition times is stationary and ergodic under the Palm probability $P_{\mathcal{T}}$. Consequently, several families of stationary processes associated with the CTMC satisfy a SLLN under $P_{\mathcal{T}}$. We apply the main result to characterize sequences of sojourn and travel times for CTMCs.

Since the CTMC X is stationary in the time parameter t, it might suggest that the sequence $\{X(\tau_n) : n \in \mathbb{Z}\}$ of X-values at the $\mathcal{T}$-transitions should be stationary in the parameter n. This sequence is not stationary under P, but it is under $P_{\mathcal{T}}$. The justification for this statement comes from the following important stationarity property for the sequence of sample paths of X at transition times defined by

$$Y_n = \mathrm{S}^{\tau_n} X = \{X(t + \tau_n) : \ t \in \mathbb{R}\}, \qquad n \in \mathbb{Z}.$$

Theorem 109. *The sequence $\{Y_n : n \in \mathbb{Z}\}$ with values in $\mathbb{D}$ is stationary and ergodic under the Palm probability $P_{\mathcal{T}}$.*

Proof. The sequence $Y = \{Y_n : n \in \mathbb{Z}\}$ is stationary if and only if

$$P_{\mathcal{T}}\{\mathrm{S}Y \in C\} = P_{\mathcal{T}}\{Y \in C\}, \qquad C \subset \mathcal{S}^{\infty},$$

where $\mathrm{S}Y = \{Y_{n+1} : n \in \mathbb{Z}\}$ is Y shifted by one time unit. To prove this equality, consider

$$\tau(x) = \inf\{t > 0 : \mathrm{S}^t x \in \mathcal{T}\}, \qquad x \in \mathcal{T}$$

which is the time between the $\mathcal{T}$-transition of x at time 0 and the first one after 0 (a $\mathcal{T}$-transition of x occurs at time 0 since $\mathrm{S}^0 x = x \in \mathcal{T}$). Next, define the transformation θ from $\mathcal{T}$ to $\mathcal{T}$ by $\theta x = \mathrm{S}^{\tau(x)} x$, for $x \in \mathcal{T}$. Finally, define the iterates θ^n by

$$\theta^n x = \theta(\theta^{n-1} x), \qquad n \ge 1.$$

Then it follows by induction, using $\mathrm{S}^{\tau_{n+1}} X = \mathrm{S}^{\tau_{n+1} - \tau_n}(\mathrm{S}^{\tau_n} X)$, that

$$\mathrm{S}^{\tau_n} x = \theta^n x, \qquad x \in \mathcal{T}, \ n \ge 0.$$

With this notation and (4.52), we have,[14] for $t \ge 0$,

[14] Here we use the shorthand $\{\mathrm{S}^{\tau_n} Y_k\} = \{\mathrm{S}^{\tau_n} Y_k : n \in \mathbb{Z}\}$, and represent other sequences similarly.

$$P_{\mathcal{T}}\{Y \in C\} = \frac{1}{t\lambda_{\mathcal{T}}} E\Big[\sum_{n=1}^{N_{\mathcal{T}}(t)} \mathbf{1}(\{\mathrm{S}^{\tau_n} Y\} \in C)\Big]$$
$$= \frac{1}{t\lambda_{\mathcal{T}}} E\Big[\sum_{n=1}^{N_{\mathcal{T}}(t)} \mathbf{1}(\{\theta^n X\} \in C)\Big].$$

By a similar argument followed by the use of the preceding, we have

$$P_{\mathcal{T}}\{\mathrm{S}Y \in C\} = \frac{1}{t\lambda_{\mathcal{T}}} E\Big[\sum_{n=1}^{N_{\mathcal{T}}(t)} \mathbf{1}(\{\theta^{n+1} X\} \in C)\Big]$$
$$= P_{\mathcal{T}}\{Y \in C\} + \frac{1}{t\lambda_{\mathcal{T}}} P\{\{\theta^{N_{\mathcal{T}}(t)+1} X\} \in C)\}$$
$$- \frac{1}{t\lambda_{\mathcal{T}}} P\{\{\theta X\} \in C, N_{\mathcal{T}}(t) \geq 1\}.$$

The $N_{\mathcal{T}}(t) \geq 1$ in the last statement is because it is implicit in the first summation. Letting $t \to \infty$ proves $P_{\mathcal{T}}\{\mathrm{S}Y \in C\} = P_{\mathcal{T}}\{Y \in C\}$.

Since the CTMC X is ergodic, one can show, as in the proof of Proposition 104, that Y is ergodic.

The preceding result yields several important SLLNs. Since $f(Y_n)$ is stationary and ergodic, by a discrete-time version of Proposition 104, the following result is a consequence of the ergodic theorem in (4.53).

Corollary 110. (SLLN Under Palm Probabilities) *For the stationary, ergodic process $Y_n = \mathrm{S}^{\tau_n} X$ in Theorem 109 and $f : \mathbb{D} \to \mathbb{R}$,*

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} f(Y_m) = E_{\mathcal{T}}[f(Y_1)] \qquad \text{a.s. under } P_{\mathcal{T}}, \tag{4.55}$$

provided the last expectation is finite.

Example 111. Special cases of the preceding convergence a.s. under $P_{\mathcal{T}}$ are

$$n^{-1} \sum_{m=1}^{n} g(X(\tau_m)) \to E_{\mathcal{T}}[g(X(\tau_1))], \qquad n^{-1} \sum_{m=1}^{n} (\tau_m - \tau_{m-1}) \to E_{\mathcal{T}}[\tau_1],$$

for $g : S \to \Re$ and $E_{\mathcal{T}}[|g(X(\tau_1))|] < \infty$. Here $X(\tau_n)$ is stationary and ergodic by Theorem 109 since $g(X(\tau_n)) = f(Y_n)$, where $f(x) = g(x(0))$. A similar statement is true for the sequence $\tau_n - \tau_{n-1}$.

Theorem 109 provides the following framework for characterizing sequences of sojourn and travel times; the general mean-value formula (4.56) is analogous to the Little law for queues in Theorem 57 in Chapter 2.

Proposition 112. (Sojourn and Travel Times) *Associated with $\mathcal{T}$-transition times τ_n of X, assume that $W_n = h(\mathrm{S}^{\tau_n} X)$ is a waiting time for a certain event to occur, where $h : \mathbb{D} \to \mathbb{R}_+$. Then under $P_{\mathcal{T}}$ the waiting time sequence $\{W_n : n \in \mathbb{Z}\}$ is stationary and ergodic, and*

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} W_m = E_{\mathcal{T}}[W_0] \qquad \text{a.s. under } P_{\mathcal{T}}.$$

In addition, this limit is given by

$$E[Y(0)] = \lambda_{\mathcal{T}} E_{\mathcal{T}}[W_0], \tag{4.56}$$

where $Y(t) = \sum_{n\in\mathbb{Z}} \mathbf{1}(\tau_n \le t < \tau_n + W_n)$, $t \in \mathbb{R}$.

Proof. The first assertion follows by Proposition 104, Theorem 109, and Corollary 110. Next, note that

$$Y(t) = \int_{\mathbb{R}} \mathbf{1}(0 \le t - u < h(\mathrm{S}^u X)) N_{\mathcal{T}}(du).$$

Then (4.56) follows since by Proposition 107

$$E[Y(0)] = \lambda_{\mathcal{T}} E_{\mathcal{T}}\Big[\int_{\mathbb{R}} \mathbf{1}(0 \le u < W_0)\, du\Big] = \lambda_{\mathcal{T}} E_{\mathcal{T}}[W_0].$$

Here are several applications of the preceding proposition.

Example 113. Waiting Times in a Set. In the context of Proposition 112, suppose the $\mathcal{T}$-transition times τ_n are the times at which items enter a set B. Let $W_n = h(\mathrm{S}^{\tau_n} X)$ be the sojourn time of X in B starting at time τ_n, where $h(x) = \inf\{t > 0 : x(t) \in B^c\}$.

Then by Propositions 112 and 86 with $Y(t) = \mathbf{1}(X(t) \in B)$,

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} W_m = \frac{\sum_{i\in B} p_i}{\sum_{i\in B^c} p_i \sum_{j\in B} q_{ij}} \qquad \text{a.s. under } P_{\mathcal{T}}.$$

Example 114. Waiting Times in an $M/M/s$ system. Suppose that X is the stationary $M/M/s$ queueing process as in Example 101, and the $\mathcal{T}$-transition times τ_n are the times at which items arrive to the system with rate $\lambda_{\mathcal{T}} = \lambda$. Let $W_n = h(\mathrm{S}^{\tau_n} X)$ denote the length of time the arrival at time τ_n spends waiting for service, where $h(x) = \inf\{t \ge 0 : x(t) < s\}$, and note that

$$s \wedge X(t) = \sum_{n\in\mathbb{Z}} \mathbf{1}(\tau_n \le t < \tau_n + W_n), \qquad t \in \mathbb{R}.$$

Therefore by Proposition 112, W_n is stationary and ergodic, where (4.56) is

$$E[s \wedge X(0)] = \lambda E_{\mathcal{T}}[W_0].$$

Next, consider the time $\hat{W}_n = W_n + \xi_n$ that the arrival at time τ_n spends in the system, where ξ_n is its service time. It follows that $\hat{W}_n$ is stationary and ergodic since (W_n, ξ_n) is. The distributions and means of W_n and $\hat{W}_n$ under $P_{\mathcal{T}}$ were presented in Example 101.

Example 115. Inter-departure Times from an $M/M/s$ system. Suppose X is the stationary $M/M/s$ queueing process as in Example 101. We know from Example 90 that the point process $N_{\mathcal{T}}(t)$ of departures is a Poisson process with rate $\lambda_{\mathcal{T}} = \lambda$ (this is, of course, with respect to the underlying probability measure P). Then the times between departures $W_n = \tau_n - \tau_{n-1}$ are i.i.d. exponential random variables with rate λ. These times are considerably different with respect to the Palm probability $P_{\mathcal{T}}$.

In fact the inter-departure times W_n satisfy the statements in Proposition 112, where $Y(t) = X(t)$ and so $E[X(0)] = \lambda E_{\mathcal{T}}[W_0]$.

Here is a more complex example of a travel time whose mean is a function of random lengths of the past and future of a Markov chain.

Example 116. Travel Times Between Two Sets. In Proposition 112, suppose the $\mathcal{T}$-transition times τ_n are the times at which the process X exits a set A and thereafter enters a set B before returning to A. That is

$$\mathcal{T} = \{x \in \mathbb{D} : x(0-) \in A, x(0) \in A^c, \eta_B(x) < \eta_A(x)\},$$

where $\eta_A(x) = \inf\{t > 0 : x(t) \in A\}$. Consider the travel time $W_n = h(\mathrm{S}^{\tau_n} X)$ of X from A to B starting at time τ_n, where

$$h(x) = \inf\{t > 0 : x(t) \in B\}, \qquad x \in \mathcal{T}.$$

Let $\gamma_i = P\{\eta_B(X) < \eta_A(X) | X_0 = i\}$, which is the probability that the embedded Markov chain X_n hits B before A starting at $i \in A^c$. These hitting probabilities are described in Section 1.7. Also, let $\bar{\gamma}_i$ the probability of hitting B before A starting at $i \in A^c$ for the embedded Markov chain of the reverse-time process $\bar{X}$ as in Proposition 63 with transition rates $\bar{q}_{ij} = p_i^{-1} p_j q_{ji}$.

Then by Propositions 112 and 86

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} W_m = \frac{\sum_{i \notin A \cup B} p_i \gamma_i \bar{\gamma}_i}{\sum_{i \in A} p_i \sum_{j \in A^c} q_{ij} \gamma_i} \qquad \text{a.s. under } P_{\mathcal{T}}$$

In this case,

$$E[Y(0)] = \sum_{i \notin A \cup B} p_i \gamma_i \bar{\gamma}_i.$$

To prove this, note that $Y(t)$ is the indicator of X being in a traverse from A to B at time t, and so

$$Y(t) = \mathbf{1}(X(t) \notin A \cup B, \eta_B(\mathrm{S}^t X) < \eta_A(\mathrm{S}^t X), \bar{\eta}_B(\mathrm{S}^t X) < \bar{\eta}_A(\mathrm{S}^t X)),$$

where $\bar{\eta}_A(x) = \sup\{t < 0 : x(t) \in A\}$, the last exit time of x from A. Then the expression above for $E[Y(0)]$ follows by using the Markov property that conditioned on $X(0) = i$, the past and future of X are independent, and $P\{\bar{\eta}_B(X) < \bar{\eta}_A(X))|X(0) = i\} = \bar{\gamma}_i$.

This concludes our discussion of Palm probabilities.

4.20 $G/G/1$, $M/G/1$ and $G/M/1$ Queues

We will now study the equilibrium behavior of several classical queueing processes. We begin with a $G/G/1$ single-server queueing system in which the arrival times form a renewal process and the service times are i.i.d. The first result establishes the convergence of the sequence of waiting times of the items. Then the equilibrium behavior of the queue lengths as well as the waiting times are characterized when the arrival process is Poisson (the $M/G/1$ model), and when the service times are exponential (the $G/M/1$ model). The queue-length processes in these two models are regenerative processes and their state at certain transition times are Markov chains that have tractable stationary distributions.

We will consider a general processing system that operates as follows. Items arrive to the system at times $0 < \tau_1 < \tau_2 < \ldots$ that form a renewal process. Denote the i.i.d. inter-arrival times by $U_n = \tau_n - \tau_{n-1}$, where $\tau_0 = 0$. The service times are i.i.d. nonnegative random variables V_n that are independent of the arrival times. The service discipline is first-come-first-served with no preemptions, and the inter-arrival and service times have finite means.

The *traffic intensity* of the process is $\rho = E[V_1]/E[U_1]$, the arrival rate $1/E[U_1]$ divided by the service rate $1/E[V_1]$.

Let $Q(t)$ denote the quantity of items in the system at time t, and let W_n denote the length of time that the item arriving at time τ_n spends in the queue before being processed. For simplicity, assume the system is empty at time 0, so $Q(0) = 0$ and $W_0 = 0$.

Definition 117. The process $\{Q(t) : t \geq 0\}$ is an $G/G/1$ *queuing process.* Special cases are:

$M/G/1$ Process — The input process is Poisson.
$G/M/1$ Process — The service times are exponentially distributed.
$M/M/1$ Process — Poisson input process and exponential service times.

The first step is to formally define the processes $Q(t)$ and W_n as functions of the system data U_n and V_n. Applying the inductive construction for analogous discrete-time waiting times in Example 24 in Chapter 1, we have the Lindley recursion

$$W_n = (W_{n-1} + V_{n-1} - U_n)^+, \qquad n \geq 1.$$

We also learned that, due to the i.i.d assumptions on the system data,

$$W_n = \max_{0 \leq m \leq n} \sum_{\ell=m+1}^{n} (V_{\ell-1} - U_\ell) \stackrel{d}{=} \max_{0 \leq m \leq n} Z_m, \tag{4.57}$$

where $Z_n = \sum_{m=1}^{n} (V_m - U_m)$ and $Z_0 = 0$.

Another quantity of interest is the time at which the nth item departs which is

$$D_n = \tau_n + W_n + V_n, \qquad n \geq 1.$$

Using these functions of the system data, the quantity of items in the system at time t is given by

$$Q(t) = \sum_{n=1}^{\infty} \mathbf{1}(\tau_n \leq t < D_n), \qquad t \geq 0.$$

We will now study the limiting behavior of these processes. We were able to describe many features of the $M/M/1$ queueing process $Q(t)$ since it is a CTMC. However, the $G/G/1$ system, and even the non-Markovian $M/G/1$ and $G/M/1$ processes are considerably more complicated, and so many of their features do not have tractable expressions.

As a first step, we show that the waiting times W_n for the $G/G/1$ system converge in distribution to a random variable W, and that this limit is a finite-valued random variable if the traffic intensity is below 1. Limits of waiting times in $G/G/1$ systems in heavy traffic (when the traffic intensity is not below 1) are described later in Section 5.16.

Theorem 118. *For the $G/G/1$ system, the waiting times W_n satisfy*

$$W_n \stackrel{d}{\to} W = \sup_{0 \leq m < \infty} Z_m \qquad \text{as } n \to \infty. \tag{4.58}$$

The limit W has the property

$$P\{W = \infty\} = \begin{cases} 0 \text{ if } \rho < 1 \\ 1 \text{ if } \rho > 1. \end{cases} \tag{4.59}$$

Furthermore, $F(t) = P\{W \leq t\}$ satisfies the Wiener-Hopf integral equation

$$F(t) = \int_{(-\infty, t]} F(t-s) dG(s), \qquad t \geq 0,$$

where $G(t) = P\{Z_1 \leq t\}$, $t \in \mathbb{R}$.

Proof. By (4.57), we know $W_n \stackrel{d}{=} \max_{0\le m\le n} Z_m$ and, as $n \to \infty$, this maximum increases a.s., and hence in distribution, to $\sup_{0\le m<\infty} Z_m$. Thus (4.58) is true.

Next, by the SLLN

$$\lim_{n\to\infty} n^{-1} Z_n \to E[Z_1] = (\rho - 1)E[U_1] \quad \text{a.s.},$$

it follows that $Z_n \to \infty$ or $-\infty$ a.s. according as $\rho > 1$ or $\rho < 1$. Then

$$P\{\sup_{0\le m<\infty} Z_m = \infty\} = \begin{cases} 0 & \text{if } \rho < 1 \\ 1 & \text{if } \rho > 1. \end{cases}$$

This proves (4.59) since $W = \max_{0\le m<\infty} Z_m$.

Finally, using this representation for W, it follows that, for $t \ge 0$,

$$\begin{aligned} F(t) &= \int_{(-\infty,t]} P\{\sup_{0\le m<\infty} (Z_m - Z_1) \le t - Z_1 | Z_1 = s\} P\{Z_1 \in ds\} \\ &= \int_{(-\infty,t]} P\{\sup_{1\le m<\infty} Z_{m-1} \le t - s\} dG(s) \\ &= \int_{(-\infty,t]} F(t-s) dG(s). \end{aligned}$$

From the preceding result, we know that the waiting times have a limit W, and the distribution of W satisfies a Wiener-Hopf integral equation. This distribution has a tractable solution for the $M/G/1$ and $G/M/1$ queues. The rest of this section derives these distributions along with properties of the queue lengths.

First suppose that $Q(t)$ is the $M/G/1$ queue-length process with Poisson arrival process M with rate λ. Consider the quantity $X_n = Q(D_n)$ of items in the system at the departure time D_n of the nth item to enter the system. This discrete-time process, which is embedded in the continuous-time process $Q(t)$, satisfies the recursion

$$X_{n+1} = X_n + Y_n - \mathbf{1}(X_n > 0), \qquad n \ge 0,$$

where $Y_n = M(D_n, D_n + V_{n+1}]$ is the number of arrivals in the interval $(D_n, D_n + V_{n+1}]$ during the service period of the $(n+1)$th item.

Because the Poisson arrival process has stationary independent increments, $Y_n \stackrel{d}{=} M(V)$, where $M(t)$ is a Poisson process with rate λ that is independent of $V \stackrel{d}{=} V_{n+1}$. Then

$$\begin{aligned} a_k &= P\{Y_n = k\} = \int_{\mathbb{R}_+} \frac{e^{-\lambda t}(\lambda t)^k}{k!} P\{V \in dt\}, \qquad (4.60) \\ E[Y_n] &= \lambda E[V] = \rho. \end{aligned}$$

This nicely structured process X_n has the following properties.

Theorem 119. ($M/G/1$ Model) *The process $\{X_n : n \geq 0\}$ is an irreducible, aperiodic Markov chain with transition matrix*

$$\mathsf{P} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

The Markov chain X_n is ergodic if and only if $\rho < 1$. In this case, its stationary distribution π has the generating function

$$G(s) = \sum_{i=0}^{\infty} \pi_i s^i = \frac{(1-\rho)(1-s)\psi(\lambda(1-s))}{\psi(\lambda(1-s)) - s}, \tag{4.61}$$

where $\psi(s) = E[e^{-sV}]$.

Proof. The first assertion follows since under the queueing assumptions

$$\begin{aligned} P\{X_{n+1} = j | X_0, \ldots, X_{n-1}, X_n = i\} &= P\{X_n + Y_n - \mathbf{1}(X_n > 0) = j | X_n = i\} \\ &= P\{Y_1 = j - i + \mathbf{1}(i > 0)\}. \end{aligned}$$

The irreducibility and aperiodicity follow since the a_i in (4.60) are positive.

Next, note that the balance equations $\pi = \pi\mathrm{P}$ are

$$\pi_i = \pi_0 a_i + \pi_1 a_i + \pi_2 a_{i-1} + \cdots + \pi_{i+1} a_0, \qquad i \geq 0. \tag{4.62}$$

Clearly $\pi_{i+1}a_0$ is a function of $\pi_0, \ldots, \pi_i$. In particular, by inductively summing the first $i+1$ equations and solving for $\pi_{i+1}a_0$ (for $i = 0, 1, \ldots$), we obtain the equations

$$\pi_{i+1} a_0 = \pi_0 b_i + \pi_1 b_{i-1} + \cdots + \pi_i b_1, \qquad i \geq 0, \tag{4.63}$$

where $b_i = 1 - \sum_{j=0}^{i} a_j$. From this it is clear that, for $\pi_0 \geq 0$, there is a unique solution π to these equations. The solution is positive when $\pi_0 > 0$ since each $b_i > 0$, and the solution is $\pi_i = 0$ when $\pi_0 = 0$.

Now, the solution π will be a distribution if and only if $\pi_0 > 0$ and $G(1) = 1$, where $G(s) = \sum_{i=0}^{\infty} \pi_i s^i$. To determine when this occurs, note that multiplying (4.62) by s^i and summing on i, we have

$$G(s) = \pi_0 H(s) + s^{-1}[G(s) - \pi_0]H(s),$$

where $H(s) = \sum_{i=0}^{\infty} a_i s^i$. Then

$$G(s) = \frac{\pi_0(1-s)H(s)}{H(s)-s}. \tag{4.64}$$

Using L'Hõpital"s rule and $H'(1) = \rho$ as noted in (4.60), it follows that $G(1) = 1$ if and only if $\pi_0 = 1-\rho$. Hence π is a stationary distribution if and only if $\rho < 1$.

We also know from (4.60) that

$$H(s) = \int_{\mathbb{R}_+} E[s^{M(t)}]P\{V \in dt\} = \psi(\lambda(1-s)),$$

where $M(t)$ is a Poisson process with rate λ and $E[s^{M(t)}] = e^{-\lambda(1-s)t}$. Substituting this expression for $H(s)$ and $\pi_0 = 1-\rho$ in (4.64) proves (4.61).

For this $M/G/1$ model with $\rho < 1$, it is clear that $Q(t)$ is a regenerative process at the times at which it enters state 0. Also, the time between regenerations has a finite mean, since the busy period discussed in Exercise 48 has a finite mean. Hence $Q(t)$ has a limiting distribution. Furthermore, one can show that its limiting distribution is the same as the limiting distribution of the embedded chain X_n (e.g., see [76]). In addition, the SLLN and central limit theorem in Chapter 2 for regenerative-increment processes apply to this queueing process.

Knowing the stationary distribution for an ergodic $M/G/1$ queueing process, we can now characterize the limiting distribution of its waiting times.

Theorem 120. *For the $M/G/1$ queueing process with $\rho < 1$, the waiting times satisfy $W_n \xrightarrow{d} W$ as $n \to \infty$, and W has the Laplace transform*

$$E[e^{-sW}] = \frac{(1-\rho)s}{s-\lambda+\lambda\psi(s)},$$

where $\psi(s) = E[e^{-sV}]$.

Proof. The nth item departing from the system leaves behind $X_n = Q(D_n)$ items in the system, and so X_n equals the number of arrivals in the time interval $(\tau_n, \tau_n + W_n + V_n]$ during which the nth item is in the system. In other words, $X_n = M(\tau_n, \tau_n + W_n + V_n]$, where M is the Poisson arrival process. Then using $E[s^{M(t)}] = e^{-\lambda t(1-s)}$, we have

$$\begin{aligned} E[s^{X_n}] &= E\left[E[s^{X_n}|V_n, W_n]\right] = E[e^{-\lambda(W_n+V_n)(1-s)}] \\ &= E[e^{-\lambda(1-s)W_n}]E[e^{-\lambda(1-s)V}]. \end{aligned}$$

By Theorem 119, $X_n \xrightarrow{d} X$, where $E[s^X] = G(s)$ in (4.61). Then from the preceding display

$$\lim_{n\to\infty} E[e^{-\lambda(1-s)W_n}] = G(s)/\psi(\lambda(1-s)).$$

From this with $G(s)$ given by (4.61), we have $W_n \xrightarrow{d} W$ where

$$E[e^{-\lambda(1-s)W}] = \frac{(1-\rho)(1-s)}{\psi(\lambda(1-s)) - s},$$

which proves the assertion.

We will now present similar results for a $G/M/1$ queueing process $Q(t)$ with exponential service rate μ. Because of its complicated structure, our focus will be on the embedded process $\hat{X}_n = Q(\tau_n -)$, which depicts the quantity in the system just prior to the arrival time τ_n of the nth item. The quantities $\hat{X}_n$ satisfy the recursion

$$\hat{X}_{n+1} = \hat{X}_n + 1 - \hat{Y}_n, \qquad n \geq 0,$$

where $\hat{Y}_n$ is the number of items that depart in the time interval $(\tau_n, \tau_{n+1}]$.

Because the exponential service times with rate μ are memoryless, the residual service time of the item in service at time τ_n is exponential with rate μ, and so the potential departures in an interval between arrivals occur according to a Poisson process $\hat{M}(t)$ with rate μ. Consequently,

$$\begin{aligned} \alpha_k &= P\{\hat{Y}_n = k | \hat{X}_n = i\} = P\{\hat{M}(U) = k\} \\ &= \int_{\mathbb{R}_+} \frac{e^{-\mu t}(\mu t)^k}{k!} P\{U \in dt\} \qquad \text{if } k < i, \end{aligned} \tag{4.65}$$

where $U \stackrel{d}{=} U_n$ is independent of the Poisson process $\hat{M}$. The only other possibility for $\hat{Y}_n$ is that all items are served in an inter-arrival time and so

$$\beta_i = P\{\hat{Y}_n = i+1 | \hat{X}_n = i\} = P\{\hat{M}(U) \geq i+1\} = 1 - \sum_{\ell=0}^{k} \alpha_\ell.$$

These observations are the basis of the following result.

Theorem 121. ($G/M/1$ Model) *The process $\{\hat{X}_n : n \geq 0\}$ is an irreducible, aperiodic Markov chain with transition matrix*

$$\mathsf{P} = \begin{bmatrix} \beta_0 & \alpha_0 & 0 & 0 & 0 & \cdots \\ \beta_1 & \alpha_1 & \alpha_0 & 0 & 0 & \cdots \\ \beta_2 & \alpha_2 & \alpha_1 & \alpha_0 & 0 & \cdots \\ \beta_3 & \alpha_3 & \alpha_2 & \alpha_1 & \alpha_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

The Markov chain X_n is ergodic if and only if $\rho < 1$. In this case, its stationary distribution is

$$\pi_i = (1-r)r^i, \qquad i \geq 0,$$

where r is the unique root in $(0,1)$ of $r = E[e^{-\mu(1-r)U}]$.

Proof. Under the queueing assumptions, $\hat{X}_n$ is a Markov chain with the specified transition matrix since

$$P\{\hat{X}_{n+1} = j | \hat{X}_0, \ldots, \hat{X}_{n-1}, \hat{X}_n = i\} = P\{\hat{Y}_n = i + 1 - j | \hat{X}_n = i\}.$$

Also, the positive α_i ensure that $\hat{X}_n$ is irreducible and aperiodic.

For this Markov chain, the balance equations $\pi = \pi \mathrm{P}$ are

$$\pi_0 = \sum_{j=0}^{\infty} \beta_j \pi_j,$$
$$\pi_i = \sum_{j=0}^{\infty} \alpha_j \pi_{j+i-1} \qquad i \geq 1.$$

To solve this difference equation, we consider a solution of the form $\pi_i = cr^i$ (this approach was used in the Gambler's Ruin Model in Example 44 in Chapter 1). Substituting this π in the last equation and dividing by r^{i-1}, we obtain $r = \phi(r)$, where

$$\phi(r) = \sum_{j=0}^{\infty} \alpha_j r^j = E[e^{-\mu(1-r)U}].$$

The last equality is due to (4.65) and $E[s^{\hat{M}(t)}] = e^{-\mu(1-s)t}$.

We have shown that $\pi_i = cr^i$ is an invariant measure, where r satisfies $r = \phi(r)$, and it is a finite measure if and only if $0 < r < 1$. Now, we know from the branching process Lemma 47 in Chapter 1 that the equation $r = \phi(r)$ has a unique solution r in $(0, 1)$ if and only if $\phi'(1) > 1$. In this case, $\phi'(1) = 1 - \rho$. Hence the Markov chain $\hat{X}_n$ is ergodic if and only if $\rho < 1$ and in that case its stationary distribution is $\pi_i = (1 - r)r^i$.

For the $G/M/1$ model, the limiting distribution of the embedded chain $\hat{X}_n$ is not necessarily the limiting distribution of its parent process $Q(t)$. In fact, $Q(t)$ does not even have a limiting distribution when the inter-arrival times U_n are a constant. However, the limiting distribution for $\hat{X}_n$ yields a tractable limiting distribution for the waiting times.

Theorem 122. *For the $G/M/1$ queueing process with $\rho < 1$, the waiting times satisfy $W_n \xrightarrow{d} W$ as $n \to \infty$, where $P\{W = 0\} = 1 - r$,*

$$P\{W > t\} = re^{-\mu(1-r)t}, \qquad t \geq 0,$$

and r is the smallest root in $(0, 1)$ of $r = E[e^{-\mu(1-r)U}]$.

Proof. We noted above that, because of the memoryless property of the exponential service distribution, the potential departure times during an

inter-arrival time form a Poisson process $\hat{M}(t)$ with rate μ. Then since the arrival at time τ_n has to wait for the $\hat{X}_n$ items ahead of it to be served, it follows that

$$P\{W_n > 0\} = P\{\hat{X}_n > 0\}$$
$$P\{W_n > t\} = P\{\hat{M}(t) < \hat{X}_n\} = \sum_{k=0}^{\infty} P\{\hat{X}_n > k\} P\{\hat{M}(t) = k\}.$$

By Theorem 121, $P\{\hat{X}_n > k\} \to r^{k+1}$ as $n \to \infty$. Applying this to the preceding display yields $W_n \xrightarrow{d} W$, where $P\{W > 0\} = r$ and

$$P\{W > t\} = \sum_{k=0}^{\infty} r^{k+1} P\{\hat{M}(t) = k\} = re^{-\mu(1-r)t}.$$

The last equality is due to $E[r^{\hat{M}(t)}] = e^{-\mu(1-r)t}$.

4.21 Markov-Renewal Processes*

We end this chapter by describing a close relative of a CTMC called a Markov-renewal process. This type of process is a jump process like a CTMC, but the sojourn time in a state has a general distribution that may depend on that state and the next state after the sojourn. Markov-renewal processes may provide more precise models than a CTMC when the exponential sojourn time assumption is not appropriate. Much of the theory of CTMCs readily extends to these processes. We will only present a brief sketch of the equilibrium behavior.

Definition 123. A jump process $\{X(t) : t \geq 0\}$ on S with embedded process (X_n, ξ_n) is a *Markov-renewal process* if it satisfies the following conditions.
(i) X_n is a discrete-time Markov chain on S with transition probabilities $\mathrm{P} = \{p_{ij}\}$, where $p_{ii} = 0$, for each i.
(ii) For nonnegative $t_0, \ldots, t_m$,

$$P\{\xi_0 \leq t_0, \ldots, \xi_m \leq t_n | X_n, n \geq 0\} = \prod_{n=0}^{m} P\{\xi_n \leq t_n | X_n, X_{n+1}\},$$

and there are distributions $F_{ij}(t)$, $i, j \in S$, such that $F_{ij}(0) = 0$ and, for each $n \geq 0$,

$$P\{\xi_n \leq t | X_n = i, X_{n+1} = j\} = F_{ij}(t), \qquad t \geq 0,\ i, j \in S.$$

Some authors refer to $X(t)$ as a *semi-Markov process* and (X_n, ξ_n) as a Markov-renewal process. A typical example is the cyclic renewal process studied in Chapter 3. The notation and many properties of CTMCs, such as classifying states, carry over to Markov-renewal processes. For instance, Proposition 4.4 justifies that the regularity of $X(t)$ implies that the sojourn-time distributions F_{ij} are P-*regular* in the sense that their respective means μ_{ij} are finite and $\sum_{n=0}^{\infty} \xi_n = \infty$ a.s.

For the following result, assume the embedded Markov chain X_n is ergodic with stationary distribution π. Then $X(t)$ is also recurrent and irreducible. As in the preceding sections, the first entrance time of the Markov-renewal process $X(t)$ to a state i is $\tau_i = \sum_{n=0}^{\nu_i - 1} \xi_n$, which is the sum of the sojourn times until X_n reaches i. From Exercise 56,

$$E_i[\tau_i] = \pi_i^{-1} \sum_{j \in S} \pi_j \sum_{\ell \in S} p_{j\ell} \mu_{j\ell}.$$

Assume this is finite. Then $X(t)$ is positive recurrent. Finally, assume for simplicity that at least one of the distributions F_{ij} is non-arithmetic, which ensures that the distribution of τ_i is also non-arithmetic. Under these assumptions, the process $X(t)$ is an ergodic Markov-renewal process with transition probabilities p_{ij} and sojourn distributions F_{ij}.

The Markov-renewal process $X(t)$ is a delayed regenerative process at the times it enters a fixed state. Therefore, its limiting distribution is essentially the same form as that for CTMCs in Theorem 39. The only difference is that the mean sojourn time in state i is now $\sum_{j \in S} p_{ij} \mu_{ij}$ instead of q_i^{-1} for CTMCs. The proof of the following is Exercise 56.

Theorem 124. *The ergodic Markov-renewal process $X(t)$ has the limiting distribution, for a fixed $i \in S$,*

$$p_j = \frac{1}{E_i[\tau_i]} E\Big[\int_0^{\tau_i} \mathbf{1}(X(t) = j)\, dt \Big] = \frac{1}{E_i[\tau_i]} \sum_{\ell \in S} p_{j\ell} \mu_{j\ell}, \qquad j \in S.$$

Furthermore,

$$p_j = \pi_j \sum_{\ell \in S} p_{j\ell} \mu_{j\ell} \Big/ \sum_{i,\ell \in S} \pi_i p_{i\ell} \mu_{i\ell}, \qquad j \in S. \tag{4.66}$$

Ergodic theorems for functionals of Markov-renewal processes are similar to those for CTMCs in Theorems 42 and 45. One simply uses (4.66) as the limiting distribution, and q_j is replaced by $1/\mu_{j\ell}$.

4.22 Exercises

Exercise 1. *Compound Poisson Processes.* Let $X(t)$ denote a compound Poisson process as in Section 3.15 in Chapter 3 with rate λ and jump-size density $f(i)$, $i \in \mathbb{Z}$. Justify that $X(t)$ is a CTMC and specify its transition rates. Assuming $X(t)$ is irreducible on $\mathbb{Z}$, classify its states when the mean of the density f is $= 0$ or is $\neq 0$.

Exercise 2. *Continuation.* Let $X_k(t)$, $k = 1, \ldots, m$, be independent compound Poisson processes with rates λ_k and jump-size densities $f_k(i)$, $i \in \mathbb{Z}$. Show that $X_1(t)+X_2(t)$, and $X_1(t)-X_2(t)$ are CTMCs and that they are also compound Poisson processes; specify their defining parameters. More generally, show that the process $X(t) = \sum_{k=1}^{m} a_k X_k(t)$, for $a_k \in \mathbb{Z}$, is a CTMC and a compound Poisson process and specify its defining parameters. Hint: the jump-size density of $X(t)$ is a mixture of the f_k.

Exercise 3. *Input-Output System: Reflected Compound Poisson Process.* Let $X(t)$ denote the value of a system (e.g., monetary account or storage area in a computer) at time t that takes values in $S = \{i \in \mathbb{Z} : a \leq i \leq b\}$, where $a < b$. The value increases "potentially" by a compound Poisson process $X_1(t)$ with rate λ_1 and nonnegative jump-sizes with density f_1, but part or all of an input is rejected (or disregarded) to the extent that it would force $X(t)$ to exceed b. Also, the value decreases "potentially" by a compound Poisson process $X_2(t)$ with rate λ_2 and nonnegative jump-sizes with density g, but part or all of a decrease is disregarded to the extent that it would force $X(t)$ to fall below a. That is, letting $Z(t) = X_1(t) - X_2(t)$ and $\Delta X(t) = X(t) - X(t-)$,

$$X(t) = X(0)+Z(t)- \sum_{0\leq s\leq t} \Big[(X(s-)+\Delta X_1(s)-b)^+-(a-X(s-)+\Delta X_2(s))^+\Big].$$

This $X(t)$ is the compound Poisson process $Z(t)$ reflected at a and b. Note that it is a continuous-time version of the reflected random walk in Example 23 in Chapter 1. Justify that $X(t)$ is a CTMC and specify its transition rates. For $a = 0$ and $b = \infty$, show that

$$\lim_{t\to\infty} P\{X(t) \leq i\} = P\{\sup_{0\leq s\leq\infty} Z(s) \leq i\}.$$

Exercise 4. *Markov Chains Driven by Clock Times.* Let $\{X_n : n \geq 0\}$ be a stochastic process on S whose dynamics are as follows. Whenever the process enters a state i, a set of independent clock times τ_{ij}, $j \in S_i$ are started, where S_i is the subset of states in S that can be reached from state i in one step. The times τ_{ij} are geometrically distributed with parameters γ_{ij} ($P\{\tau_{ij} > m\} = \gamma_{ij}^m$). Then the sojourn time in state i is the minimum $\tau_i = \min_{j\in S_i} \tau_{ij}$, and, at the end of the sojourn, the process jumps to the state $j \in S_i$ for which $\tau_{ij} = \tau_i$. Find the distribution of τ_i.

Think of τ_{ij} as the time to the next "potential" transition from i to $j \in S_i$ with probability p_{ij}, and the clock time j that is the smallest of these times "triggers" a transition from i to j. Show that X_n is a Markov chain with transition probabilities

$$P\{X_{n+1} = j | X_m, m < n, X_n = i\} = \gamma_{ij} / \sum_{k \in S_i} \gamma_{ik}.$$

Exercise 5. *Multiclass Exponential Clocks.* Consider a jump process $\{X(t) : t \geq 0\}$ with countable state space S that evolves as follows. For each pair of states i, j, there is a countable set of sources $\mathrm{Y}(i,j)$ that may trigger a transition from i to j, provided that such a transition is possible. Specifically, whenever the process $X(t)$ is in state i, the time for source y to "potentially" trigger a transition to j is exponentially distributed with rate $q_y(i,j)$, independent of everything else. Then the time for a transition from i to j is the minimum of these independent exponential times, and so the potential transition time from i to j is exponentially distributed with rate

$$q_{ij} = \sum_{y \in \mathrm{Y}(i,j)} q_y(i,j).$$

Thus, as in Example 9, we know that $X(t)$ is a CTMC with transition rates q_{ij}, provided they are regular with respect to $p_{ij} = q_{ij}/q_i$, where $q_i = \sum_j q_{ij}$.

In this setting, the sources that trigger the transitions are of interest, especially if there are costs or rewards associated with the sources. By properties of exponential variables, the probability that source y is the one that triggers the transition is $q_y(i,j)/q_{ij}$. Let Y_n denote the source that triggers the transition at time T_{n+1}. Consider the process

$$Y(t) = Y_{n+1}, \qquad \text{if } t \in [T_n, T_{n+1}) \text{ for some } n.$$

This $Y(t)$ is the source that triggers the next transition at or after time t.
(a) Show that $(X(t), Y(t))$ is a CTMC on the set $\hat{S} = \{(i,y) : i \in S,\ y \in \cup_{i,j \in S} \mathrm{Y}(i,j)\}$. Specify its exponential sojourn rates $q_{(i,y)}$ and its transition probabilities $p_{(i,y),(j,y')}$.
(b) Show that (X_n, Y_n, T_n) is a discrete time Markov chain and specify the following transition probabilities: For $s,\ t \geq 0,\ (i,y),\ (j,y') \in \hat{S}$,

$$P\{X_{n+1} = j, Y_{n+1} = y', T_{n+1} > s + t | X_n = i, Y_n = y, T_n = s\}.$$

Exercise 6. *$M/M/s$ System With Feedback.* Consider the $M/M/s$ queueing system in Example 60 with the modification that upon completing its service, an item is either fed back for another service with probability r, or it exits the system with probability $1 - r$. In this case, the process $X(t)$ representing the number of items in the system would not change state at a feed-back departure — it experiences a fictitious jump. Justify that $X(t)$ is a CTMC

with parameters $\hat{p}_{i,i} = r\mu_i/\hat{q}_i$,

$$\hat{p}_{i,i+1} = \lambda/\hat{q}_i, \qquad \hat{p}_{i,i-1} = (1-r)\mu_i/\hat{q}_i, \qquad i \geq 1,$$

and $\hat{q}_i = \lambda + \mu_i$, where $\mu_i = \min\{i, s\}$. Show that $X(t)$ can also be represented as a CTMC with transition rates $q_{i,i+1} = \lambda$, $q_{i,i-1} = (1-r)\mu_i$. Specify its parameters p_{ij} and q_i under this alternative representation.

Exercise 7. *Independence of Future and Past Given the Present.* Suppose that $X(t)$ is a CTMC and for $t > 0$, let $Y(t)$ be a random variable generated by the past $\{X(s) : s < t\}$ and let $Z(t)$ be a random variable generated by the future $\{X(u) : u > t\}$. Show that

$$E[Y(t)Z(t)|X(t)] = E[Y(t)|X(t)]E[Z(t)|X(t)].$$

Exercise 8. Two types of jobs, labeled 1 and 2, arrive to a processor according to independent Poisson processes with rates λ_1 and λ_2. Let $X(t)$ denote the type of the last job arrival before time t. Show that $X(t)$ is a CTMC and specify its transition rates. Show that $p_{ij}(t) = a + be^{-(\lambda_1+\lambda_2)t}$, and specify the coefficients a and b. Show that the $X(t)$ is ergodic and find its stationary distribution.

Exercise 9. *Yule Process.* Consider a pure birth process with exponential sojourn rates $q_i = i\lambda$, for some $\lambda > 0$. Find an expression for $p_{ij}(t)$. Hint:A special case is $p_{1j}(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{j-1}$, $j \geq 1$.

Exercise 10. *Search Process.* Items arrive to a system of m cells labeled $1, \ldots, m$ at times that form a Poisson process with rate λ. Each arrival independently enters cell i with probability α_i and remains there until it is deleted by a search. Independently of the arrivals, searches occur at times that form a Poisson process with rate μ, and each search is performed at cell i with probability δ_i. If items are in the search cell, one item is deleted; otherwise no items are deleted and the search is terminated. Let $X(t) = (X_1(t), \ldots, X_m(t))$ denote the numbers of items in the cells at time t. Justify that $X(t)$ is a CTMC and classify its states. Find an invariant measure for it. Prove that $X(t)$ is positive recurrent if and only if $\alpha_i\lambda \leq \delta_i\mu$, $1 \leq i \leq m$. Find the mean and variance of the number of items in node i in equilibrium (that is $E[X_i(0)]$ and $\mathrm{Var}[X_i(0)]$ when $X(t)$ is stationary).

Exercise 11. *Continuation.* In the search model in the preceding exercise, suppose the system is constrained to have $X_1(t) \leq X_2(t) \leq \cdots \leq X_m(t)$. Find an invariant measure for the resulting process $X(t)$ with this constraint.

Exercise 12. *Merging Process.* Two types of items arrive to a merging station at times that form two independent Poisson processes with respective

rates λ_1 and λ_2. The units queue up and merge into pairs (one of each type) as follows. Whenever a type 1 item arrives to the station, it either merges with a type 2 item that is waiting at the station, or it enters a queue if are no type 2 items present. Similarly, a type 2 arrival either merges with a type 1 item or it enters a queue. Let $X_k(t)$ denote the number of type k items at the station at time t ($k = 1$ or 2). Note that either $X_1(t)$ or $X_2(t)$ is 0 at any time. Assume that the station can contain at most m items (which are necessarily of one type), and when this capacity is reached, additional items of the type in the queue are turned away. Examples of such a system are automatically guided vehicles meeting products to be transported, or taxis and customers pairing up at a station.

When the system is in equilibrium, find the following quantities:
- The probability of i_k type k items at the station.
- The probability of more than i items at the station.
- The expected number of type i_k items at the station.

Also, find the average number of type k items that are turned away and not served. Hint: model the system by the process $X(t) = X_1(t) - X_2(t)$.

Exercise 13. For an ergodic CTMC $X(t)$, show that, for any state $i \in S$,

$$\lim_{n\to\infty} n^{-1}T_n = E_i[\tau_i].$$

Then show that $t^{-1}T_{N(t)} \to 1$ a.s. as $t \to \infty$.

Exercise 14. *Balking and reneging in an $M/M/s$ system.* Suppose that $X(t)$ is an $M/M/s$ queueing process with arrival rate λ and service rate μ. Consider the variation in which items balk at entering, such that when n items are in the system, the arrival rate is $b_i\lambda$. For instance, b_i could be the probability that an arrival decides to enter the system when i items are in the system (e.g., $b_i = \mathbf{1}(i \le K)$). Assume $b_i = e^{-\alpha i/\mu}$, where i/μ is an estimate for the average waiting time when i items are present and $\alpha > 0$. Specify the transition rates for the system, give a criterion for the process to be ergodic, and determine its stationary distribution.

Next, consider the $M/M/s$ process $X(t)$ in which an item in the system may renege and depart from the system with rate r_i. That is, the time to the next departure is the minimum of the service times of the items being served and an independent exponential random variable with rate r_i. Therefore, when i items are in the system, the next potential departure is exponentially distributed with rate $\mu i \wedge s + r_{i-1}$. Determine when $X(t)$ is ergodic, and find its stationary distribution.

Finally, determine conditions under which $X(t)$ is ergodic, and find its stationary distribution if there is both balking and reneging as above.

Exercise 15. *Multiclass Service System with Blocking.* Consider a service system that processes m classes of items, but it can serve only one class

at any time. While it is serving items of class c, any arrivals of other classes cannot enter the system and are turned away, but new type c arrivals may enter. Assume class c items enter and depart such that the number in the system behaves as an ergodic CTMC on the nonnegative integers with transition rates $q_c(x, y)$, and its stationary distribution $p_c(x)$ is known.

The system is represented as a CTMC $X(t)$ with states $x = (x_1, \ldots, x_m)$, where x_c is the nonnegative number of class c items in the system — at most, one of the x_c's is positive. Assume the system is empty at time 0. Now, the state space S consists of a center $S_0 = \{0\}$ and point sets

$$S_c = \{(x_1, \ldots, x_m) : x_c > 0, x_l = 0, l \neq c\}, \qquad c = 1, \ldots, m,$$

such that the process $X(t)$ can transfer from a state in S_c to a state in $S_{c'}$, $c' \neq c$, only by passing through 0.

Under the preceding assumptions, the transition rates of $X(t)$ are

$$q(x, y) = \begin{cases} q_c(x, y) & \text{if } x, y \in S_0 \cup S_c, \text{for some } c, \text{ and } y = e_c \text{ if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $X(t)$ is ergodic and its stationary distribution is

$$\pi(x) = \pi(0) p_c(x_c)/p_c(0), \qquad x \in S_c,\ c \neq 0,$$

and $\pi(0) = [1 + \sum_{c \neq 0}(p_c(0)^{-1} - 1)]^{-1}$. A related discrete-time model is in Example 107 in Chapter 1.

Exercise 16. *A Batch-Service System.* Suppose that $X(t)$ is a CTMC that denotes the number of customers in a service system at time t whose transition rates are

$$q_{ij} = \lambda 1(j = i + 1) + \mu 1(j = \max\{0, i - K\}).$$

Here λ, μ, and K are positive, and $\lambda < K\mu$. This represents a system in which customers arrive by a Poisson process with rate λ and are served in batches as follows. Whenever there are $i \geq K$ customers in the system, batches of K customers depart at the rate μ; and whenever $i < K$ customers are present, all of the customers depart at the rate μ. Show that $X(t)$ is ergodic and its stationary distribution is $p_i = r^i(1 - r)$, $i \geq 0$, where r is the unique root in $(0, 1)$ of the equation $\mu r^{K+1} - (\lambda + \mu)r + \lambda = 0$.

Exercise 17. *Continuation.* Assume the batch-service process in the preceding exercise is stationary. Let $N_{\mathcal{T}}$ denote the point process of $\mathcal{T}$-transition times at which batches of size K depart from the system. Describe the set $\mathcal{T}$, and show that N is a Poisson process with rate $\mu + \lambda(1 - r^{-1})$.

Exercise 18. *Average Sojourn Time in a Set.* Let $W_n(A)$ denote the amount of time the CTMC spends in a set A on its nth sojourn in that set. The

average sojourn or waiting time in A is

$$W(A) = \lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} W_m(A) \quad \text{a.s.}$$

Show that $W(A) = \lambda(A)^{-1} \sum_{i\in A} p_i$. Use this to find the average duration of time that an $M/M/1$ spends with less than K items in the system.

Exercise 19. *$M/M/s$ system.* Let $X(t)$ denote a stationary $M/M/s$ queueing process as in Example 60, with $s < \infty$, traffic intensity $\rho = \lambda/s\mu < 1$, and stationary distribution p_i. Let $Y(t)$ denote the number of items in the queue waiting to be served (not in service) at time t. Show that $Y(t)$ is a stationary process and specify its distribution. Also, show that

$$E[Y(0)] = \frac{p_0\rho(s\rho)^s}{(1-\rho)^2 s!}.$$

Consider the waiting time $W(t) = \min\{u \geq t : X(u) < s\}$ until a server is available at or after time t (sometimes called the virtual waiting time in the queue of an arrival at time t before it can enter service). Justify that $W(t)$ is a stationary process, and find $E[W(0)]$. Show that

$$P\{W(0) \leq t | X(0) = s\} = 1 - e^{-rt},$$

and specify the rate r. Show that $P\{W(0) > 0\} = P\{X(0) > s\}$ and

$$P\{W(0) > t\} = \sum_{i=s}^{\infty} p_i P_i\{M(t) < i - s\} = P\{X(0) > s\} e^{-(s\mu-\lambda)},$$

where $M(t)$ is a Poisson process with rate $s\mu$ denoting departure times when s servers are busy. Finally, prove $P\{X(0) \geq s + m\} = (\lambda/s\mu)^m$.

Exercise 20. *Passage Times.* Let $X(t)$ be a CTMC with transition rates q_{ij} (here $q_{ii} = -q_i$). For a fixed set B in the state space S, consider the first passage time $\tau_B = \inf\{t > \xi_0 : X(t) \in B\}$. Show that the hitting probabilities $v_i = P_i\{\tau_B < \infty\}$ of set B are the smallest y_i in $[0, 1]$ that satisfy

$$\sum_j q_{ij} y_j = 0, \qquad i \in B^c.$$

Show that the mean passage times $v_i = E_i[\tau_B]$ are the smallest nonnegative y_i that satisfy $1 + \sum_{j\in B^c} q_{ij} y_j = 0$, $i \in B^c$.

Exercise 21. *Passage Values for Uniformized Chains.* In the setting of the preceding exercise, for $f : S \to \mathbb{R}$, consider the mean values

$$v_i = E_i\left[\int_0^{\tau_B} f(X(t))dt\right], \qquad i \in S.$$

Assuming $\lambda = \sup_i q_i < \infty$, show that

$$v_i = \lambda^{-1} E_i \left[\sum_{n=0}^{\nu_B - 1} f(\hat{X}_n) \right].$$

where $\hat{X}_n$ is a Markov chain and $\nu_B = \inf\{n \geq 1 : \hat{X}_n \in B\}$. Specify the transition probabilities $\hat{p}_{ij}$ for $\hat{X}_n$ in terms of the rates q_{ij} for $X(t)$. Show that the vector v has the form

$$v = e^{\hat{Q}} r = \Big(\sum_j \sum_{n=0}^{\infty} \hat{Q}_{ij}^n r_j : i \in S \Big)$$

and specify the matrix $\hat{Q}$ and the vector r.

Exercise 22. *Asymptotic Stationarity.* Suppose that $X(t)$ is an ergodic CTMC with stationary distribution p and, for $t \geq 0$, let $Z(t)$ be a real-valued random variable generated by the future $\{X(u) : u > t\}$. Let $\bar{X}(t)$ and $\bar{Z}(t)$ denote these processes when $\bar{X}(t)$ is a stationary version of $X(t)$. Show that (conditioning on $X(t)$), $Z(t) \xrightarrow{d} \bar{Z}(0)$, as $t \to \infty$.

(This asymptotic stationarity statement says in particular that $\mathrm{S}^t X \xrightarrow{d} \bar{X}$. A discrete-time version of this is in Proposition 63 in Chapter 1.)

Prove the *asymptotic independence property* that, for $t < u$, if $t \to \infty$ and $u - t \to \infty$, then

$$P\{X(t) = i, X(u) = j\} \to p_i p_j.$$

For $f : S \to \mathbb{R}$ and $a < b$, consider the process $Y(t) = \int_a^b f(X(u+t))du$. Show that there exists a random variable Y such that $Y(t) \xrightarrow{d} Y$ and give a formula for $E[Y]$.

Exercise 23. Let $\hat{X}(t)$ and $X(t)$ on S denote jump processes as described prior to Proposition 23 with respective parameters $(\hat{\alpha}_i, \hat{p}_{ij}, \hat{q}_i)$ and (α_i, p_{ij}, q_i), where $\hat{p}_{ii}$ and p_{ii} are in $[0, 1)$. Assume that $\hat{\alpha}_i = \alpha_i$, $q_i \geq \hat{q}_i$, and

$$p_{ij} = \hat{q}_i \hat{p}_{ij} / q_i, \qquad j \neq i, \qquad p_{ii} = 1 - \hat{q}_i (1 - \hat{p}_{ii}) / q_i.$$

Show that $\hat{X}(t)$ and $X(t)$ are CTMCs that have the same distribution.

Exercise 24. *Sampling a CTMC.* Let $X(t)$ be an ergodic CTMC on S with stationary distribution p_i. Suppose this CTMC is sampled (or observed) at times $T_1 < T_2 < \ldots$ that are occurrence times of a renewal process with inter-renewal distribution F. Show that the sampled values $X(T_n)$ form a Markov chain and specify its transition probabilities. Show that $X(T_n)$ is ergodic with the same stationary distribution p_i that $X(t)$ has. How would you estimate the p_i from observations $X(T_0), X(T_1), \ldots, X(T_n)$? That is, find an estimator $\hat{p}_i(n)$ of p_i and show that it is consistent in that $\hat{p}_i(n) \to p_i$,

for each $i \in S$, as $n \to \infty$. Is the estimator the same if the sampling times formed a Poisson process or if the inter-sampling times were a constant?

Exercise 25. $M/M/1$ *Production-Inventory System.* Consider the system shown in Figure 4.2 consisting of an $M/M/1$ system $\mathcal{M}$ that produces units of a product and a warehouse $\mathcal{W}$ that houses the units until they are requested. Demands for the product occur at times that form a Poisson process with rate λ. An arriving demand is satisfied from the warehouse if a unit is available, otherwise the demand waits outside of $\mathcal{W}$ until a unit arrives from $\mathcal{M}$ and then it is satisfied. In either case, the arrival also triggers a unit to be produced at $\mathcal{M}$, where the service rate of single server is $\mu > \lambda$. Assume the warehouse has a capacity L and, at time 0, $\mathcal{W}$ is full and $\mathcal{M}$ is empty. Let $X(t)$ denote the number of units in $\mathcal{M}$ at time t; this is also the number of demands that are waiting for units. The number of units in $\mathcal{W}$ is $W(t) = L - X(t)$. What type of process is $X(t)$? Specify its stationary distribution.

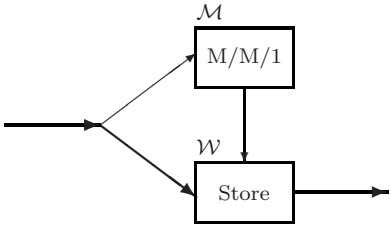

Fig. 4.2 $M/M/1$ Production-Inventory System

The system receives a reward at the arrival time of each demand: the reward is R if the demand is satisfied immediately and is r if the demand incurs a wait to be satisfied. Also, there is an inventory cost of h per unit time of holding one unit in inventory. Find the average reward for the system and the average holding cost. Then find the warehouse capacity L that maximizes the average reward minus holding cost.

Exercise 26. Suppose (X_n, ξ_n) is a Markov chain as in Proposition 23 and as in its proof let $\nu_0 = 0$ and $\nu_{n+1} = \min\{m > \nu_n : X_m \neq X_{\nu_n}\}$, $n \geq 0$. Show that each ν_n is a stopping time of the chain (X_n, ξ_n).

Exercise 27. Let $X(t)$ be a birth-death process with state-dependent birth and death rates λ_i and μ_i. Let $\nu_i = \min\{n \geq 1 : X_n = i\}$. For a cost (or value) function $f : S \times \mathbb{R}_+ \to \mathbb{R}$, show that

$$E_0\Big[\sum_{n=1}^{\nu_k} f(X_n, \xi_n)\Big] = \sum_{j=0}^{k-1} \frac{1}{\lambda_j \eta_j} \sum_{i=0}^{j} (\lambda_i + \mu_i) E_i[f(i, \xi_1)]\eta_i, \tag{4.67}$$

where $\eta_j = \prod_{i=1}^{j} \lambda_{i-1}/\mu_i$. First derive an expression for

$$v_j = E_j\Big[\sum_{n=1}^{\nu_{j+1}} f(X_n, \xi_n)\Big], \qquad 0 \le j \le k,$$

based on a first-step analysis and a recursive equation for the v_j.

Use (4.67) to find an expression for $E_0[T_{\nu_i}]$, the mean first passage time to state i.

Exercise 28. *Counterexamples.* Let $X(t)$ be an irreducible CTMC with sojourn rates q_i whose embedded chain X_n is a random walk on $S = \mathbb{Z}_+$ with transition probabilities $p_{00} = q$ and

$$p_{ij} = p\mathbf{1}(j = i + 1) + q\mathbf{1}(j = i - 1).$$

Show that any q_i are P-regular if $p \le 1/2$.

Assuming $p < q$, show that X_n is ergodic and specify its stationary distribution. In this case, find q_i such that $X(t)$ is not ergodic.

Assuming $p = 1/2$, show that X_n is not ergodic, but $X(t)$ is ergodic for $q_i = 1/a^i$ and $0 < a < 1$.

Exercise 29. *Tandem Network.* Suppose $X(t)$ is a Jackson network process for the tandem network shown in Figure 4.3, where arrivals enter node 1 according to a Poisson process with rate λ and each node i consists of a single server with service rate μ_i. Show that $X(t)$ is ergodic if and only if $\lambda < \mu_i$ for each i. When it is ergodic, establish its stationary distribution.

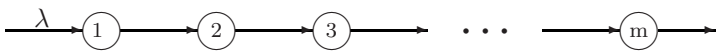

Fig. 4.3 Tandem Network

Now, assume $X(t) = (X_1(t), \ldots, X_m(t))$ is stationary. Use the result in the preceding exercise to find the average duration of time that there are more than L items in the network. Find an expression for $P\{\max_{1\le i\le m} X_i(0) > a\}$. Assume $\mu_i = \mu > \lambda$, for each i, and determine $P\{\sum_{i=1}^m X_i(0) > L\}$.

Exercise 30. *Open Acyclic Jackson Network.* Suppose $X(t)$ is an ergodic Jackson process representing the open network shown in Figure 4.4.

Let p_{ij} denote the probability that an item is routed from node i to node j. Show that a solution to the traffic equations is $w_1 = p_{01}$,

$$w_2 = p_{02} + w_1 p_{12}, \quad w_3 = p_{03} + w_1 p_{13}, \quad w_4 = w_2 p_{24}, \quad w_5 = w_3.$$

Assume $X(t)$ is stationary and let $N_{ij}(t)$ denote the number of times an item moves from node i to node j in a time interval $[0, t]$. Show that N_{ij} is a

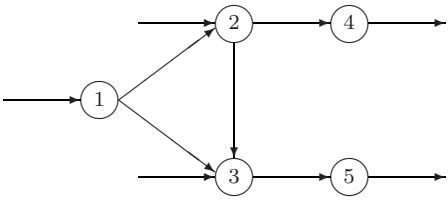

Fig. 4.4 Open Acyclic Jackson Network

Poisson process with rate $w_i\lambda_{ij}$. In other words, the flow between each pair of nodes in the network is a Poisson process. This property is true for any network in which each item can visit a node at most once. While some of the flows are independent, some of them are dependent.

Exercise 31. *Input-Output System with Batch Arrivals.* Let $X(t)$ denote the quantity of items in an input-output system that operates as follows. Batches of items arrive at times that form a Poisson process with rate λ, and the batch sizes are i.i.d. with geometric distribution $(1-\alpha)\alpha^{n-1}$. The items are served by a single server and the service times are independent exponentially distributed with rate $\mu > \lambda/(1-\alpha)$. Justify that $X(t)$ is a CTMC and specify its transition rates. Show that its stationary distribution is

$$p_i = p_0 \frac{\lambda}{\mu} \left(\frac{\lambda + \alpha\mu}{\mu} \right)^{i-1} \qquad i \geq 1.$$

Exercise 32. *Continuation: Batch Services.* Consider the system described in the preceding example with the difference that at the end of a service time with i items in the system, $\min\{i, K\}$ items depart, where K is the batch service capacity. So those items present at the start of a service plus all the arrivals during a service time up to the amount K depart at the end of the service. Here $K\mu > \lambda/(1-\alpha)$. Justify that the quantity of items in the system $X(t)$ is a CTMC with transition rates

$$\begin{aligned} q_{ij} = \lambda(1-\alpha)\alpha^{k-1}\mathbf{1}(j = i+k) \\ +\mu[\mathbf{1}(j = i-B, i > K) + \mathbf{1}(j = 0, i \leq K)]. \end{aligned}$$

Show that its stationary distribution is

$$p_i = p_0 \lambda r^{i-1} / (\mu \sum_{k=0}^{K-1} r^k), \qquad i \geq 1,$$

where r is the unique solution (which you need not prove) in $(0,1)$ of

$$\mu r^K + \mu(1-\alpha) \sum_{k=0}^{K-1} r^k = \lambda + \mu.$$

Exercise 33. *Central Limit Theorem.* Let $X(t)$ denote an ergodic CTMC with limiting distribution $p_i = \pi_i/q_i / \sum_{j\in S} \pi_j/q_j$, for $i \in S$, where π is the stationary distribution for X_n, which is ergodic. Consider the functional $Z(t) = \int_0^t f(X(s))\, ds$, where $f(i)$ denotes a value per unit time when $X(t)$ is in state i. Theorem 42 showed that $t^{-1}Z(t) \to a = \sum_{i\in S} p_i f(i)$, a.s., provided the sum is absolutely convergent. For simplicity, fix $i \in S$ and assume $X(0) = i$. Specify conditions, based on Theorem 65 in Chapter 2, under which

$$(Z(t) - at)/t^{1/2} \xrightarrow{d} N(0, \sigma^2), \qquad \text{as } t \to \infty.$$

Give an expression for σ^2 using ideas in Example 68 in Chapter 3 and

$$Z(T_1) - aT_1 = \sum_{k=1}^{\nu_i} [f(X_k)Y_k - aY_k],$$

where $\nu_i = \min\{n \geq 1 : X_n = i\}$ and Y_n is the sojourn time of $X(t)$ in X_n.

Exercise 34. *Reversible Multiple Instantaneous Jumps.* Let $\tilde{X}(t)$ be a CTMC on S with transition rates $\tilde{q}_{ij}$. Define a CTMC $X(t)$ with transition rates $q_{ij} = \sum_{m=1}^{n} \tilde{q}_{ij}^m$, where $\tilde{Q}^m = \{\tilde{q}_{ij}^m\}$ is the mth product of the matrix $\tilde{Q} = \{\tilde{q}_{ij}\}$ with each $q_{ii} = 0$ and n is fixed. This process is a variation of $\tilde{X}(t)$ in which each transition consists of up to n transitions of $\tilde{X}(t)$ occurring simultaneously. The compound rate q_{ij}^m represents a "macro" transition rate for m "instantaneous jumps" of $\tilde{X}$. Show that if $\tilde{X}(t)$ is reversible with respect to γ, then $X(t)$ is reversible with respect to γ.

Exercise 35. *Networks with Variable Waiting Spaces.* Consider an m-node open network process $X_t = (X_t^1, \ldots, X_t^m)$ that represents the numbers of items at the nodes at time t. Suppose the waiting spaces at the nodes vary such that $Y_t = (Y_t^1, \ldots, Y_t^m)$ is the maximum numbers of items allowed at the nodes at time t. Suppose $\{(X_t, Y_t) : \ t \geq 0\}$ is an irreducible CTMC on $S = \{(x, y) \in S_X \times S_Y : \ x \leq y\}$, where $S_X = \{x : |x| < \infty\} = S_Y$. Assume that its transition rates are

$$q((x,y),(x',y')) = \begin{cases} \lambda_{jk}\phi_j(x_j) & \text{if } x' = T_{jk}x,\ y' = y \\ & \text{and } x_k < y_k \text{ for some } j,k \in M \\ q_Y(y,y') & \text{if } x' = x \text{ and } y' \geq x'. \\ 0 & \text{otherwise.} \end{cases}$$

The X is an open Jackson process whose node populations are restricted by the process Y with transition rates q_Y. Assume that the routing rates λ_{jk} are reversible with respect to w_j, and that q_Y is reversible with respect to π_Y. Show that the process (X, Y) is reversible with respect to $\pi(x, y) = \pi_Y(y) \prod_{j=1}^m w_j \prod_{n=1}^{x_j} \phi_(n)^{-1}$.

Exercise 36. Suppose the transition rates $\tilde{q}(x, y)$ represent a CTMC on S that is reversible with respect to $\tilde{\gamma}$. Assume the process is subject to the constraint that a transition is possible if and only if $h(x, y) \leq b$, for some $h : S^2 \to \mathbb{R}_+$ and $b > 0$, and denote the resulting process by $X(t)$. Show that this is a reversible CTMC and specify an invariant measure for it.

Exercise 37. Prove that a Jackson process is irreducible if and only if its routing process is irreducible. Recall that λ_{ij} is irreducible if and only if, for any fixed $i \neq j$ in M, there are $i_1, \ldots, i_\ell$ in M such that $\lambda_{ii_1}\lambda_{i_1i_2}\cdots\lambda_{i_\ell j} > 0$.

Exercise 38. *Parameters for Closed Jackson Process.* The convolution $f \star g$ of two real-valued functions g and h on $\mathbb{Z}_+$ is defined by

$$g \star h(n) = \sum_{i=0}^{n} g(i)h(n-i), \quad n \geq 0.$$

For a sequence of such functions $g_1, g_2, \ldots$, show by induction that

$$g_1 \star \cdots \star g_m(n) = \sum_{x:|x|=n} \prod_{i=1}^{m} g_i(x_i), \quad n \geq 0,\ m \geq 1.$$

Show that the normalizing constant c in Theorem 74 for a closed Jackson network process has the representation $c^{-1} = f_1 \star \cdots \star f_m(\nu)$.

Associated with a sector of nodes J, define $x_J = \sum_{j\in J} x_j$, and let f_J denote the convolution of the functions $\{f_j,\ j \in J\}$. Assume the network process $X(t)$ is stationary and let $X_J(t) = \sum_{j\in J} X_j(t)$. Show that, for any disjoint sectors $J_1, \ldots, J_\ell$ whose union is M, the joint equilibrium distribution of $n_1, \ldots, n_\ell$ items in these sectors is

$$P\{X_{J_1}(0) = n_1, \ldots, X_{J_\ell}(0) = n_\ell\} = c\prod_{i=1}^{\ell} f_{J_i}(n_i), \quad n_1 + \cdots + n_\ell = \nu.$$

From these distributions, one can obtain means, variances, covariances, and other items of interest such as expected costs for the process. In particular, show that the mean number of items in a sector J in equilibrium is

$$E[X_J(0)] = c\sum_{n=1}^{\infty} n f_J(n) f_{J^c}(\nu - n).$$

Exercise 39. *Jackson Networks with Feedbacks at Nodes*. Consider a Jackson process under the usual assumption that, whenever it is in state x, the time to the next departure from node i is exponentially distributed with rate $\phi_i(x_i)$, but the routing is a little different. Assume that an item departing from node i enters node j with probability $\bar{p}_{ij}$, independently of everything else, where the probability $\bar{p}_{ii}$ of a feedback may be positive. Justify that the

resulting process $X(t)$ is a CTMC and specify its transition rates $q(x, T_{ij}x)$ and exponential sojourn rate $q(x)$ in state x.

Exercise 40. *Busing System.* Items arrive to a waiting station at times that form a Poisson process with rate λ. "Buses" arrive to the station at times that form a Poisson process with rate μ to take items immediately from the system. If a bus arrives and finds the system empty, it departs immediately. Busing is common in computer systems and material handling systems. Assume that the number of items each bus can take is a random variable with the geometric distribution $p^{n-1}(1-p)$, $n \geq 1$. Also, when there are no items in the queue and an item arrives, then with probability p there is a bus available to take the arrival without delay.

Show that if a bus arrives and finds i items waiting, then the actual number Y that departs in a batch has the truncated geometric distribution

$$P\{Y = n\} = p^{n-1}(1-p)\mathbf{1}(n < i) + p^{n-1}\mathbf{1}(n = i).$$

Let $X(t)$ denote the number of items in the system at time t. Show that it is a CTMC with transition rates

$$q_{i,i+1} = \lambda(1-p)\mathbf{1}(i = 0) + \lambda\mathbf{1}(i \geq 1),$$
$$q_{i,i-n} = \mu p^{n-1}(1-p)\mathbf{1}(1 \leq n \leq i-1) + \mu p^{i-1}\mathbf{1}(n = i).$$

Show that $X(t)$ is ergodic if and only if $\lambda < \mu + p\lambda$, and in this case, its stationary distribution is

$$p_i = p_0(1-p)\lambda^i/(\mu + p\lambda)^i, \quad i \geq 1.$$

Exercise 41. *Star-Shaped Network.* Let $X(t)$ denote a closed Jackson process with ν items for the *star-shaped* or *central-processor* network shown in Figure 4.5. Node 1 is the center node and nodes $2, \ldots, m$ are points of the star such that the routing rates λ_{1j} and λ_{i1} are positive. All the other routing rates are 0. Suppose node 1 operates like an $M/M/\infty$ node with service rate μ_1, and the other nodes operate like $M/M/1$ nodes with service rate μ_i for node i. Find a solution w_i of the traffic equations, and then find the stationary distribution for $X(t)$. Is this process reversible? Find the equilibrium probability that there are no items at node 1.

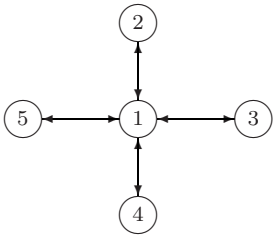

Fig. 4.5 Star-Shaped Network

Exercise 42. Prove that a CTMC $X(t)$ is reversible with respect to p if and only if

$$p_i p_{ij}(t) = p_j p_{ji}(t), \qquad i, j \in S. \tag{4.68}$$

Hint: Consider the sum $p_{ij}(t) = \sum_{n=0}^{\infty} p_{ij}(t,n)$, where $p_{ij}(t,n)$ is the probability of $X(t)$ starting in state i and being in j at time t at the n-th state visited by the process. To prove (4.68), it suffices to show (by induction) that

$$p_i p_{ij}(t,n) = p_j p_{ji}(t,n), \qquad i, j \in S,\ n \geq 0.$$

To prove the converse use $p_{ij}(t) = q_{ij}t + o(t)$, as $t \downarrow 0$ from Theorem 18.

Exercise 43. *Sojourn Times in an $M/M/1$ System.* In the context of Example 101, show that the sojourn time $\hat{W}$ for an arriving item in equilibrium and its waiting time W prior to service satisfy

$$P_{\mathcal{T}}\{\hat{W} \leq t\} = P_{\mathcal{T}}\{W \leq t | W > 0\} = 1 - e^{-(\mu-\lambda)t}.$$

Exercise 44. *Finite-Capacity Kelly Network.* Consider the network process in Theorem 81 with the modification that arrivals from outside are dependent on the network such that $q(x, x + e_{r1}) = \lambda(r)\phi_{r0}(|x|)$. This would allow for a finite capacity network by assuming $\phi_{r0}(n) = 0$ for $n = \nu_r$. For this more general arrival rate, show that the invariant measure as in Theorem 81 would be

$$p(x) = \prod_{rs \in M} \lambda(r)^{x_r} f_{rs}(x) \prod_{k=1}^{|x|} \phi_{r0}(k-1), \quad x \in S.$$

Exercise 45. *Multiclass Jackson Network Process.* Suppose $X(t)$ is the multiclass network process $X(t)$ in Theorem 82 with service intensities $\phi_{\alpha i}(x) = \mu_{\alpha i}(x_{\alpha i})$, $i \neq 0$, that do not depend on the quantity of items at the nodes. Find an invariant measure for it. Next, consider the modification that α items arrive to the network from outside with rate

$$\phi_{\alpha 0}(x) = g_0(|x|)h_{\alpha 0}(|x_\alpha|),$$

where $|x_\alpha| = \sum_{i=1}^{m} x_{\alpha i}$ is the number of α-items in the network. Show that an invariant measure for $X(t)$ has the form $p(x) = \prod_{\alpha i \in M} w_{\alpha i}^{x_{\alpha i}} f_{\alpha i}(x)$, with

$$f_{\alpha 0}(x) = \prod_{k=0}^{|x|-1} g_0(k) \prod_{k'=0}^{|x_\alpha|-1} h_{\alpha 0}(k').$$

Exercise 46. *Throughput Rates.* Suppose $X(t)$ is the multiclass network process $X(t)$ in Theorem 82 and assume it is ergodic. Find a simple expression for the throughput (or average number of items that jump) from αi to βj

$$\rho_{\alpha i,\beta j} = \lim_{t\to\infty} t^{-1} \sum_{s\le t} \mathbf{1}(X(s) = X(s-) - e_{\alpha i} + e_{\beta j}).$$

Next, assume the network is closed with ν items in it. Specify the stationary distribution for $X(t)$, and derive the throughput formula

$$\rho_{\alpha i,\beta j} = c_\nu c_{\nu-1}^{-1} w_{\alpha i} \lambda_{\alpha i,\beta j},$$

where c_ν is the normalization constant for the network with ν items in it.

Exercise 47. *Continuation.* Suppose $X(t)$ is the open multiclass network process $X(t)$ in the preceding exercise. Explain how the solutions $w_{\alpha i}$ to the traffic equations and throughput $\rho_{\alpha i,\beta j}$ is changed or simplifies under the following two scenarios.
(a) Each item carries a class label that does not change: $\lambda_{\alpha i,\beta j} = 0$ if $\alpha \neq \beta$.
(b) The class changes are independent of the routing in that $\lambda_{\alpha i,\beta j} = \tilde{\lambda}_{\alpha\beta}\bar{\lambda}_{ij}$, where $\tilde{\lambda}_{\alpha\beta}$ and $\bar{\lambda}_{ij}$ are irreducible transition rates for class changes and node changes, respectively.

Exercise 48. *Busy Period in an $M/G/1$ System.* Suppose $Q(t)$ is an $M/G/1$ queueing process with Poisson arrival times $0 < \tau_1 < \tau_2 < \cdots$, $Q(0) = 0$ and $\rho < 1$. Consider $T = \inf\{t > \tau_1 : Q(t) = 0\}$, which is the time at which the system first becomes empty. Now $T = \tau_1 + Y$, where Y is the duration of the *busy period* for the server. Find $E[T]$ and show that $E[Y] = \rho/\lambda(1-\rho)$.

Exercise 49. *Scheduling Patients.* Prior to having an operation at a hospital, a patient is given a set of tests depending on the type of procedure (e.g., EKG, sonogram, blood work). This has to be done at least one week before the operation. Patients used to come in at their convenience, usually near noon or late afternoon. To avoid congestion, the hospital required that patients make an appointment for the test. The number of tests was such that patients were scheduled to arrive every u minutes. (The following model is similar to an actual study by Georgia Tech students for a hospital in Atlanta.)

As an idealized model, assume that patients do indeed arrive each u minutes, and the durations of the tests are independent exponentially distributed with rate $\mu < 1/u$. The tests are done one at a time and a patient arriving when another one is being tested waits in a queue. Let $Q(t)$ denote the number of patients in the system (waiting or being tested) at time t, and let W_n denote the length of time the nth patient waits in the queue before being tested. Find the distributions of the equilibrium queue length $\tilde{Q}$ and waiting time $\tilde{W}$. Determine what the time between tests u should be under the following criteria.
(a) Find the shortest u such that $P\{\tilde{W} \le w\} = .90$ for fixed w.
(b) Find the shortest u such that $E[\tilde{W}] \le w$ and $P\{\tilde{Q} > m\} \le .10$ for fixed w and m.

Exercise 50. *Fork-Join Processing System.* Consider an m-node fork-join network that processes jobs as follows. Jobs arrive every u time items (u is a constant) and each job splits into m tasks, which are simultaneously assigned to the m nodes for processing. The nodes operate independently, and each node processes jobs like a single-server $G/M/1$ system with independent exponential service times with rate μ. When all the m tasks for a job are finished, the job is complete and exits the system. The network is shown in Figure 1.4 in Chapter 1, where the operating rules were different. Assume the system is empty at time 0. Let $X(t) = (X_1(t), \ldots, X_m(t))$ denote the numbers of tasks at the m nodes at time t., and find its limiting distribution.

Let W_n^i denote the time to complete the task at node i for the nth job. Then the sojourn time in the system for the nth job (i.e., the time to process the job) is $W_n = \max\{W_n^1, \ldots, W_n^m\}$. Show that $W_n \xrightarrow{d} W$ as $n \to \infty$, and determine the distribution of W (which is a product of exponential distributions). Find the distribution of W when G is an exponential distribution.

Exercise 51. *Extreme-value Process.* Claims arrive at an insurance company at times T_n that form a Poisson process N with rate λ. The size Y_n of the nth claim that arrives at time T_n has an exponential distribution with rate μ and the claim sizes are independent of their arrival times. The maximum claim up to time t is $X(t) = \max_{k \le N(t)} Y_k$. Justify that $X(t)$ is a CTMC and specify its defining parameters. Show that $X(t) \to \infty$ a.s. as $t \to \infty$, and that

$$\mu X(t) - \log(\lambda t) \xrightarrow{d} Z,$$

where $P\{Z \le x\} = \exp\{-e^{-x}\}$, which is the Gumbel distribution. Evaluate the distribution by conditioning on $N(t)$ and using the exponential property that if $na_n \to a$, then $(1 - a_n)^n \to e^{-a}$ as $n \to \infty$. As an intermediate step, justify that $\mu X(T_n) - \log(\lambda n) \xrightarrow{d} Z$ by evaluating the distribution of $\mu X(T_n)$.

Exercise 52. *Batch-Service System.* Consider a batch-service system as in Section 2.12 that processes items as follows. Items arrive to the station according to a Poisson process with rate λ and they enter a queue where they wait to be served. Items are processed in batches, and the number of items in a batch can be any number less than or equal to K (the service capacity). The service times of the batches are independent, exponentially distributed with rate μ independently of everything else. Only one batch can be served at a time and, during a service, additional arrivals join the queue. Batches are served when and only when the queue length is equal or greater than m (a control limit). In particular, if at the end of a service there are $i \ge m$ items in the queue, then a batch of $i \wedge K$ items is served; and when the queue length is $m - 1$ and an arrival occurs, then a batch of size m is served. Let X_n denote the queue length at the end of the nth service.

Show that the probability of n arrivals during a service is qp^n, where $p = \lambda/(\lambda + \mu)$ and $q = 1 - p$. Justify that X_n is a Markov chain with

transition probabilities

$$p_{ij} = \begin{cases} qp^j & \text{if } i < K \\ qp^{j+k-i}(1-p) & \text{if } K \le i \le j-K \end{cases}$$

and $p_{ij} = 0$ otherwise. Why don't these probabilities depend on m? Assuming $\lambda < K\mu$, prove that X_n is ergodic with stationary distribution $\pi_i = (10r)r^i$, $i \ge 0$, where r is the unique solution of $qr^{K+1} - r + p = 0$.

Exercise 53. *Continuation.* In the preceding batch-service model, let T_n denote the time of the nth service completion, where $T_0 = 0$. Show that

$$E[T_1|X_0 = i] = (m-i)\lambda^{-1}\mathbf{1}(i \le m) + \mu^{-1}.$$

Assume that $C + ci$ is the cost for serving a batch of size i, and hi is the cost per unit time for holding i items in the queue. Show that the expected service plus holding cost in a time interval $(T_n, T_{n+1}]$ given $X_n = i$ is

$$f(i) = C + ci + hi\mu^{-1} + h\lambda\mu^{-2}, \quad i > m,$$

and for $i \le m$,

$$f(i) = \frac{1}{2}hm(m-1)\lambda^{-1} - \frac{1}{2}hi(i-1)\lambda^{-1} + C + cm + hm\mu^{-1} + h\lambda\mu^{-2}.$$

Justify that the average cost for the system is

$$\sum_{i=0}^{\infty} f(i)\pi_i / \sum_{i=0}^{\infty} E[T_1|X_0 = i]\pi_i.$$

This cost is a tractable function $\phi(m)$ of the control level m. This cost is minimized at the smallest integer $m \le K$ such that $D_m \ge 0$, see [33], where

$$D_m = m\Big[\frac{1}{2}(m+1) + \lambda/\mu - c\Big] - c^2 r^m + c(c - \lambda/\mu) - C\lambda/h.$$

Exercise 54. *Markov/Poisson Particle System.* Consider a particle system in a countable space S similar to the one in Section 3.11 with the following modifications. Each particle moves independently in continuous time according to an ergodic CTMC with transition probabilities $p_{ij}(t)$ and stationary distribution p_i, $i \in S$. That is, $p_{ij}(t)$ is the probability that a particle starting in state i is in state j at time t. Assume the system is empty at time 0 and that particles enter the system according to a space-time Poisson process M on $\mathbb{R}_+ \times S$, where $M((0,t] \times B)$ is the number of arrivals in $(0,t]$ that enter $B \subseteq S$, and $E[M((0,t] \times \{i\})] = \lambda t p_i$. Let $Q_i(t)$ denote the quantity of particles in state i at time t. Show that

$$(Q_i(t) : i \in S) \xrightarrow{d} (Q_i : i \in S), \qquad \text{as } t \to \infty,$$

where Q_i are independent Poisson random variables with $E[Q_i] = \lambda p_i$.

Exercise 55. *Continuation.* In the setting of the preceding exercise, suppose at time 0 the number of particles in the system is a point process with intensity μ that is independent of the space-time arrival process M of other particles and all the particles move independently as above. The quantity of particles in state i at time t is $X_i(t) = Q_i^0(t) + Q_i(t)$, where $Q_i^0(t)$ denotes the quantity of particles in i at time t that were in the system at time 0. Show that $E[Q_i^0(t)] = \sum_{j\in S} p_{ji}(t)\mu(i)$, and find $\alpha_i = \lim_{t\to\infty} E[X_i(t)]$. Prove

$$Q_i^0(t) \xrightarrow{d} Q_i^0, \qquad \text{as } t \to \infty, \text{ for } i \in S,$$

where Q_i^0 are independent Poisson random variables with $E[Q_i] = p_i$. Show that $\lim_{t\to\infty} P\{X_i(t) = n\} = e^{-\alpha_i}(\alpha_i)^n/n!$.

Exercise 56. Let $X(t)$ denote the ergodic Markov-renewal process as in Theorem 124. Arguing as in Proposition 41, show that

$$E_i\Big[\int_0^{\tau_i} \mathbf{1}(X(t) = j)\, dt\Big] = \pi_i^{-1}\pi_j \sum_{\ell\in S} p_{j\ell}\mu_{j\ell},$$

$$E_i[\tau_i] = \pi_i^{-1} \sum_{j\in S} \pi_j \sum_{\ell\in S} p_{j\ell}\mu_{j\ell}$$

$$E_i\Big[\int_0^{\tau_i} f(X(t))\, dt\Big] = \pi_i^{-1} \sum_{j\in S} \pi_j f(j) \sum_{\ell\in S} p_{j\ell}\mu_{j\ell}.$$

Use these formulas to prove Theorem 124.

Chapter 5
Brownian Motion

Brownian motion processes originated with the study by the botanist Brown in 1827 of the movements of particles suspended in water. As a particle is occasionally hit by the surrounding water molecules, it moves continuously in three dimensions. Assuming the infinitesimal displacements of the particle are independent and identically distributed, the central limit theorem would imply that the size of a typical displacement (being the sum of many small ones) is normally distributed. Then the continuous trajectory of the particle in $\mathbb{R}^3$ would have increments that are stationary, independent and normally distributed. These are the defining properties of Brownian motion. This diffusion phenomenon, commonly encountered in other contexts as well, gave rise to the theory of Brownian motion and more general diffusion processes.

Brownian motion is one of the most prominent stochastic processes. Its importance is due in part to the central limit phenomenon that sums of random variables such as random walks, considered as processes in time, converge to a Brownian motion or to functions of it. Moreover, Brownian motion plays a key role in stochastic calculus involving integration with respect to Brownian motion and semimartingales. This calculus is used to study dynamical systems modeled by stochastic differential equations. For instance, in the area of stochastic finance, stochastic differential equations are the basis for pricing of options by Black-Scholes and related models. Brownian motion is an important example of a diffusion process and it is a Gaussian process as well. Several variations of Brownian motion arise in specific applications, such as Brownian bridge in statistical hypothesis testing. In operations research, the major applications of Brownian motion have been in approximations for queueing systems, and there have also been applications in various areas such as financial models and supply chains.

This chapter begins by introducing a Brownian motion as a Markov process that satisfies the strong Markov property, and then characterizes a Brownian motion as a Gaussian process. The second part of the chapter is a study of hitting times of Brownian motion and its cumulative maximum process. This

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

includes a reflection principle for Brownian sample paths, and an introduction to martingales and the optional stopping theorem for them.

The next major results are limit theorems: a strong law of large numbers for Brownian motion and its maximum process, a law of the iterated logarithm for Brownian motion, and Donsker's functional limit theorem showing that Brownian motion is an approximation to random walks. Applications of Donsker's theorem yield similar Brownian approximations for Markov chains, renewal and regenerative-increment processes, and $G/G/1$ queueing systems.

Other topics include peculiarities of Brownian sample paths, geometric Brownian motion, Brownian bridge processes, multidimensional Brownian motion, Brownian/Poisson particle process, and Brownian motion in a random environment.

5.1 Definition and Strong Markov Property

Recall that a random walk in discrete time on the integers is a Markov chain with stationary independent increments. An analogous process in continuous time on $\mathbb{R}$ is a Brownian motion. This section introduces Brownian motion as a real-valued Markov process on the nonnegative time axis. Its distinguishing features are that it has stationary, independent, normally-distributed increments and continuous sample paths. It also satisfies the strong Markov property.

We begin by describing a "standard" Brownian motion, which is also called a *Wiener process*.

Definition 1. A real-valued stochastic process $B = \{B(t) : t \in \mathbb{R}_+\}$ is a *Brownian motion* if it satisfies the following properties.
(i) $B(0) = 0$ a.s.
(ii) B has independent increments and, for $s < t$, the increment $B(t) - B(s)$ has a normal distribution with mean 0 and variance $t - s$.
(iii) The paths of B are continuous a.s.

Property (ii) says that a Brownian motion B has stationary, independent increments. From this one can show that B is a Markov process. Consequently, a Brownian motion is a diffusion process — a Markov process with continuous sample paths. The next section establishes the existence of Brownian motion as a special type of Gaussian process. An introduction to Brownian motion in $\mathbb{R}^d$ is in Section 5.14.

Because the increments of a Brownian motion B are stationary, independent and normally distributed, its finite-dimensional distributions are tractable. The normal density of $B(t)$ with mean 0 and variance t is

$$f_{B(t)}(x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}.$$

Denoting this density by $\varphi(x;t)$, it follows by induction and properties (i) and (ii) that, for $0 = t_0 < t_1 < \cdots < t_n$ and $x_0 = 0$, the joint density of $B(t_1), \ldots, B(t_n)$ is

$$f_{B(t_1),\ldots,B(t_n)}(x_1, \ldots, x_n) = \prod_{m=1}^{n} \varphi(x_m - x_{m-1}; \sqrt{t_m - t_{m-1}}). \tag{5.1}$$

Another nice feature is that the covariance between $B(s)$ and $B(t)$ is

$$E[B(s)B(t)] = s \wedge t. \tag{5.2}$$

This follows since, for $s < t$,

$$\begin{aligned} \mathrm{Cov}(B(s), B(t)) = E[B(s)B(t)] &= E[B(s)[(B(t) - B(s)) + B(s)]] \\ &= E[B(s)^2] = s. \end{aligned}$$

Several elementary functions of a Brownian motion are also Brownian motions; see Exercise 1. Here is an obvious example.

Example 2. Symmetry Property. The process $-B(t)$, which is B reflected about 0, is a Brownian motion (i.e., $-B \stackrel{d}{=} B$).

As a generalization of a standard Brownian motion B, consider the process

$$X(t) = x + \mu t + \sigma B(t), \qquad t \geq 0.$$

Any process equal in distribution to X is a *Brownian motion with drift*: x is its initial value, μ is its *drift* coefficient, and $\sigma > 0$ is its *variation*. Many properties of a Brownian motion with drift readily follow from properties of a standard Brownian motion. For instance, X has stationary, independent increments and $X(t+s) - X(s)$ is normally distributed with mean μt and variance $\sigma^2 t$. Drift and variability parameters may be useful in Brownian models for representing certain trends and volatilities.

Now, let us see how Brownian motions are related to diffusion processes. Generally speaking, a diffusion process is a Markov process with continuous paths. Most diffusion processes in applications, however, have the following form. Suppose that $\{X(t) : t \geq 0\}$ is a real-valued Markov process with continuous paths a.s. that satisfies the following properties: For each $x \in \mathbb{R}$, $t \geq 0$, and $\varepsilon > 0$,

$$\begin{aligned} \lim_{h \downarrow 0} h^{-1} P\{|X(t+h) - X(t)| > \epsilon | X(t) = x\} &= 0, \\ \lim_{h \downarrow 0} h^{-1} E[X(t+h) - X(t) | X(t) = x] &= \mu(x,t), \\ \lim_{h \downarrow 0} h^{-1} E[(X(t+h) - X(t))^2 | X(t) = x] &= \sigma(x,t), \end{aligned}$$

where μ and σ are functions on $\mathbb{R} \times \mathbb{R}_+$. The X is a *diffusion process* on $\mathbb{R}$ with drift parameter $\mu(x,t)$ and diffusion parameter $\sigma(x,t)$.

As a prime example, a Brownian motion with drift $X(t) = \mu t + \sigma B(t)$, is a diffusion process whose drift and diffusion parameters μ and σ are independent of x and t. Many functions of Brownian motions are also diffusions (e.g., the Ornstein-Uhlenbeck and Bessel Processes in Examples 8 and 64).

We end this introductory section with the strong Markov property for Brownian motion. Suppose that B is a Brownian motion on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_t^B \subseteq \mathcal{F}$ be the σ-field generated by $\{B(s) : s \in [0,t]\}$, and assume $\mathcal{F}_0^B$ includes all sets of P-probability 0 to make it complete. A *stopping time* of the *filtration* $\mathcal{F}_t^B$ is a random time τ, possibly infinite, such that $\{\tau \le t\} \in \mathcal{F}_t^B$, $t \in \mathbb{R}_+$. The σ-field of events up to time τ is

$$\mathcal{F}_\tau^B = \{A \in \mathcal{F} : A \cap \{\tau \le t\} \in \mathcal{F}_t,\ t \in \mathbb{R}_+\}.$$

If τ_1 and τ_2 are two $\mathcal{F}_t^B$-stopping times and $\tau_1 \le \tau_2$, then $\mathcal{F}_{\tau_1}^B \subseteq \mathcal{F}_{\tau_2}^B$.

Theorem 3. *If τ is an a.s. finite stopping time for a Brownian motion B, then the process $B'(t) = B(\tau + t) - B(\tau)$, $t \in \mathbb{R}_+$, is a Brownian motion independent of $\mathcal{F}_\tau^B$.*

Proof. We will prove this only for a stopping time τ that is a.s. bounded ($\tau \le u$ a.s. for some $u > 0$). Clearly B' has continuous sample paths a.s. It remains to show that the increments of B' are independent and independent of $\mathcal{F}_\tau^B$, and $B'(s+t) - B'(s)$ is normally distributed with mean 0 and variance t. These properties will follow upon showing that, for any $0 \le t_0 < \cdots < t_n$, and $u_1, \ldots, u_n$ in $\mathbb{R}_+$,

$$E[e^{S_n} | \mathcal{F}_\tau^B] = e^{\frac{1}{2}\sum_{i=1}^n u_i^2 (t_i - t_{i-1})} \quad \text{a.s.}, \tag{5.3}$$

where $S_n = \sum_{i=1}^n u_i [B'(t_i) - B'(t_{i-1})]$.

The proof of (5.3) will be by induction. First note that

$$E[e^{S_{n+1}} | \mathcal{F}_\tau^B] = E\Big[e^{S_n} E[e^{S_{n+1} - S_n} | \mathcal{F}_{\tau + t_n}^B] \Big| \mathcal{F}_\tau^B\Big]. \tag{5.4}$$

Now, since $\tau + t_n$ is a bounded stopping time, using Example 26 below,

$$\begin{aligned} E[e^{S_{n+1} - S_n} | \mathcal{F}_{\tau + t_n}^B] &= E[e^{u_{n+1}[B(\tau + t_{n+1}) - B(\tau + t_n)]} | \mathcal{F}_{\tau + t_n}^B] \\ &= e^{\frac{1}{2} u_{n+1}^2 (t_{n+1} - t_n)}. \end{aligned}$$

This expression with $n = 0$ and $S_0 = 0$ proves (5.3) for $n = 1$. Next assuming (5.3) is true for some n, then using the last display and (5.3) in (5.4) yields (5.3) for $n + 1$.

5.2 Brownian Motion as a Gaussian Process

This section shows that Brownian motion is a special type of Gaussian Process. Included is a proof of the existence of Gaussian processes, which leads to the existence of Brownian motion.

We begin with a discussion of multivariate normal distributions. Suppose that $X_1, \ldots, X_n$ are normally distributed (not necessarily independent) random variables with means $m_1, \ldots, m_n$. Then clearly, for $u_1, \ldots, u_n \in \mathbb{R}$,

$$E\Big[\sum_{i=1}^{n} u_i X_i\Big] = \sum_i u_i m_i, \qquad \mathrm{Var}\Big[\sum_{i=1}^{n} u_i X_i\Big] = \sum_i \sum_j u_i u_j c_{ij}, \tag{5.5}$$

where $c_{ij} = \mathrm{Cov}(X_i, X_j)$. The vector $(X_1, \ldots, X_n)$ is said to have a *multivariate normal* (or Gaussian) distribution if $\sum_{i=1}^{n} u_i X_i$ has a normal distribution for any $u_1, \ldots, u_n$ in $\mathbb{R}$. In light of (5.5), the $(X_1, \ldots, X_n)$ has a multivariate normal distribution if and only if its moment generating function has the form

$$E\left[e^{\sum_{i=1}^{n} u_i X_i}\right] = \exp\Big\{\sum_i u_i m_i + \frac{1}{2}\sum_i \sum_j u_i u_j c_{ij}\Big\}, \quad u_i \geq 0. \tag{5.6}$$

The vector (or distribution) associated with the moment generating function (5.6) is called *nondegenerate* if the $n \times n$ matrix $C = \{c_{ij}\}$ has rank n. In this case, the joint density of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\Big\{-\frac{1}{2}\sum_i \sum_j \hat{c}_{ij}(x_i - m_i)(x_j - m_j)\Big\}, \tag{5.7}$$

where $\{\hat{c}_{ij}\}$ is the inverse of C and $|C|$ is its determinant.

It turns out that any multivariate normal vector can be represented by a nondegenerate one as follows. When C does not have rank n, it follows by a property of symmetric matrices that there exists a $k \times n$ matrix A with transpose A^t, where $k \leq n$, such that $C = A^t A$. Let X denote the multivariate normal vector as a $1 \times n$ matrix with mean vector m. Suppose that Y is a $1 \times k$ nondegenerate multivariate vector of i.i.d. random variables $Y_1, \ldots, Y_k$ that are normally distributed with mean 0 and variance 1. Then the multivariate normal vector, has the representation

$$X \stackrel{d}{=} m + YA. \tag{5.8}$$

This equality in distribution follows because the moment generating function of $m + YA$ is equal to (5.6). Indeed,

$$E\left[\exp\left\{\sum_{i=1}^{n} u_i m_i + \sum_{i=1}^{n} u_i (\sum_{j=1}^{k} a_{ji}) Y_j\right\}\right] = \exp\left\{\sum_i u_i m_i + \frac{v}{2}\right\},$$

where interchanging the summations and using $C = A^t A$,

$$\begin{aligned} v = \mathrm{Var}\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n} u_i a_{ji}\right) Y_j\right] &= \sum_{j=1}^{k}\left(\sum_{i=1}^{n} u_i a_{ji}\right)^2 \\ &= \sum_{j=1}^{k}\left(\sum_{i=1}^{n} u_i a_{ij}^t\right)\left(\sum_{\ell=1}^{n} u_\ell a_{j\ell}\right) = \sum_i \sum_j u_i u_j c_{ij}. \end{aligned}$$

A major characteristic of a Brownian motion is that its finite-dimensional distributions are multivariate normal. Other stochastic processes with this property are as follows.

Definition 4. A stochastic process $X = \{X(t) : t \in \mathbb{R}_+\}$ is a *Gaussian process* if $(X(t_1), \ldots, X(t_n))$ has a multivariate normal distribution for any $t_1, \ldots, t_n$ in $\mathbb{R}_+$. Discrete-time Gaussian processes are defined similarly.

A process X is Gaussian, of course, if and only if $\sum_{i=1}^{n} u_i X(t_i)$ has a normal distribution for any $t_1, \ldots, t_n$ in $\mathbb{R}_+$ and $u_1, \ldots, u_n$ in $\mathbb{R}$. A Gaussian process X is called *nondegenerate* if its covariance matrix $c_{ij} = \mathrm{Cov}(X(t_i), X(t_j))$ has rank n, for any $t_1, \ldots, t_n$ in $\mathbb{R}_+$. In that case, $\big(X(t_1), \ldots, X(t_n)\big)$ has a multivariate normal density as in (5.7).

The next result establishes the existence of Gaussian processes. It also shows that the distribution of a Gaussian process is determined by its mean and covariance functions. So two Gaussian processes are equal in distribution if and only if their mean and covariance functions are equal. Let $c(s,t)$ be a real-valued function on $\mathbb{R}_+^2$ that satisfies the following properties:

$$c(s,t) = c(t,s), \qquad s,t, \in \mathbb{R}_+. \quad \text{(Symmetric)}$$

For any finite set $I \subset \mathbb{R}_+$ and $u_t \in \mathbb{R}$,

$$\sum_{t\in I}\sum_{s\in I} u_s u_t c(s,t) \geq 0. \quad \text{(Nonnegative-definite)}$$

Theorem 5. *For any real-valued function $m(t)$ and the function $c(s,t)$ described above, there exists a Gaussian process $\{X(t) : t \in \mathbb{R}_+\}$ defined on a probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^{\mathbb{R}_+}, \mathcal{B}^{\mathbb{R}_+}, P)$, with $E[X(t)] = m(t)$ and*

$$\mathrm{Cov}(X(s), X(t)) = c(s,t), \qquad s,t, \in \mathbb{R}_+.$$

Furthermore, the distribution of this process is determined by the functions $m(t)$ and $c(s,t)$.

Proof. We begin by defining finite-dimensional probability measures μ_I for a process on $(\mathbb{R}^{\mathbb{R}_+}, \mathcal{B}^{\mathbb{R}_+}, P)$. For any finite subset I in $\mathbb{R}_+$, let μ_I be the probability measure specified by $\mu_I(\times_{t\in I} A_t)$, $A_t \in \mathcal{B}$, $t \in I$, that has the joint normal moment generating function

$$G(u_I) = \exp\Big\{ \sum_{t\in I} u_t m(t) + \frac{1}{2} \sum_{t\in I} \sum_{s\in I} u_s u_t c(s,t) \Big\}, \quad u_I = (u_t : t \in \mathbb{R}_+).$$

Note that for $I \subseteq J$,

$$G(u_J) = G(u_I), \qquad \text{if } u_t = 0,\ t \in J\backslash I.$$

Consequently, the joint normal distributions μ_I satisfy the consistency condition that, for any $I \subseteq J$ for finite J and $A_t \in \mathcal{B}$, $t \in J$,

$$\mu_J(\times_{t\in J} A_t) = \mu_I(\times_{t\in I} A_t), \qquad \text{if } A_t = \mathbb{R} \text{ for } t \in J\backslash I. \tag{5.9}$$

Then it follows by Kolmogorov's extension theorem (Theorem 5 in the Appendix), that there exists a stochastic process $\{X(t) : t \in \mathbb{R}_+\}$ defined on the probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^{\mathbb{R}_+}, \mathcal{B}^{\mathbb{R}_+}, P)$, whose finite-dimensional probability measures are given by μ_I. Since the μ_I are determined by $m(t)$ and $c(s,t)$, so is the distribution of X. Moreover, from the moment generating function for the μ_I, it follows that

$$E[X(t)] = m(t), \qquad \mathrm{Cov}(X(s), X(t)) = c(s,t).$$

Brownian motion is a quintessential example of a Gaussian process.

Proposition 6. *A Brownian motion with drift $X(t) = \mu t + \sigma B(t)$, $t \geq 0$, is a Gaussian process with continuous sample paths a.s. starting at $X(0) = 0$ and its mean and covariance functions are*

$$E[X(t)] = \mu t, \qquad \mathrm{Cov}(X(s), X(t)) = \sigma^2(s \wedge t), \qquad s, t \in \mathbb{R}_+.$$

Proof. For any $0 = t_0 < t_1 < \cdots < t_n$, letting $Y_i = X(t_i) - X(t_{i-1})$, we have

$$\sum_{i=1}^{n} u_i X(t_i) = \sum_{i=1}^{n} u_i \sum_{k=1}^{i} Y_i = \sum_{k=1}^{n} \left(\sum_{i=k}^{n} u_i \right) Y_k, \qquad u_1, \ldots, u_n \in \mathbb{R}.$$

Now the increments Y_i are independent, normally distributed random variables with mean 0 and variance $\sigma^2(t_i - t_{i-1})$. Then the last double-sum term has a normal distribution, and so $(X(t_1), \ldots, X(t_n))$ has a multivariate normal distribution. Hence X is a Gaussian process, and its mean and variance are clearly as shown.

The preceding characterization is useful for verifying that a process is a Brownian motion, especially when the multivariate normality condition

is easy to verify (as in Exercise 2). There are other interesting Gaussian processes that do not have stationary independent increments; see Example 8 below and Exercise 10.

One approach for establishing the existence of a Brownian motion is to construct it as a Gaussian process as follows.

Theorem 7. *There exists a stochastic process* $\{B(t) : t \geq 0\}$ *defined on a probability space* $(\Omega, \mathcal{F}, P) = (\mathbb{R}^{\mathbb{R}_+}, \mathcal{B}^{\mathbb{R}_+}, P)$ *such that* B *is a Brownian motion.*

Sketch of Proof. Let $\{B(t) : t \geq 0\}$ be a Gaussian process as constructed in the proof of Theorem 5 with the special Brownian functions $m(t) = 0$ and $c(s,t) = s \wedge t$. A major result (whose proof is omitted) says that this process has stationary independent increments, and $B(t) - B(s)$, for $s < t$, is normally distributed with mean 0 and variance $t - s$. A second step is needed, however, to justify that such a process has continuous sample paths.

Since $B(t) - B(s) \stackrel{d}{=} (t-s)^{1/2} B(1)$, for $s < t$, the process satisfies

$$E[|B(t) - B(s)|^a] = (t-s)^{a/2} E[|B(1)|^a] < \infty, \quad a > 0.$$

Using this property, another major result shows that B can be chosen so that its sample paths are continuous, and hence it is a Brownian motion. The results that complete the preceding two steps are proved in [64].

A Brownian motion is an example of a Markov process with continuous paths that is a Gaussian process. Are there Markov processes with continuous paths (i.e., diffusion processes), other than Brownian motions, that are Gaussian? Yes there are — here is an important example.

Example 8. An *Ornstein-Uhlenbeck Process* is a stationary Gaussian process $\{X(t) : t \geq 0\}$ with continuous sample paths whose mean function is 0 and whose covariance function is

$$\mathrm{Cov}(X(s), X(t)) = \frac{\sigma^2}{2\alpha} e^{-\alpha|s-t|}, \qquad s, t \geq 0,$$

where α and σ are positive. This process as proved in [61] is the only stationary Gaussian process with a continuous covariance function that is a Markov process. (Exercise 9 shows that a Gaussian process X is stationary if and only if its mean function is a constant and its covariance function $\mathrm{Cov}(X(s), X(t))$ only depends on $|t - s|$.)

It is interesting that the process X is also a function of a Brownian motion B in that X is equal in distribution to the process

$$Y(t) = \sigma e^{-\alpha t} B(e^{\alpha t}/2\alpha), \qquad t \geq 0.$$

To see this, note that Y has continuous sample paths and clearly it is Gaussian since B is. In addition, $E[Y(t)] = 0$ for each t and, for $s < t$,

$$\begin{aligned}\mathrm{Cov}(Y(s), Y(t)) &= \sigma^2 e^{-\alpha(s+t)} E\Big[B(e^{\alpha s}/2\alpha)B(e^{\alpha t}/2\alpha)\Big] \\ &= \frac{\sigma^2}{2\alpha} e^{-\alpha(t-s)}.\end{aligned}$$

Consequently, Y is a stationary Gaussian process and it has the same mean and covariance functions as X. Hence $Y \stackrel{d}{=} X$.

The Ornstein-Uhlenbeck process X defined above satisfies the stochastic differential equation

$$dX(t) = -\alpha X(t)dt + \sigma dB(t). \tag{5.10}$$

The example above assumes that $X(0)$ has a normal distribution with mean 0 and variance σ^2. The solution of this equation is

$$X(t) = X(0)e^{-\alpha t} + \sigma \int_0^t e^{-\alpha(t-s)} dB(s).$$

The stochastic differential equation and the integral with respect to Brownian motion, which is beyond the scope of this work, is discussed in [61, 64].

Scientists realized that the Brownian motion representation for a particle moving in a medium was an idealized model in that it does not account for friction in the medium. To incorporate friction in the model, Langevin 1908 proposed that the Ornstein-Uhlenbeck process X could represent the velocity of a particle undergoing a Brownian motion subject to friction. He assumed that the rate of change in the velocity satisfies (5.10), where $-\alpha X(t)dt$ models the change due to friction; the friction works in the opposite direction to the velocity and α is the coefficient of friction divided by the mass of the particle. The stochastic process for this model was formalized later by Ornstein and Uhlenbeck 1930 and Doob 1942.

5.3 Maximum Process and Hitting Times

For a Brownian motion B, its cumulative *maximum process* is

$$M(t) = \max_{s \le t} B(s), \qquad t \ge 0.$$

This process is related to the hitting times

$$\tau_a = \inf\{t > 0 : B(t) = a\}, \qquad a \in \mathbb{R}.$$

Namely, for each $a \ge 0$ and t,

$$\{\tau_a \le t\} = \{M(t) \ge a\}. \tag{5.11}$$

In other words, the distribution of the maximum process is determined by that of the hitting times and vice versa. This section presents expressions for these distributions and an important property of the hitting times.

We begin with a preliminary fact.

Remark 9. The hitting time τ_a is an a.s. finite stopping time of B.

The τ_a is a stopping time since B has continuous paths a.s. Its finiteness follows by Theorem 32 below, which is proved by the martingale optional stopping theorem in the next section. The finiteness also follows by the consequence (5.30) of the law of the iterated logarithm in Theorem 38 below.

The first result is a reflection principle that an increment $B(t) - B(\tau)$ after a stopping time τ has the same distribution as the reflected increment $-(B(t) - B(\tau))$. This is basically the symmetry property $B \stackrel{d}{=} -B$ in Example 2 manifested at the stopping time τ. A version of this principle for stochastic processes is in Exercises 20 and 21.

Proposition 10. (Reflection Principle) *If τ is an a.s. finite stopping time of B, then, for any a and t,*

$$P\{B(t) - B(\tau) \le a,\, \tau \le t\} = P\{B(t) - B(\tau) \ge -a,\, \tau \le t\}.$$

Proof. Letting $B'(t) = B(\tau + t) - B(\tau)$, $t \ge 0$, we can write

$$B(t) - B(\tau) = B'(t - \tau), \qquad \text{for } \tau \le t. \tag{5.12}$$

By the strong Markov property in Theorem 3, B' is a Brownian motion independent of $\mathcal{F}_\tau$. Using this and (5.12) along with the symmetry property $B' \stackrel{d}{=} -B'$ and $\{\tau \le t\} \in \mathcal{F}_\tau$,

$$\begin{aligned} P\{B(t) - B(\tau) \le a, \tau \le t\} &= E\Big[P\{B'(t-\tau) \le a | \tau \le t, \mathcal{F}_\tau\}\Big] P\{\tau \le t\} \\ &= E\Big[P\{-B'(t-\tau) \le a | \tau \le t\}\Big] P\{\tau \le t\} \\ &= P\{B'(t-\tau) \ge -a, \tau \le t\}. \end{aligned}$$

Then using (5.12) in the last probability completes the proof.

We will now apply the reflection principle to obtain an expression for the joint distribution of $B(t)$ and $M(t)$.

Theorem 11. *For $x < y$ and $y \ge 0$,*

$$\begin{aligned} P\{B(t) \le x, M(t) \ge y\} &= P\{B(t) \ge 2y - x\}, \\ P\{M(t) \ge y\} &= 2P\{B(t) \ge y\}. \end{aligned} \tag{5.13}$$

Furthermore, $M(t) \stackrel{d}{=} |B(t)|$ for each t, and the density of $M(t)$ is

$$f_{M(t)}(x) = \frac{2}{\sqrt{2\pi t}} e^{-x^2/2t}, \qquad x \geq 0.$$

Hence

$$E[M(t)] = \sqrt{2t/\pi}, \qquad \text{Var}[M(t)] = (1 - 2/\pi)t. \tag{5.14}$$

Proof. Assertion (5.13) follows since by (5.11) and Proposition 10 with $\tau = \tau_y$, $B(\tau) = y$, and $a = x - y$, we have, for $x \leq y$ and $y \geq 0$,

$$\begin{aligned} P\{B(t) \leq x, M(t) \geq y\} &= P\{B(t) \leq x, \tau_y \leq t\} \\ &= P\{B(t) \geq 2y - x, \tau_y \leq t\} \\ &= P\{B(t) \geq 2y - x\}. \end{aligned}$$

The last equality is because $2y - x \geq y$ and

$$\{B(t) \geq 2y - x\} \subseteq \{B(t) \geq y\} \subseteq \{\tau_y \leq t\}.$$

Next, using what we just proved with $x = y$, we have

$$\begin{aligned} P\{M(t) \geq y\} &= P\{B(t) \leq y, M(t) \geq y\} + P\{B(t) \geq y, M(t) \geq y\} \\ &= 2P\{B(t) \geq y\}. \end{aligned}$$

Taking the derivative of this with respect to y yields the density of $M(t)$. In addition, $2P\{B(t) \geq y\} = P\{|B(t)| \geq y\}$ implies $M(t) \stackrel{d}{=} |B(t)|$. Exercise 11 proves (5.14).

Even though $M(t) \stackrel{d}{=} |B(t)|$ for each t, the processes M and $|B|$ are not equal in distribution; M is nondecreasing while $|B|$ is not. Exercise 19 points out the interesting equality in distribution $M \stackrel{d}{=} M - B$ for the processes.

Because of the reflection property $-B \stackrel{d}{=} B$ of a Brownian motion B, its minima is also a reflection of its maxima.

Remark 12. The *minimum process* for B is

$$\overline{M}(t) = \min_{s \leq t} B(s), \qquad t \geq 0.$$

It is related to the maximum process M by $\overline{M} \stackrel{d}{=} -M$. Hence

$$P\{\overline{M}(t) \leq a\} = 2P\{B(t) \geq -a\}, \qquad a < 0.$$

That $\overline{M} \stackrel{d}{=} -M$ follows by $\overline{M}(t) = -\max_{s \leq t} -B(t)$ and the reflection property $-B \stackrel{d}{=} B$. Also, Theorem 11 yields the distribution of $\overline{M}(t)$.

We will now obtain the distribution of the hitting time τ_a from Theorem 11. This result is also a special case of Theorem 32 for hitting times for a Brownian motion with drift, which also contains the Laplace transform of τ_a and shows that $E[\tau_a] = \infty$.

Corollary 13. *For any $a \geq 0$,*

$$P\{\tau_a \leq t\} = 2[1 - \Phi(a/\sqrt{t})], \qquad t > 0,$$

where Φ is the standard normal distribution. Hence, the density of τ_a is

$$f_{\tau_a}(t) = \frac{a}{\sqrt{2\pi t^3}} e^{-a^2/2t}, \qquad t > 0.$$

Proof. From (5.11), Theorem 11, and $B(t) \stackrel{d}{=} \sqrt{t}B(1)$, we have

$$\begin{aligned} P\{\tau_a \leq t\} &= P\{M(t) \geq a\} \\ &= 2P\{B(t) \geq a\} \;=\; 2[1 - \Phi(a/\sqrt{t})]. \end{aligned}$$

Taking the derivative of this yields the density of τ_a.

The family of hitting times $\{\tau_a : a \geq 0\}$ for B is an important process in its own right. It is the non-decreasing *left-continuous inverse process* of the maximum process M since

$$\tau_a = \inf\{t : B(t) = a\} = \inf\{t : M(t) = a\}.$$

By Corollary 13, we know the density of τ_a and that $E[\tau_a] = \infty$. Here is more information about these hitting times.

Proposition 14. *The process $\{\tau_a : a \geq 0\}$ has stationary independent increments and, for $a < b$, the increment $\tau_b - \tau_a$ is independent of $\mathcal{F}\tau_a$ and it is equal in distribution to $\tau_{(b-a)}$.*

Proof. Since $\tau_b - \tau_a = \inf\{t : B(\tau_a + t) - B(\tau_a) = b - a\}$, it follows by the strong Markov property at τ_a that $\tau_b - \tau_a$ is independent of $\mathcal{F}_{\tau_a}$ and it is equal in distribution to τ_{b-a}. Also, it follows by an induction argument that $\{\tau_a : a \geq 0\}$ has stationary independent increments.

5.4 Special Random Times

In this section, we derive arc sine and arc cosine probabilities for certain random times of Brownian motion by applying properties of the maximum process.

We first consider two random times associated with a Brownian motion B on the interval $[0, 1]$ and its maximum process $M(t) = \max_{s \leq t} B(s)$. These times have the same arc sine distribution, which is discussed in Exercise 14.

Theorem 15. (Lévy Arc Sine Law) *For a Brownian motion B on $[0, 1]$, the time $\tau = \inf\{t \in [0, 1] : B(t) = M(1)\}$ has the arc sine distribution*

$$P\{\tau \le t\} = \frac{2}{\pi} \arcsin \sqrt{t}, \qquad t \in [0,1]. \tag{5.15}$$

In addition, the time $\tau' = \sup\{t \in [0,1] : B(t) = 0\}$ *has the same distribution.*

Proof. First note that, for $t \le 1$,

$$\tau \le t \quad \Longleftrightarrow \quad \max_{s \le t} B(s) - B(t) \ge \max_{t \le s \le 1} B(s) - B(t).$$

Denote the last inequality as $Y_1 \ge Y_2$ and note that these random variables are independent since B has independent increments. Now, by the translation and symmetry properties of B and Theorem 11,

$$\begin{aligned} Y_1 &\stackrel{d}{=} M(t) \stackrel{d}{=} |B(t)| \stackrel{d}{=} t^{1/2}|Z_1|, \\ Y_2 &\stackrel{d}{=} M(1-t) \stackrel{d}{=} |B(1-t)| \stackrel{d}{=} (1-t)^{1/2}|Z_2|, \end{aligned}$$

where Z_1 and Z_2 are normal random variables with mean 0 and variance 1. From these observations, we have

$$\begin{aligned} P\{\tau \le t\} = P\{Y_1 \ge Y_2\} &= P\{tZ_1^2 \ge (1-t)Z_2^2\} \\ &= P\{Z_2^2/(Z_1^2 + Z_2^2) \le t\}, \end{aligned} \tag{5.16}$$

where we may take Z_1 and Z_2 to be independent. Then (5.15) follows since the last probability, due to the symmetry property of the normal distribution, is the arsine distribution by Exercise 14.

Next, note that by Remark 12 on the minimum process, we have

$$\begin{aligned} P\{\tau' < t\} &= P\{\max_{t \le s \le 1} B(s) > 0\} + P\{\min_{t \le s \le 1} B(s) < 0\} \\ &= 2P\{Y > -B(t)\}. \end{aligned}$$

where $Y = \max_{t \le s \le 1} B(s) - B(t)$ is independent of $B(t)$. By the symmetry of B and Theorem 11, we have

$$\begin{aligned} -B(t) &\stackrel{d}{=} B(t) \stackrel{d}{=} t^{1/2} Z_1, \\ Y &\stackrel{d}{=} M(1-t) \stackrel{d}{=} |B(1-t)| \stackrel{d}{=} (1-t)^{1/2}|Z_2|, \end{aligned}$$

where Z_1 and Z_2 are normal random variables with mean 0 and variance 1. Assuming Z_1 and Z_2 are independent, the preceding observations and (5.16) yield

$$\begin{aligned} P\{\tau' < t\} &= 2P\{(1-t)^{1/2}|Z_2| < t^{1/2} Z_1\} \\ &= P\{(1-t)Z_2^2 < tZ_1^2\} \;=\; P\{\tau \le t\}. \end{aligned}$$

This proves that τ' also has the arc sine distribution.

Next, we consider the event that a Brownian motion returns to the origin 0 in a future time interval.

Theorem 16. *The event A that a Brownian motion B hits 0 in a time interval $[t, u]$ has the probability*

$$P(A) = \frac{2}{\pi} \arccos \sqrt{t/u}, \qquad \textit{where } 0 < t < u. \tag{5.17}$$

Proof. For $t < u$ and $u = 1$, it follows by Theorem 15 that

$$P(A) = P\{\tau > t\} = 1 - \frac{2}{\pi} \arcsin \sqrt{t} \;=\; \frac{2}{\pi} \arccos \sqrt{t}.$$

The proof for $u \neq 1$ is Exercise 15.

5.5 Martingales

A martingale is a real-valued stochastic process defined by the property that the conditional mean of an "increment" of the process conditioned on past information is 0. A random walk and Brownian motion whose mean step sizes are 0 have this property. However, the increments of a martingale are generally dependent, unlike the independent increments of a random walk or Brownian motion. Martingales are used for proving convergence theorems, analyzing hitting times of processes, finding optimal stopping rules, and providing bounds for processes. Moreover, they are key tools in the theory of stochastic differential equations.

In this section, we introduce martingales and discuss several examples associated with Brownian motion and compound Poisson processes. We also present the important submartingale convergence theorem. The next two sections cover the optional stopping theorem for martingales and its applications to Brownian motion.

Throughout this section, $X = \{X(t) : t \geq 0\}$ will denote a real-valued continuous-time stochastic process that has right-continuous paths and $E[\,|X(t)|\,] < \infty$, $t \geq 0$. Associated with the underlying probability space $(\Omega, \mathcal{F}, P)$ for the process X, there is a *filtration* $\{\mathcal{F}_t : t \geq 0\}$, which is a family of σ-fields contained in $\mathcal{F}$ that is increasing ($\mathcal{F}_s \subseteq \mathcal{F}_t$, $s \leq t$) and right-continuous ($\mathcal{F}_t^B = \cap_{u>t} \mathcal{F}_u^B$), and $\mathcal{F}_0$ contains all events with P-probability 0. Furthermore, the process X is *adapted* to the filtration $\mathcal{F}_t$ in that $\{X(t) \leq x\} \in \mathcal{F}_t$, for each t and x.

Definition 17. The process X is a *martingale* with respect to $\mathcal{F}_t$ if

$$E[X(t)|\mathcal{F}_s] = X(s) \quad \text{a.s.} \qquad 0 \leq s < t. \tag{5.18}$$

The process X is a *submartingale* if

$$E[X(t)|\mathcal{F}_s] \geq X(s) \quad \text{a.s.} \qquad 0 \leq s < t.$$

If the inequality is reversed, then X is a *supermartingale*.

Taking the expectation of (5.18) yields the characteristic of a martingale that

$$E[X(t)] = E[X(s)], \qquad s \leq t.$$

The martingale condition (5.18) is equivalent to

$$E[X(t) - X(s)|\mathcal{F}_s] = 0,$$

which says that the conditional mean of an increment conditioned on the past is 0.

A classic illustration of a martingale is the value $X(t)$ of an investment (or the fortune of a gambler) at time t in a marketplace described by the events in $\mathcal{F}_t$. The martingale property (5.18) says that the investment is subject to a "fair market" in that its expected value at any time t conditioned on the environment $\mathcal{F}_s$ up to some time $s < t$ is the same as the value $X(s)$.

On the other hand, the submartingale property implies that the market is biased toward "upward" movements of the value X resulting in $E[X(t)] \geq X(s)$ a.s., for $s \leq t$. Similarly, the supermartingle property implies "downward" movements resulting in $E[X(t)] \leq X(s)$ a.s.

In typical applications, $\mathcal{F}_t = \mathcal{F}_t^Y$, which is the σ-field generated by the events of a right-continuous process $\{Y(s) : s \leq t\}$ on a general state space. In this setting, we say that X is a martingale *with respect to* Y. In some instances, it is natural that X is a martingale with respect to the filtration $\mathcal{F}_t = \mathcal{F}_t^X$ of its own history.

Martingales in discrete time are defined similarly. In particular, real-valued random variables X_n with $E[\,|X_n|\,] < \infty$ form a martingale with respect to increasing σ-fields $\mathcal{F}_n$ if

$$E[X_{n+1}|\mathcal{F}_n] = X_n, \qquad n \geq 0.$$

The X_n is a submartingale or supermartingale if the equality is replaced by $\geq$ or $\leq$, respectively. Standard examples are sums and products of independent random variables; see Exercise 30.

Note that a Brownian motion B is a martingale with respect to itself since, for $s \leq t$,

$$E[B(t)|\mathcal{F}_s^B] = E\Big[B(t) - B(s)\Big|B(s)\Big] + B(s) = B(s).$$

Similarly, if $X(t) = x + \mu t + \sigma B(t)$ is a Brownian motion with drift, then

$$E[X(t)|\mathcal{F}_s^B] = \mu(t-s) + X(s), \qquad s \leq t.$$

Therefore, X is a martingale, submartingale, or supermartingale with respect to B according as μ is $= 0$, > 0, or < 0.

We will also encounter several functions of Brownian motion that are martingales of the following type.

Example 18. Martingales For Processes with Stationary Independent Increments. Suppose that Y is a real-valued process that has stationary independent increments and, for simplicity, assume that $Y(0) = 0$. Suppose that the moment generating function $\psi(\alpha) = E[e^{\alpha Y(1)}]$ exists for α in a neighborhood of 0, and that $E[e^{\alpha Y(t)}]$ as a function of t is continuous at 0, for fixed α.

Then by Exercise 7, $E[e^{\alpha Y(t)}] = \psi(\alpha)^t$ and

$$E[Y(t)] = at, \qquad \text{Var}[Y(t)] = bt,$$

where $a = E[Y(1)]$ and $b = \text{Var}[Y(1)]$. For instance, Y may be Brownian motion with drift, a Poisson process or a compound Poisson process.

An easy check shows that two martingales with respect to Y are

$$Y(t) - at, \quad \text{and} \quad (Y(t) - at)^2 - bt, \qquad t \geq 0.$$

The means of these martingales are 0.

Next, consider the process

$$Z(t) = e^{\alpha Y(t)}/\psi(\alpha)^t, \qquad t \geq 0.$$

Clearly $Z(t)$ is a deterministic, nonnegative function of $\{Y(s) : s \leq t\}$, and $E[Z(t)] = 1$. Then Z is a martingale (sometimes called an *exponential* martingale) with respect to Y. Indeed,

$$E[Z(t)|\mathcal{F}_s^Y] = Z(s)\frac{E[e^{\alpha(Y(t)-Y(s))}|Z(s)]}{\psi(\alpha)^{t-s}} = Z(s).$$

Example 19. Martingales for Brownian Motion. For a Brownian motion with drift $Y(t) = x + \mu t + \sigma B(t)$, the preceding example justifies that the following functions of Y are martingales with respect to B:

$$(Y(t) - x - \mu t)^2 - \sigma^2 t, \qquad e^{c[Y(t)-x-\mu t]-e^{-c^2\sigma^2 t/2}}, \qquad t \geq 0,\ c \neq 0.$$

In particular, $B(t)^2 - t$ and $e^{cB(t)-c^2t/2}$ are martingales with respect to B.

Having a constant mean suggests that a martingale should also have a nice limiting behavior. According to the next major theorem, many submartingales as well as martingales converge a.s. to a limit. This result for discrete-time processes also holds in continuous-time.

Theorem 20. (Submartingale Convergence) *If X_n is a martingale, or a submartingale that satisfies $\sup_n E[X_n^+] < \infty$, then there exists a random variable X with $E[\,|X|\,] < \infty$ such that $X_n \to X$ a.s. as $n \to \infty$.*

This convergence can be viewed as an extension of the fact that a nondecreasing sequence of real numbers that is bounded converges to a finite limit. For a submartingale, the nondecreasing tendency is $E[X_{n+1}|\mathcal{F}_n] \le X_n$, and a bound on $E[X_n^+]$ is enough to ensure convergence a.s. — the submartingale itself need not be nondecreasing.

The theorem establishes the existence of the limit X, but it does not specify its distribution. Properties of X can sometimes be derived in specific cases depending on characteristics of X_n.

In addition to the convergence $X_n \to X$ a.s., it follows by Theorem 15 in the Appendix that $E[\,|X_n - X|\,] \to 0$ as $n \to \infty$ when the X_n are uniformly integrable.

Although Theorem 20 is a major result, we will only give the following example since it is not needed for our results. For a proof and other applications, see for instance [37, 61, 62, 64].

Example 21. Doob Martingale. Let Z be a random variable with $E[\,|Z|\,] < \infty$, and let $\mathcal{F}_n$ be an increasing filtration on the underlying probability space for Z. The conditional expectation

$$X_n = E[Z\,|\mathcal{F}_n], \qquad n \ge 1,$$

is a martingale with respect to $\mathcal{F}_n$. Then by Theorem 20

$$X = \lim_{n\to\infty} E[Z|\mathcal{F}_n] \quad \text{exists a.s.}$$

That X_n is a martingale follows since

$$E[\,|X_n|\,] \le E\Big[E[\,|Z|\,|\mathcal{F}_n]\Big] = E[\,|Z|\,] < \infty,$$
$$E[X_{n+1}|\mathcal{F}_n] = E\Big[E[Z\,|\mathcal{F}_{n+1}]\,\Big|\mathcal{F}_n\Big] = E[Z\,|\mathcal{F}_n] = X_n.$$

Consider the case $X_n = E[Z\,|Y_1, \ldots, Y_n]$ in which X_n is a martingale with respect to Y_n. For instance, X_n could be an estimate for the mean of Z based on observations $Y_1, \ldots, Y_n$ associated with Z. By an additional argument it follows that the limit of X_n is $X = E[Z\,|Y_1, Y_2, \ldots]$. Therefore,

$$E[Z\,|Y_1, \ldots, Y_n] \to E[Z\,|Y_1, Y_2, \ldots] \quad \text{a.s. as } n \to \infty.$$

In particular, if Z is the indicator of an event A in the σ-field generated by $Y_1, Y_2, \ldots$, then

$$P(A\,|Y_1, \ldots, Y_n) \to P(A) \quad \text{a.s. as } n \to \infty.$$

5.6 Optional Stopping of Martingales

This section presents the optional stopping theorem for martingales. It was instrumental in the proof of the strong Markov property for Brownian motion in Theorem 3; the next section uses it to analyze hitting times of Brownian motion.

For the following discussion, suppose that X is a martingale with respect to a filtration $\mathcal{F}_t$, and that τ is a stopping time of the filtration: $\{\tau \le t\} \in \mathcal{F}_t$, for each t.

We will now address the following questions: Is the martingale property, $E[X(t)] = E[X(0)]$ also true when t is a stopping time? More generally, is $E[X(\sigma)] = E[X(\tau)]$ true for any stopping times σ and τ?

The optional stopping theorem below says that $E[X(\tau)] = E[X(0)]$ is indeed true for a bounded stopping time τ. A corollary is that the equality is also true for a finite stopping time when X satisfies certain bounds. This would imply, for instance, that the expected value of an investment in a fair market at the stopping time is the same as the initial value. In other words, in this fair market, there would be no benefit for the investor to choose to stop and freeze his investment at a time depending only on the past history of the market (independent of the future).

There are several optional stopping theorems with slightly different assumptions. For our purposes, we will use the following version from [61]. Its discrete-time version is Theorem 28 below.

Theorem 22. *Associated with the martingale X, assume that σ and τ are stopping times of $\mathcal{F}_t$ such that τ is bounded a.s. Then*

$$X(\sigma \wedge \tau) = E[X(\tau)|\mathcal{F}_\sigma] \quad a.s. \tag{5.19}$$

Hence $E[X(\tau)] = E[X(0)]$. If σ is also bounded, then $E[X(\sigma)] = E[X(\tau)]$.

Proof. Our proof for this continuous time setting uses an approximation based on the analogous discrete-time result in Theorem 28 below. For a fixed $n \in \mathbb{Z}_+$, let $\bar{X}_k = X(k2^{-n})$, $k \in \mathbb{Z}_+$. Clearly $\bar{X}_k$ is a discrete-time martingale with respect to $\bar{\mathcal{F}}_k = \mathcal{F}_{k2^{-n}}$. Define

$$\sigma_n = \lfloor 2^n\sigma + 1 \rfloor / 2^n, \qquad \tau_n = \lfloor 2^n\tau + 1 \rfloor / 2^n.$$

Now $\sigma'_m = 2^n\sigma_m$ for fixed m, and $\tau'_n = 2^n\tau_n$ are integer-valued stopping times of $\bar{\mathcal{F}}_k$. Then by Theorem 28 below,

$$\bar{X}_{\sigma'_m \wedge \tau'_n} = E[\bar{X}_{\tau'_n}|\bar{\mathcal{F}}_{\sigma'_m}] \quad \text{a.s.}$$

This expression in terms of the preceding definitions is

$$X(\sigma_m \wedge \tau_n) = E[X(\tau_n)|\mathcal{F}_{\sigma_m}] \quad \text{a.s.}$$

Letting $m \to \infty$ in this expression results in $\sigma_m \to \sigma$ and

$$X(\sigma \wedge \tau_n) = E[X(\tau_n)|\mathcal{F}_\sigma] \quad \text{a.s.}$$

Then letting $n \to \infty$ in this expression yields (5.19). The justifications of the last two limit statements, which are in [61], will not be covered here.

The assertion $E[X(0)] = E[X(\tau)]$ follows by taking expectations in (5.19) with $\sigma = 0$. Finally, when σ as well as τ is bounded, then (5.19) and this expression with the roles of σ and τ reversed yield

$$E[X(\tau)|\mathcal{F}_\sigma] = X(\sigma \wedge \tau) = E[X(\sigma)|\mathcal{F}_\tau] \quad a.s.$$

Then expectations of these terms give $E[X(\sigma)] = E[X(\tau)]$.

Theorem 22 can also be extended to stopping times that are a.s. finite, but not necessarily bounded. To see this, suppose that τ is an a.s. finite stopping time of $\mathcal{F}_t$. The key idea is that, for fixed s and t, the $\tau \wedge s$ and $\tau \wedge t$ are a.s. bounded stopping times of $\mathcal{F}_t$. Then by Theorem 22,

$$X(\tau \wedge s) = E[X(\tau \wedge t)|\mathcal{F}_{\tau \wedge s}], \qquad s < t.$$

This property justifies the following fact, which is used in the proof below.

Remark 23. Stopped Martingales. The stopped process $X(\tau \wedge t)$ is a martingale with respect to $\mathcal{F}_t$.

Corollary 24. *Associated with the martingale X, suppose that τ is an a.s. finite stopping time of $\mathcal{F}_t$, and that either one of the following conditions is satisfied.*

(i) $E\Big[\sup_{t \le \tau} |X(t)|\Big] < \infty$.

(ii) $E[\,|X(\tau)|\,] < \infty$, *and* $\lim_{u \to \infty} E[\,|X(u)|1(\tau > u)] = 0$.

Then $E[X(\tau)] = E[X(0)]$.

Proof. Since $X(\tau \wedge t)$ is a martingale with respect to $\mathcal{F}_t$, Theorem 22 implies $E[X(\tau \wedge u)] = E[X(0)]$, for $u > 0$. Now, we can write

$$\begin{aligned} |E[X(\tau)] - E[X(0)]| &= |E[X(\tau)] - E[X(\tau \wedge u)]| \\ &\le E[\,|X(\tau) - X(u)|\mathbf{1}(\tau > u)]\,|. \end{aligned}$$

If (i) holds, then $|X(\tau) - X(u)| \le 2Z$, where $Z = \sup_{t \le \tau} |X(t)|$. Since τ is finite a.s., $\mathbf{1}(\tau > u) \to 0$ a.s. as $u \to \infty$, and so by the dominated convergence theorem,

$$|E[X(\tau)] - E[X(0)]| \le 2E[Z\mathbf{1}(\tau > u)] \to 0.$$

On the other hand, if (ii) holds, then

$$|E[X(\tau)] - E[X(0)]| \le E\Big[\Big(|X(\tau)| + |X(u)|\Big)\mathbf{1}(\tau > u)\Big] \to 0.$$

Thus, $E[X(\tau)] = E[X(0)]$ if either (i) or (ii) is satisfied.

The next proposition and example illustrate computations involving optional stopping.

Proposition 25. (Wald Identities) *Let X be a process with stationary independent increments as in Example 18, with $E[\,|X(1)|\,] < \infty$ and $\psi(\alpha) = E[e^{\alpha X(1)}]$. Suppose τ is an a.s. finite stopping time of X. Then*

$$E[X(\tau)] = E[X(1)]E[\tau].$$

If in addition τ is bounded or $E\Big[\sup_{t\le\tau} |X(t)|\Big] < \infty$, then

$$E[e^{\alpha X(\tau)}\psi(\alpha)^{-\tau}] = 1, \quad \textit{for any } \alpha \textit{ with } \psi(\alpha) \ge 1. \tag{5.20}$$

Proof. Example 18 establishes that $X(t) - E[X(1)]t$ is a martingale with respect to X. Now, $\tau \wedge t$ is a bounded stopping time of X, and so by the optimal stopping theorem, $E[X(\tau \wedge t) - E[X(1)](\tau \wedge t)] = 0$. Letting $t \to \infty$ in this expression, the dominated and monotone convergence theorems yield $E[X(\tau)] = E[X(1)]E[\tau]$.

Similarly, $Z(t) = e^{\alpha X(t)}\psi(\alpha)^{-t}$ is a martingale with respect to X, and under the assumptions the optional stopping theorem or Corollary 24 imply that $E[Z(\tau)] = E[Z(0)] = 1$, which gives (5.20).

Example 26. Brownian Optional Stopping. For a Brownian motion B, we know by Example 19 that the processes $B(t)$ and $B(t)^2 - t$ are martingales with respect to B with zero means. Then as in the preceding proposition, we have the following result.

If τ is an a.s. finite stopping time of B, then $E[B(\tau)] = 0$. In addition, $E[\tau] = E[B(\tau)^2]$ if τ is bounded a.s.

Example 19 also noted that $X(t) = e^{cB(t)-c^2t/2}$ is a martingale with respect to B with mean 1. If τ is an a.s. bounded stopping time of B, then $E[X(\tau)] = 1$ by the optional stopping theorem. Consequently, the conditional moment generating function for an increment of B following τ is

$$E[e^{c[B(\tau+u)-B(\tau)]}|\mathcal{F}_\tau] = e^{c^2u/2} = E[e^{cB(u)}].$$

This was the key step in proving the strong Markov property of B for bounded stopping times.

The rest of this section is devoted to proving the discrete-time optional stopping theorem used in the proof of Theorem 22. We begin with a preliminary result.

Proposition 27. *Let X and Y be random variables on a probability space, and let $\mathcal{F}$ and $\mathcal{G}$ be two σ-fields on the space. Suppose there is an event $A \in \mathcal{F} \cap \mathcal{G}$ such that $X = Y$ a.s. on A and $\mathcal{F} = \mathcal{G}$ on A ($A \cap \mathcal{F} = A \cap \mathcal{G}$). Then $E[X|\mathcal{F}] = E[Y|\mathcal{G}]$ a.s. on A.*

Proof. Let $Z = E[X|\mathcal{F}] - E[Y|\mathcal{G}]$ and $C = A \cap \{Z > 0\}$. Under the hypotheses, $C \in \mathcal{F} \cap \mathcal{G}$ and

$$E[Z\mathbf{1}_C] = E\Big[E[X|\mathcal{F}]\mathbf{1}_C - E[Y|\mathcal{G}]\mathbf{1}_C\Big] = E[X\mathbf{1}_C - Y\mathbf{1}_C] = 0.$$

Here $\mathbf{1}_C$ is the random variable $\mathbf{1}(\omega \in C)$. Because a nonnegative random variable V is 0 a.s. if and only if $E[V] = 0$, it follows that $Z\mathbf{1}_C = 0$ a.s., which implies $Z \leq 0$ a.s. on A. A similar argument with $C = A \cap \{Z < 0\}$, shows $Z \geq 0$ a.s. on A. This proves the assertion.

Theorem 28. *Suppose that $\{X_n : n \in \mathbb{Z}_+\}$ is a martingale with respect to $\mathcal{F}_n$, and that σ and τ are stopping times of $\mathcal{F}_n$ such that τ is bounded a.s. Then*

$$X_{\sigma\wedge\tau} = E[X_\tau|\mathcal{F}_\sigma] \quad a.s..$$

Hence $E[X_\tau] = E[X_0]$. If σ is also bounded, then $E[X_\sigma] = E[X_\tau]$.

Proof. For $m \leq n$, one can show that $\mathcal{F}_\tau = \mathcal{F}_m$ on $\{\tau = m\}$. Then by Proposition 27 and the martingale property,

$$E[X_n|\mathcal{F}_\tau] = E[X_n|\mathcal{F}_m] = X_m = X_\tau, \quad \text{a.s. on } \{\tau = m\}.$$

Since this is true for each $m \leq n$, we have

$$E[X_n|\mathcal{F}_\tau] = X_\tau, \quad \text{a.s. if } \tau \leq n \quad \text{a.s.} \tag{5.21}$$

Now, consider the case $\sigma \leq \tau \leq n$ a.s. Then $\mathcal{F}_\sigma \subseteq \mathcal{F}_\tau$. Using this and (5.21) for τ and for σ, we get

$$E[X_\tau|\mathcal{F}_\sigma] = E\Big[E[X_n|\mathcal{F}_\tau]\Big|\mathcal{F}_\sigma\Big] = E[X_n|\mathcal{F}_\sigma] = X_\sigma \quad \text{a.s.}$$

In addition, $E[X_\tau|\mathcal{F}_\sigma] = X_\tau$ a.s. if $\tau \leq \sigma \wedge n$.

For the general case, similar reasoning using the preceding two results and Proposition 27 yield

$$\begin{aligned} E[X_\tau|\mathcal{F}_\sigma] &= E[X_\tau|\mathcal{F}_{\sigma\wedge\tau}] = X_{\sigma\wedge\tau} \quad \text{a.s. on } \{\sigma \leq \tau\} \\ E[X_\tau|\mathcal{F}_\sigma] &= E[X_{\sigma\wedge\tau}|\mathcal{F}_\sigma] = X_{\sigma\wedge\tau} \quad \text{a.s. on } \{\sigma > \tau\}. \end{aligned}$$

This proves $X_{\sigma\wedge\tau} = E[X_\tau|\mathcal{F}_\sigma]$ a.s. The other assertions follow as in the proof of Theorem 22.

5.7 Hitting Times for Brownian Motion with Drift

We will now address the following questions for a Brownian motion with drift. What is the probability that the process hits b before it hits a, for $a < b$?

What is the distribution and mean of the time for the process to hit b? We answer these questions by applications of the material in the preceding section on martingales and optional stopping.

Consider a Brownian motion with drift $X(t) = x + \mu t + \sigma B(t)$, where B is a standard Brownian motion. For $a < x < b$, let τ_a and τ_b denote the times at which X hits the respective states a and b. In addition, let $\tau = \tau_b \wedge \tau_a$, which is the time at which X escapes from the open strip (a, b). Our focus will be on properties of these hitting times.

Remark 29. Finiteness of Hitting Times. If $\mu \geq 0$, then τ_b is finite a.s. since using Remark 9,

$$\tau_b = \inf\{t \geq 0 : X(t) = b\} \leq \inf\{t \geq 0 : B(t) = (b - x)/\sigma\} < \infty \quad \text{a.s.}$$

Similarly, τ_a is finite a.s. if $\mu \leq 0$. Also, τ is finite since either τ_a or τ_b is necessarily finite.

We begin with a result for a Brownian motion with no drift.

Theorem 30. *The probability that the process $X(t) = x + \sigma B(t)$ hits b before a is*

$$P\{\tau_b < \tau_a\} = (x - a)/(b - a). \tag{5.22}$$

Also, $E[\tau] = (x - a)(b - x)/\sigma^2$.

Proof. By Example 19, X is a martingale with respect to B with mean x. Also, $E[\sup_{t \leq \tau} |X(t)|]$ is finite since $X(t) \in (a, b)$ for $t \leq \tau$. Then by the optional stopping theorem (Theorem 22) for τ,

$$E[X(\tau)] = E[X(0)] = x. \tag{5.23}$$

Now, since $\tau = \tau_a \wedge \tau_b$, we can write

$$X(\tau) = a\mathbf{1}(\tau_a \leq \tau_b) + b\mathbf{1}(\tau_b < \tau_a). \tag{5.24}$$

Then

$$E[X(\tau)] = a[1 - P\{\tau_b < \tau_a\}] + bP\{\tau_b < \tau_a\}.$$

Substituting this in (5.23) and solving for $P\{\tau_b < \tau_a\}$, we obtain (5.22).

Next, we know by Example 19 that $Z(t) = (X(t) - x)^2 - \sigma^2 t$ is a martingale with respect to B. Then the optional stopping theorem for the bounded stopping time $\tau \wedge t$ yields $E[Z(\tau \wedge t)] = E[Z(0)] = 0$. That is,

$$\sigma^2 E[\tau \wedge t] = E[(X(\tau \wedge t) - x)^2].$$

Now since $\tau \wedge t \uparrow \tau$ and $X(t)$ is bounded for $t \leq \tau$, it follows by the monotone and bounded convergence theorems that

$$\sigma^2 E[\tau] = E[(X(\tau) - x)^2].$$

Then using (5.24) in the last expectation followed by (5.22), we have

$$\begin{aligned}\sigma^2 E[\tau] &= (a-x)^2 P\{\tau_a \le \tau_b\} + (b-x)^2 P\{\tau_b < \tau_a\} \\ &= (x-a)(b-x).\end{aligned}$$

The preceding result for a Brownian motion with no drift has the following analogue for a Brownian motion $X(t) = x + \mu t + \sigma B(t)$ with drift $\mu \neq 0$.

Theorem 31. *The probability that the process X hits b before a is*

$$P\{\tau_b < \tau_a\} = \frac{e^{\alpha x} - e^{\alpha a}}{e^{\alpha b} - e^{\alpha a}}, \tag{5.25}$$

where $\alpha = -2\mu/\sigma^2$. In addition,

$$E[\tau] = \mu^{-1}\Big[(a-x) + (b-a)P\{\tau_b < \tau_a\}\Big]. \tag{5.26}$$

Proof. As in Example 19, $Z(t) = \exp\{cX(t) - cx - (c\mu + c^2\sigma^2/2)t\}$ is a martingale with respect to B. Letting $c = \alpha$, this martingale reduces to $Z(t) = e^{\alpha X(t) - \alpha x}$.

Now, $E[\sup_{t \le \tau} |Z(t)|]$ is finite, since $X(t) \in [a, b]$ for $t \le \tau$. Then by Corollary 24 on optional stopping,

$$1 = E[Z(0)] = E[Z(\tau)] = e^{-\alpha x} E[e^{\alpha X(\tau)}].$$

Now, using $X(\tau) = a\mathbf{1}(\tau_a \le \tau_b) + b\mathbf{1}(\tau_b < \tau_a)$ in this expression yields

$$e^{\alpha x} = e^{\alpha a}(1 - P\{\tau_b < \tau_a\}) + e^{\alpha b} P\{\tau_b < \tau_a\}.$$

This proves (5.25).

To determine $E[\tau]$, we apply the optional stopping theorem to the martingale $B(t) = \sigma^{-1}[X(\tau) - x - \mu\tau]$ and the bounded stopping time $\tau \wedge t$ to get $0 = E[B(\tau \wedge t)]$. That is,

$$\mu E[\tau \wedge t] + x = E[X(\tau \wedge t)].$$

Letting $t \to \infty$ in this expression, we have $\tau \wedge t \uparrow \tau$, and so the monotone and bounded convergence theorems yield

$$\mu E[\tau] + x = E[X(\tau)] = bP\{\tau_b < \tau_a\} + a(1 - P\{\tau_b < \tau_a\}).$$

This proves (5.26).

The last result of this section characterizes the distribution of the hitting time τ_b for a Brownian motion $X(t) = \mu t + \sigma B(t)$ with drift μ.

Theorem 32. *Let τ_b denote the time at which the Brownian motion X hits $b > 0$. If $\mu \ge 0$, then $E[\tau_b] = -b/\mu$ (which is ∞ if $\mu = 0$) and the Laplace transform and density of τ_b are*

$$E[e^{-\lambda\tau_b}] = \exp\{-b\sigma^{-2}[\sqrt{\mu^2 + 2\sigma^2\lambda} - \mu]\}, \tag{5.27}$$

$$f_{\tau_b}(x) = \frac{b}{\sigma\sqrt{2\pi x^3}} \exp\{-(b - \mu x)^2/(2\sigma^2 x\}. \tag{5.28}$$

If $\mu < 0$, then τ_b may be infinite and $P\{\tau_b < \infty\} = e^{2b\mu/\sigma^2}$.

Proof. For the case $\mu < 0$, it follows by (5.25) (with $x = 0$ and $\alpha > 0$) that

$$P\{\tau_b < \infty\} = \lim_{a \to -\infty} P\{\tau_b < \tau_a\} = e^{2b\mu/\sigma^2}.$$

Next, consider the case $\mu \geq 0$. For positive constants α and λ, consider the process $Z(t) = e^{\alpha X(t) - \lambda t}$. This is a martingale (see Example 19) and it reduces to $Z(t) = e^{cB(t) - c^2 t/2}$, where

$$c = \alpha\sigma, \qquad \alpha = \sigma^{-2}[\sqrt{\mu^2 + 2\sigma^2\lambda} - \mu].$$

This choice of α ensures that $\alpha^2\sigma^2/2 + \alpha\mu - \lambda = 0$.

Now, applying the optional stopping theorem to the martingale $Z(t)$ and the bounded stopping time $\tau_b \wedge t$, we obtain

$$1 = E[Z(0)] = E[Z(\tau_b \wedge t)] = E[e^{\alpha X(\tau_b \wedge t) - \lambda(\tau_b \wedge t)}].$$

Since X is continuous a.s., we have

$$\lim_{t \to \infty} [\alpha X(\tau_b \wedge t) - \lambda(\tau_b \wedge t)] = \alpha b - \lambda\tau_b \quad \text{a.s.}$$

Then by the preceding displays and the bounded convergence theorem,

$$1 = E[\lim_{t \to \infty} Z(\tau_b \wedge t)] = e^{\alpha b} E[e^{-\lambda\tau_b}].$$

This proves (5.27). Inverting this Laplace transform yields the density formula (5.28). Finally, the derivative of this transform at $\lambda = 0$ yields $E[\tau_b] = -b/\mu$.

5.8 Limiting Averages and Law of the Iterated Logarithm

This section contains strong laws of large numbers for Brownian motion and its maximum process, and a law of the iterated logarithm for Brownian motion.

As usual B will denote a standard Brownian motion. The strong law of large numbers for it is as follows.

Theorem 33. *A Brownian motion B has the limiting average*

$$\lim_{t\to\infty} t^{-1}B(t) = 0 \quad a.s.$$

Proof. Since B has stationary independent increments, it has regenerative increments with respect to the deterministic times $T_n = n$. Then the assertion will follow by the SLNN in Theorem 54 in Chapter 2 for processes with regenerative increments upon showing that $n^{-1}B(n) \to 0$ a.s., and

$$E\Big[\max_{n-1\le t\le n} |B(t)|\Big] < \infty. \tag{5.29}$$

Now, the SLLN for i.i.d. random variables ensures that

$$n^{-1}B(n) = n^{-1}\sum_{m=1}^{n}[B(m) - B(m-1)] \to E[B(1)] = 0, \quad \text{a.s.}$$

Also, (5.29) follows since $E[M(1)] < \infty$ and

$$\max_{n-1\le t\le n} |B(t)| \stackrel{d}{=} \max_{0\le t\le 1} |B(t)| \le M(1) - \overline{M}(1),$$

where $\overline{M}(t) = \min_{s\le t} B(s) \stackrel{d}{=} -M(t)$ by Remark 12.

If a real-valued process X, such as a Brownian motion or a functional of a Markov process, has a limiting average $t^{-1}X(t) \to c$ a.s., you might wonder if its maximum $M(t) = \sup_{s\le t} X(s)$ also satisfies $t^{-1}M(t) \to c$ a.s. Wonder no longer. The answer is given by the next property, which is analogous to the elementary fact that if $n^{-1}c_n \to c$, then $n^{-1}\sum_{k=1}^{n} c_k \to c$.

Proposition 34. *Let $x(t)$ and $a(t)$ be real-valued functions on $\mathbb{R}_+$ such that*

$$0 \le a(t) \to \infty, \qquad a(t)^{-1}x(t) \to c, \quad \text{as } t \to \infty.$$

Then the maximum $m(t) = \sup_{s\le t} x(s)$ satisfies $\lim_{t\to\infty} a(t)^{-1}m(t) = c$.

Proof. For any $\varepsilon > 0$, let t' be such that $a(t)^{-1}x(t) < c + \varepsilon$, for $t \ge t'$. Then for $t \ge t'$,

$$\begin{aligned} a(t)^{-1}x(t) \le a(t)^{-1}m(t) &= \max\Big\{a(t)^{-1}m(t'), a(t)^{-1}\sup_{t'\le s\le t} x(s)\Big\} \\ &\le \max\big\{a(t)^{-1}m(t'), c+\varepsilon\big\}. \end{aligned}$$

Letting $t \to \infty$ in this display, it follows that

$$c \le \liminf_{t\to\infty} a(t)^{-1}m(t) \le \limsup_{t\to\infty} a(t)^{-1}m(t) \le c + \varepsilon.$$

Since this is true for any ε, we have $a(t)^{-1}m(t) \to c$.

We will now apply this result to the maximum process $M(t) = \max_{s\le t} X(s)$ for a Brownian motion with drift $X(t) = x + \mu t + \sigma B(t)$.

Proposition 35. *The Brownian motion with drift X and its maximum process have the limiting averages*

$$t^{-1}X(t) \to \mu, \qquad t^{-1}M(t) \to \mu \quad \textit{a.s. as } t \to \infty.$$

Proof. This follows since the SLLN $t^{-1}B(t) \to 0$ implies that

$$t^{-1}X(t) = t^{-1}x + \mu + \sigma t^{-1}B(t) \to \mu \quad \text{a.s.},$$

and then $t^{-1}M(t) \to \mu$ a.s. follows by Proposition 34.

The preceding result implies that $M(t) \to \infty$ or $-\infty$ a.s. according as the drift μ is positive or negative. This tells us something about the maximum

$$M(\infty) = \sup_{t \in \mathbb{R}_+} X(t)$$

on the entire time axis. First, we have the obvious result

$$M(\infty) = \lim_{t \to \infty} M(t) = \infty \quad \text{a.s. when } \mu > 0.$$

Second, $M(\infty) = \infty$ a.s. when $\mu = 0$ by the law of the iterated logarithm in (5.30) below.

For the remaining case of $\mu < 0$, we have the following result.

Theorem 36. *If $\mu < 0$ and $X(0) = 0$, then $M(\infty)$ has an exponential distribution with rate $-2\mu/\sigma^2$.*

Proof. The assertion follows, since letting $t \to \infty$ in $\{M(t) > b\} = \{t_b < t\}$ and using Theorem 32,

$$P\{M(\infty) > b\} = P\{\tau_b < \infty\} = e^{2\mu b/\sigma^2}.$$

We will now consider fluctuations of Brownian motions that are described by a law of the iterated logarithm. Knowing that the limiting average of Brownian motion B is 0 as $t \to \infty$, a follow-on issue is to characterize its fluctuations about 0. These fluctuations, of course, can be described for a "fixed" t by the normal distribution of $B(t)$; e.g., $P\{|B(t)| \le 2\sqrt{t}\} \approx .95$.

However, to get a handle on rare fluctuations as $t \to \infty$, it is of interest to find constants $h(t)$ such that

$$\limsup_{t \to \infty} \frac{B(t)}{h(t)} = 1 \quad \text{a.s.}$$

In other words, $h(t)$ is the maximum height of the fluctuations of $B(t)$ above 0, and $B(t)$ gets near $h(t)$ infinitely often (i.o.) as $t \to \infty$ in that

$$P\{B(t) \in [h(t) - \varepsilon, h(t)] \quad \text{i.o.}\} = 1, \qquad \varepsilon > 0.$$

Since the reflection $-B$ is a Brownian motion, the preceding would also yield

$$\liminf_{t\to\infty} \frac{B(t)}{h(t)} = -1 \quad \text{a.s.}$$

These fluctuations are related to those as $t \downarrow 0$ as follows.

Remark 37.

$$\limsup_{t\to\infty} \frac{B(t)}{h(t)} = 1 \quad \text{a.s.} \quad \Longleftrightarrow \quad \limsup_{t\downarrow 0} \frac{B(t)}{th(1/t)} = 1 \quad \text{a.s.}$$

This is because the time-inversion process $X(t) = tB(1/t)$ is a Brownian motion by Exercise 2. Indeed, the equivalence is true since, using $s = 1/t$,

$$\limsup_{t\to\infty} \frac{B(t)}{h(t)} = \limsup_{s\downarrow 0} \frac{X(s)}{sh(1/s)}.$$

Remark 37 says that $h(t)$ is the height function for fluctuations of B as $t \to \infty$ if and only if $th(1/t)$ is the height function for fluctuations as $t \downarrow 0$. The height functions for both of these cases are as follows. The proof, due to Khintchine 1924, is in [37, 61, 64].

Theorem 38. (Law of the Iterated Logarithm)

$$\limsup_{t\downarrow 0} \frac{B(t)}{\sqrt{2t\log\log(1/t)}} = 1, \qquad \limsup_{t\to\infty} \frac{B(t)}{\sqrt{2t\log\log t}} = 1 \quad \textit{a.s.}$$

$$\liminf_{t\downarrow 0} \frac{B(t)}{\sqrt{2t\log\log(1/t)}} = -1, \qquad \liminf_{t\to\infty} \frac{B(t)}{\sqrt{2t\log\log t}} = -1 \quad \textit{a.s.}$$

Note that the $\limsup_{t\downarrow 0}$ result implies that $B(t) > 0$ i.o. near 0, and so

$$\inf\{t > 0 : B(t) > 0\} = 0 \quad \text{a.s.}$$

Similarly, the $\liminf_{t\downarrow 0}$ result implies that $B(t) < 0$ i.o. near 0 a.s. Consequently, $B(t) = 0$ i.o. near 0 a.s. because B has continuous paths a.s.

The other results for $t \to \infty$ imply that, for any fixed $a > 0$, we have $B(t) > a$ and $B(t) < -a$ i.o. a.s. as $t \to \infty$, and so B passes through $[-a, a]$ i.o. a.s. Furthermore, the extremes of B are

$$\sup_{t\in\mathbb{R}_+} B(t) = \infty \quad \text{a.s.}, \qquad \inf_{t\in\mathbb{R}_+} B(t) = -\infty \quad \text{a.s.} \tag{5.30}$$

5.9 Donsker's Functional Central Limit Theorem

By the classical central limit theorem (Theorem 63 in Chapter 2), we know that a random walk under an appropriate normalization converges in distribution to a normal random variable. This section extends this result to stochastic processes. In particular, viewing a random walk as a process in continuous time, if the time and space parameters are rescaling appropriately, then the random walk process converges in distribution to a Brownian motion. This result, called Donsker's functional central limit theorem, also establishes that many functionals of random walks can be approximated by corresponding functionals of Brownian motion.

Throughout this section $S_n = \sum_{i=1}^n \xi_k$ will denote a random walk in which the step sizes ξ_n are i.i.d. with mean 0 and variance 1. For each n, consider the stochastic process

$$X_n(t) = n^{-1/2} S_{\lfloor nt \rfloor}, \qquad t \in [0,1].$$

That is,

$$X_n(t) = n^{-1/2} S_k \quad \text{if } k/n \le t < (k+1)/n \text{ for some } k < n.$$

This process is a continuous-time representation of the random walk S_k in which the location S_k is rescaled (or shrunk) to the value $n^{-1/2}S_k$, and the time scale is rescaled such that the walk takes $[nt]$ steps in time t. Then as n becomes large the steps become very small and frequent and, as we will show, X_n converges in distribution to a standard Brownian motion B as $n \to \infty$.

We begin with the preliminary observation that the finite-dimensional distributions of X_n converge in distribution to those of B. That is, for any fixed $t_1 < \cdots < t_k$,

$$(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{d} (B(t_1), \ldots, B(t_k)), \quad \text{as } n \to \infty. \tag{5.31}$$

In particular, for each fixed t, we have $X_n(t) \xrightarrow{d} B(t)$, as $n \to \infty$.

The latter follows since $n^{-1/2} S_n \xrightarrow{d} B(1)$ by the classical central limit theorem, and so

$$X_n(t) = (\lfloor nt \rfloor / n)^{1/2} \lfloor nt \rfloor^{-1/2} S_{\lfloor nt \rfloor} \xrightarrow{d} t^{1/2} B(1) \stackrel{d}{=} B(t).$$

Similarly, (5.31) follows by a multivariate central limit theorem.

Expression (5.31) only provides a partial description of the convergence in distribution of X_n to B; we will now give a complete description of the convergence that includes sample path information.

Throughout this section, $D = D[0,1]$ will denote the set of all functions $x : [0,1] \to \mathbb{R}$ that are right-continuous with left-hand limits. Assume that the σ-field associated with D is the smallest σ-field under which the projection

map $x \to x(t)$ is measurable, for each t. Almost every sample path of X_n is a function in D, and so the process X_n is a D-valued random variable (or a random element in D).

We will consider D as a metric space in which the distance between two functions x and y is $\|x - y\|$, based on the uniform or supremum norm

$$\|x\| = \sup_{t \le 1} |x(t)|.$$

Other metrics for D are discussed in [11, 115]. Convergence in distribution of random elements in D, as in other metric spaces, is as follows. Random elements X_n in a metric S *convergence in distribution* to X in S as $n \to \infty$, denoted by $X_n \xrightarrow{d} X$ in S, if

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

for any bounded continuous function $f : S \to \mathbb{R}$. The convergence $X_n \xrightarrow{d} X$ is equivalent to the weak convergence of their distributions

$$P\{X_n \in \cdot\} \xrightarrow{w} P\{X \in \cdot\}. \tag{5.32}$$

Several criteria for this convergence are in the Appendix.

An important consequence of $X_n \xrightarrow{d} X$ in S is that it readily leads to the convergence in distribution of a variety of functionals of the X_n as follows.

Theorem 39. (Continuous Mapping) *Suppose that $X_n \xrightarrow{d} X$ in S as $n \to \infty$, and $f : S \to S'$ is a measurable mapping, where S' is another metric space. If $C \subseteq S$ is in the σ-field of S such that f is continuous on C and $X \in C$ a.s., then $f(X_n) \xrightarrow{d} f(X)$ in S' as $n \to \infty$.*

Proof. Recall that $X_n \xrightarrow{d} X$ is equivalent to (5.32), which we will denote by $\mu_n \xrightarrow{w} \mu$. Then $f(X_n) \xrightarrow{d} f(X)$ is equivalent to $\mu_n f^{-1} \xrightarrow{w} \mu f^{-1}$ since

$$P\{f(X_n) \in A\} = P\{X_n \in f^{-1}(A)\} = \mu_n f^{-1}(A).$$

Also note that by Theorem 10 in the Appendix, $\mu_n \xrightarrow{w} \mu$ is equivalent to

$$\liminf_{n\to\infty} \mu_n(G) \ge \mu(G), \qquad \text{for any open } G \subseteq S.$$

Now using this characterization, for any open set $G \subseteq S'$,

$$\liminf_{n\to\infty} \mu_n f^{-1}(G) \ge \liminf_{n\to\infty} \mu_n(f^{-1}(G)^\circ) \ge \mu(f^{-1}(G)^\circ).$$

Here A° is the interior of the set A. Clearly $f^{-1}(G)^\circ \supset C \cap f^{-1}(G)$, and $\mu(C) = 1$ by the assumption $X \in C$ a.s. Then $\mu(f^{-1}(G)^\circ) = \mu f^{-1}(G)$. Using

this in the preceding display yields $\liminf_{n\to\infty} \mu_n f^{-1}(G) \geq \mu f^{-1}(G)$, which proves $\mu_n f^{-1} \xrightarrow{w} \mu f^{-1}$, and hence $f(X_n) \xrightarrow{d} f(X)$.

We are now ready to present the functional central limit theorem proved by Donsker in 1951 for the continuous-time random walk process

$$X_n(t) = n^{-1/2} S_{\lfloor nt \rfloor}, \qquad t \in [0,1].$$

Theorem 40. (Donsker's FCLT) *For the random walk process X_n defined above, $X_n \xrightarrow{d} B$ in D as $n \to \infty$, where B is a standard Brownian motion.*

The proof of this theorem will follow after a few observations and preliminary results. Donsker's theorem is called a "functional central limit theorem" because, under the continuous-mapping theorem, many functionals of the random walk also converge in distribution to the corresponding functionals of the limiting Brownian motion. Two classic examples are as follows; we cover other examples later.

Example 41. If $X_n \xrightarrow{d} B$ in D, then, for $t_1 < \cdots < t_k \leq 1$,

$$(n^{-1/2} S_{\lfloor nt_1 \rfloor}, \ldots, n^{-1/2} S_{\lfloor nt_k \rfloor}) \xrightarrow{d} (B(t_1), \ldots, B(t_k)). \tag{5.33}$$

This convergence is equivalent to (5.31). Now (5.33) says $f(X_n) \xrightarrow{d} f(B)$, where $f : D \to R^k$ is defined, for fixed $t_1 < \cdots < t_k$, by

$$f(x) = (x(t_1), \ldots, x(t_k)).$$

Clearly f is continuous on the set C of continuous functions in D and $B \in C$ a.s. Then (5.33) follows from the continuous-mapping theorem.

Example 42. The convergence $X_n \xrightarrow{d} B$ implies

$$n^{-1/2} \max_{m \leq n} S_m \xrightarrow{d} \max_{s \leq 1} B(s).$$

The distribution of the limit is given in Theorem 11. The convergence follows by the continuous-mapping theorem since the function $f : D \to \mathbb{R}_+$ defined by $f(x) = \max_{s \leq 1} x(s)$ is continuous in that $\|x_n - x\| \to 0$ implies $\max_{s \leq 1} x_n(s) \to \max_{s \leq 1} x(s)$.

Donsker's FCLT is also called an *invariance principle* because in the convergence $X_n \xrightarrow{d} B$, the Brownian motion limit B is the same for "any" distribution of the step size of the random walk, provided it has a finite mean and variance. When the mean and variance are not 0 and 1, respectively, the result applies with the following change in notation.

Remark 43. If ξ_n are i.i.d. random variables with finite mean μ and variance σ^2, then $(\xi_k - \mu)/\sigma$ are i.i.d. with mean 0 and variance 1, and hence Donsker's theorem holds for

$$X_n(t) = n^{-1/2} \sum_{k=1}^{\lfloor nt \rfloor} (\xi_k - \mu)/\sigma, \qquad t \in [0,1].$$

Consequently, the random walk $S_n = \sum_{k=1}^{n} \xi_k$, for large n, is approximately equal in distribution to a Brownian motion with drift. In particular, using $n^{1/2} B(t) \stackrel{d}{=} B(nt)$,

$$S_{\lfloor nt \rfloor} \stackrel{d}{\approx} \mu \lfloor nt \rfloor + \sigma B(nt), \qquad S_n \stackrel{d}{\approx} \mu n + \sigma B(n).$$

Does the convergence in distribution in Donsker's theorem hold for processes defined on the entire time axis $\mathbb{R}_+$? To answer this, consider the space $D[0,T]$ of all functions $x : [0,T] \to \mathbb{R}$ that are right-continuous with left-hand limits, for fixed $T > 0$. Similarly to $D[0,1]$, the $D[0,T]$ is a metric space with the supremum norm. Now let $D(\mathbb{R}_+)$ denote the space of all functions $x : \mathbb{R}_+ \to \mathbb{R}$ that are right-continuous with left-hand limits. Consider $D(\mathbb{R}_+)$ as a metric space in which convergence $x_n \to x$ in $D(\mathbb{R}_+)$ holds if $x_n \to x$ in $D[0,T]$ holds for each T that is a continuity point of x.

Remark 44. Convergence in $D(\mathbb{R}_+)$. Donsker's convergence $X_n \stackrel{d}{\to} B$ holds in $D[0,T]$, for each T, and in $D(\mathbb{R}_+)$ as well. The proof for $D[0,T]$ is exactly the same as that for $D[0,1]$. The convergence also holds in $D(\mathbb{R}_+)$, since B is continuous a.s.

Donsker's approach for proving Theorem 40 is to prove the convergence (5.31) of the finite-dimensional distributions and then establish a certain tightness condition. This proof is described in Billingsley 1967; his book and one by Whitt 2002 cover many fundamentals of functional limit theorems and weak convergence of probability measures on metric spaces.

Another approach for proving Theorem 40, which we will now present, is by applying Skorohod's embedding theorem. The gist of this approach is that one can construct a Brownian motion B and stopping times τ_n for it such that $\{S_n\} \stackrel{d}{=} \{B(\tau_n)\}$. Then further analysis of X_n and B defined on the same probability space establishes $\|X_n - B\| \stackrel{P}{\to} 0$, which yields $X_n \stackrel{d}{\to} B$.

The key embedding theorem for this analysis is as follows. It says that any random variable ξ with mean 0 and variance 1 can be represented as $B(\tau)$ for an appropriately defined stopping time τ. Furthermore, any i.i.d. sequence ξ_n of such variables can be represented as an embedded sequence $B(\tau_n) - B(\tau_{n-1})$ in a Brownian motion B for appropriate stopping times τ_n. The proof is in [37, 61].

Theorem 45. (Skorohod Embedding) *Associated with the random walk S_n, there exists a standard Brownian motion B with respect to a filtration and*

stopping times $0 = \tau_0 \le \tau_1 \le \dots$ *such that* $\tau_n - \tau_{n-1}$ *are i.i.d. with mean* 0 *and variance* 1, *and* $\{S_n\} \stackrel{d}{=} \{B(\tau_n)\}$.

Another preliminary leading to Donsker's theorem is the following Skorohod approximation result that the uniform difference between the random walk and a Brownian motion on $[0,t]$ is $o(t^{1/2})$ a.s. as $t \to \infty$. This material and the proof of Donsker's theorem below is from Kallenberg 2004.

Theorem 46. (Skorohod Approximation of Random Walks) *There exists a standard Brownian motion* B *on the same probability space as the random walk* S_n *such that*

$$t^{-1/2} \sup_{s \le t} |S_{\lfloor s \rfloor} - B(s)| \stackrel{P}{\to} 0, \qquad \text{as } t \to \infty. \tag{5.34}$$

Proof. Let B and τ_n be as in Theorem 45, and define them on the same probability space as S_n (which is possible) so $S_n = B(\tau_n)$ a.s. Define

$$D(t) = t^{-1/2} \sup_{s \le t} |B(\tau_{\lfloor s \rfloor}) - B(s)|.$$

Then (5.34) is equivalent to $P\{D(t) > \varepsilon\} \to 0$ for $\varepsilon > 0$.

To prove this convergence, let $\delta_t = \sup_{s \le t} |\tau_{\lfloor s \rfloor} - s|$, $t \ge 0$. For a fixed $\gamma > 0$, consider the inequality

$$P\{D(t) > \varepsilon\} \le P\{D(t) > \varepsilon,\, t^{-1}\delta_t \le \gamma\} + P\{t^{-1}\delta_t > \gamma\}. \tag{5.35}$$

Note that $n^{-1}\tau_n \to 1$ a.s. by the strong law of large numbers for i.i.d. random variables, and so $t^{-1}|\tau_{\lfloor t \rfloor} - t| \to 0$ a.s. Then the limiting average of the supremum of these differences is $t^{-1}\delta_t \to 0$ a.s. by Proposition 34.

Next, consider the *modulus of continuity* of $f : \mathbb{R}_+ \to \mathbb{R}$, which is

$$w(f,t,\gamma) = \sup_{r,s \le t,\, |r-s| \le \gamma} |f(r) - f(s)|, \qquad t \ge 0.$$

Clearly

$$D(t) \le w(B, t + t\gamma, t\gamma), \qquad \text{when } t^{-1}\delta_t \le \gamma.$$

Using this observation in (5.35) and $\{t^{-1/2}B(r) : r \ge 0\} \stackrel{d}{=} \{B(rt) : r \ge 0\}$ from the scaling property in Exercise 2, we have

$$\begin{aligned} P\{D(t) > \varepsilon\} &\le P\{t^{-1/2} w(B, t + t\gamma, t\gamma) > \varepsilon\} + P\{t^{-1}\delta_t > \gamma\} \\ &= P\{w(B, 1 + \gamma, \gamma) > \varepsilon\} + P\{t^{-1}\delta_t > \gamma\}. \end{aligned}$$

Letting $t \to \infty$ ($t^{-1}\delta_t \to 0$ a.s.), and then letting $\gamma \to 0$ (B has continuous paths a.s.), the last two probabilities tend to 0. Thus $P\{D(t) > \varepsilon\} \to 0$, which proves (5.34).

We will now obtain Donsker's theorem by applying Theorem 46.

Proof of Donsker's Theorem. Let B and $S_n = B(\tau_n)$ a.s. be as in the proof of Theorem 46, and define $B_n(t) = n^{-1/2}B(nt)$. Clearly

$$\|X_n - B_n\| = n^{-1/2} \sup_{t \le 1} |S_{\lfloor nt \rfloor} - B(nt)| = n^{-1/2} \sup_{s \le n} |S_{\lfloor s \rfloor} - B(s)|.$$

Then $\|X_n - B_n\| \xrightarrow{P} 0$ by Theorem 46.

Next, note that, by Exercise 1, the scaled process B_n is a Brownian motion. Now, as in [61], one can construct $\tilde{X}_n$ and a Brownian motion $\tilde{B}$ on the same probability space such that $(\tilde{X}_n, \tilde{B}) \stackrel{d}{=} (X_n, B_n)$. Then we have

$$\|X_n - B\| \stackrel{d}{=} \|\tilde{X}_n - \tilde{B}\| \stackrel{d}{=} \|X_n - B_n\| \xrightarrow{P} 0.$$

This proves $X_n \xrightarrow{d} B$.

5.10 Regenerative and Markov FCLTs

This section presents an extension of Donsker's FCLT for processes with regenerative increments. This in turn yields FCLTs for renewal processes and ergodic Markov chains in discrete and continuous time.

For this discussion, suppose that $\{Z(t) : t \ge 0\}$ is a real-valued process with $Z(0) = 0$ that is defined on the same probability space as a renewal process $N(t)$ whose renewal times are denoted by $0 = T_0 < T_1 < \ldots$ The increments of the two-dimensional process $(N(t), Z(t))$ in the interval $[T_{n-1}, T_n)$ are denoted by

$$\zeta_n = (T_n - T_{n-1}, \{Z(t) - Z(T_{n-1}) : t \in [T_{n-1}, T_n)\}).$$

Recall from Section 2.10 that $Z(t)$ has regenerative increments over T_n if ζ_n are i.i.d.

Theorem 65 in Chapter 2 is a central limit theorem for processes with regenerative increments. An analogous FCLT is as follows. Assuming they are finite, let

$$\mu = E[T_1], \qquad a = E[Z(T_1)]/\mu, \qquad \sigma^2 = \mathrm{Var}[Z(T_1) - aT_1],$$

and assume $\sigma > 0$. In addition, let

$$M_n = \sup_{T_n < t \le T_{n+1}} |Z(t) - Z(T_n)|, \qquad n \ge 0,$$

and assume $E[M_1]$ and $E[T_1^2]$ are finite. For $r > 0$, consider the process

$$X_r(t) = \frac{Z(rt) - art}{\sigma\sqrt{r/\mu}}, \qquad t \in [0,1].$$

This is the regenerative-increment process Z with space-time scale changes analogous to those for random walks. A real-valued parameter r instead of an integer is appropriate since Z is a continuous-time process.

Theorem 47. (Regenerative Increments) *For the normalized regenerative-increment process X_r defined above, $X_r \xrightarrow{d} B$ as $r \to \infty$, where B is a standard Brownian motion.*

The proof uses the next two results. Let D_1 denote the subspace of functions x in D that are nondecreasing with $x(0) = 0$ and $x(t) \uparrow 1$ as $t \to 1$. The *composition mapping* from the product space $D \times D_1$ to D, denoted by $(x, y) \to x \circ y$, is defined by $x \circ y(t) = x(y(t))$, $t \in [0,1]$. Let C and C_1 denote the subspaces of continuous functions in D and D_1, respectively.

Proposition 48. *The composition mapping from $D \times D_1$ to D is continuous on the subspace $C \times C_1$.*

Proof. Suppose $(x_n, y_n) \to (x, y)$ in $D \times D_1$ such that $(x, y) \in C \times C_1$. Using the sup norm and the triangle inequality,

$$\|x_n \circ y_n - x \circ y\| \le \|x_n \circ y_n - x \circ y_n\| + \|x \circ y_n - x \circ y\|.$$

Now, the last term tends to 0 since $x \in C$ is uniformly continuous. Also,

$$\|x_n \circ y_n - x \circ y_n\| = \|x_n - x\| \to 0.$$

Thus $x_n \circ y_n \to x \circ y$ in D, which proves the assertion.

The continuity of composition mappings under weaker assumptions is discussed in [11, 115]. The importance of the composition mapping is illustrated by the following result. In the setting of Theorem 47, the regenerative- increment property of Z implies that

$$\xi_n = Z(T_n) - Z(T_{n-1}) - a(T_n - T_{n-1})$$

are i.i.d. with mean 0 and variance σ^2.

Lemma 49. *Under the preceding assumptions, define the process*

$$X_r'(t) = \frac{1}{\sigma\sqrt{r/\mu}} \sum_{k=1}^{N(rt)} \xi_k, \qquad t \in [0,1].$$

Then $X_r' \xrightarrow{d} B$ as $r \to \infty$.

Proof. Letting $\tilde{X}_r(t) = \frac{1}{\sigma\sqrt{r/\mu}} \sum_{k=1}^{\lfloor rt \rfloor} \xi_k$, it follows by Donsker's theorem that $\tilde{X}_r \xrightarrow{d} \mu^{1/2} B$ as $r \to \infty$. With no loss in generality, assume $\mu^{-1} < 1$. Consider the process

$$Y_r(t) = \begin{cases} N(rt)/r & \text{if } N(r)/r \le \mu^{-1} \\ t/\mu & \text{if } N(r)/r > \mu^{-1}. \end{cases}$$

Note that

$$\tilde{X}_r \circ Y_r(t) = \frac{1}{\sigma\sqrt{r/\mu}} \sum_{k=1}^{\lfloor rY_r(t) \rfloor} \xi_k.$$

This equals $X_r'(t)$ when $N(r)/r \le \mu^{-1}$, and so for any $\varepsilon > 0$,

$$P\{\|X_r' - \tilde{X}_r \circ Y_r\| > \varepsilon\} \le P\{N(r)/r > \mu^{-1}\} \to 0.$$

The convergence follows since $N(r)/r \to \mu^{-1}$ a.s. by the SLLN for renewal processes (Corollary 11 in Chapter 2). This proves $X_r' - \tilde{X}_r \circ Y_r \xrightarrow{d} 0$. Then to prove $X_r' \xrightarrow{d} B$, it suffices by Exercise 53 to show that $\tilde{X}_r \circ Y_r \xrightarrow{d} B$.

Letting $I(t) = t$, $t \in [0,1]$, note that

$$\begin{aligned} \|Y_r - \mu^{-1} I\| &\le \sup_{t \le 1} |N(rt)/r - \mu^{-1} t| \\ &= r^{-1} \sup_{s \le r} |N(s) - \mu^{-1} s| \to 0 \quad \text{a.s.} \end{aligned}$$

The convergence follows by Proposition 34 since the SLLN for N implies $r^{-1}|N(r) - \mu^{-1} r| \to 0$ a.s. Now, we have $(\tilde{X}_r, Y_r) \xrightarrow{d} (\mu^{1/2} B, \mu^{-1} I)$, where the limit functions are continuous. Then Proposition 48 and Exercise 1 yield

$$\tilde{X}_r \circ Y_r \xrightarrow{d} \mu^{1/2} B \circ \mu^{-1} I \stackrel{d}{=} B.$$

Thus $\tilde{X}_r \circ Y_r \xrightarrow{d} B$, which completes the proof.

Remark 50. The assertion in Lemma 49 implies that

$$X_r'(1) = \frac{1}{\sigma\sqrt{r/\mu}} \sum_{k=1}^{N(r)} \xi_k \xrightarrow{d} B(1),$$

which is Anscombe's result in Theorem 64 in Chapter 2.

We now establish the convergence of $X_r(t) = (Z(rt) - art)/(\sigma\sqrt{r/\mu})$.

Proof of Theorem 47. We can write

$$X_r(t) = X_r'(t) + \frac{\sqrt{\mu}}{\sigma} V_r(t), \tag{5.36}$$

where

$$X_r'(t) = \frac{Z(T_{N(rt)}) - aT_{N(rt)}}{\sigma\sqrt{r/\mu}},$$

$$V_r(t) = r^{-1/2}\Big[Z(rt) - Z(T_{N(rt)}) - a(rt - T_{N(rt)})\Big].$$

Recognizing that X_r' is the process in Lemma 49, we have $X_r' \xrightarrow{d} B$. Then the proof of $X_r \xrightarrow{d} B$ will be complete upon showing that $V_r \xrightarrow{d} 0$.

Letting

$$\bar{\xi}_n = \sup_{T_n < t \le T_{n+1}} |Z(t) - Z(T_n)| + a(T_{n+1} - T_n),$$

it follows that

$$\|V_r\| \le r^{-1/2} \sup_{t\le 1} \bar{\xi}_{N(rt)} = \sqrt{N(r)/r}\Big(N(r)^{-1/2} \sup_{k\le N(r)} \bar{\xi}_k\Big).$$

The regenerative-increment property of Z implies that the $\bar{\xi}_n$ are i.i.d. Then

$$n^{-1/2}\bar{\xi}_n \stackrel{d}{=} n^{-1/2}\bar{\xi}_1 \to 0 \quad \text{a.s.}$$

Now $N(r)^{-1/2} \sup_{k\le N(r)} \bar{\xi}_k \xrightarrow{P} 0$ by Proposition 34. Applying this to the preceding display and using $N(r)/r \to \mu^{-1}$ a.s., we get $\|V_r\| \xrightarrow{d} 0$. □

Since renewal processes and ergodic Markov chains are regenerative processes, FCLTs for them are obtainable by Theorem 47. To see this, first note that a renewal process $N(t)$ has regenerative increments over its renewal times T_n, and the parameters above are $M_n = 1$,

$$a = E[N(T_1)]/\mu = \mu^{-1}, \quad \text{Var}[N(T_1) - \mu^{-1}T_1] = \mu^{-2}\text{Var}[T_1].$$

Then the following is an immediate consequence of Theorem 47.

Corollary 51. (Renewal Process) *Suppose $N(t)$ is a renewal process whose inter-renewal times have mean μ and variance σ^2, and define*

$$X_r(t) = \frac{N(rt) - rt/\mu}{\sigma\sqrt{r/\mu^3}}, \qquad t \in [0,1].$$

Then $X_r \xrightarrow{d} B$ as $r \to \infty$.

The particular case $X_r(1) \xrightarrow{d} B(1)$ is the classical central limit theorem for renewal processes, which we saw in Example 67 in Chapter 2; namely

$$\frac{N(r) - r/\mu}{\sigma\sqrt{r/\mu^3}} \xrightarrow{d} B(1).$$

For the next result, suppose that Y is an ergodic CTMC on a countable state space S with stationary distribution p. For a fixed state i, assume that

$Y(0) = i$ and let $0 = T_0 < T_1 < \dots$ denote the times at which Y enters state i. Assume $E_i[T_1^2] < \infty$ and let $\mu = E_i[T_1]$. For $f : S \to \mathbb{R}$, assuming the following integral exists, consider the process

$$Z(t) = \int_0^t f(Y(s))ds, \qquad t \geq 0.$$

This has regenerative increments over the T_n and, assuming the sum is absolutely convergent, Corollary 40 in Chapter 4 yields

$$a = E_i[Z(T_1)]/\mu = \sum_j f(j)p_j.$$

Assume $E_i[M_1]$ and $\sigma^2 = \text{Var}[Z(T_1) - aT_1]$ are finite, and $\sigma > 0$. Then Theorem 47 for the CTMC functional Z is as follows. An analogous result for discrete-time Markov chains is in Exercise 48.

Corollary 52. (CTMC) *Under the preceding assumptions, for $r > 0$, define the process*

$$X_r(t) = \frac{\int_0^{rt} f(Y(s))ds - art}{\sigma\sqrt{r/\mu}}, \qquad t \in [0, 1].$$

Then $X_r \xrightarrow{d} B$ as $r \to \infty$.

5.11 Peculiarities of Brownian Sample Paths

While sample paths of a Brownian motion are continuous a.s., they are extremely erratic. This section describes their erratic behavior.

Continuous functions are typically monotone on certain intervals, but this is not the case for Brownian motion paths.

Proposition 53. *Almost every sample path of a Brownian motion B is monotone on no interval.*

Proof. For any $a < b$ in $\mathbb{R}_+$, consider the event $A = \{B \text{ is nondecreasing on } [a, b]\}$. Clearly $A = \cap_{n=1}^{\infty} A_n$, where

$$A_n = \cap_{i=1}^{n} \{B(t_i) - B(t_{i-1}) \geq 0\}$$

and $t_i = a + i(b - a)/n$. The A is measurable since each A_n is. Because $P\{B(t_i) - B(t_{i-1}) \geq 0\} = 1/2$ and the increments of B are independent, we have $P(A_n) = 2^{-n}$, and so $P(A) \leq \lim_{n\to\infty} P(A_n) = 0$. This conclusion is also true for the event $A = \{B \text{ is nonincreasing on } [a, b]\}$. Thus B is monotone on no interval a.s.

For the next result, we say that for a Brownian motion B on a closed interval I, its *local maximum* is $\sup_{t\in I} B(t)$, and its *local minimum* is $\inf_{t\in I} B(t)$. There are processes that have local maxima on two disjoint intervals that are equal with a positive probability, but this is not the case for Brownian motion.

Proposition 54. *The local maxima and minima of a Brownian motion B are a.s. distinct.*

Proof. It suffices to show that, for disjoint closed intervals I and J in $\mathbb{R}_+$,

$$M_I \neq M_J \quad \text{a.s.},$$

where each of the quantities M_I and M_J is either a local minimum or a local maximum.

First, suppose M_I and M_J are both local maxima. Let u denote the right endpoint of I and $v > u$ denote the left endpoint of J.

$$M_J - M_I = \sup_{t\in J}[B(t) - B(v)] - \sup_{t\in I}[B(t) - B(u)] + B(v) - B(u).$$

The three terms on the right are independent, and the last one is nonzero a.s. (since the increments are normally distributed). Therefore, $M_I \neq M_J$ a.s.

This result is also true by similar arguments when each of the quantities M_I and M_J are both local minima, or when one is a local minimum and the other is a local maximum.

We now answer the question: How much time does a Brownian motion spend in a particular state?

Proposition 55. *The amount of time that a Brownian motion B spends in a fixed state a over the entire time horizon is the Lebesgue measure L_a of the time set $\{t \in \mathbb{R}_+ : B(t) = a\}$, and $L_a = 0$ a.s.*

Proof. Since L_a is nonnegative, it suffices to show $E[L_a] = 0$. For $n \in \mathbb{Z}_+$, consider the process $X_n(t) = B(\lfloor nt \rfloor/n)$, $t \geq 0$. Clearly $X_n(t) \to B(t)$ a.s. as $n \to \infty$ for each t. Then by Fubini's theorem,

$$\begin{aligned} E[L_a] = \int_{\mathbb{R}_+} P\{B(t) = a\}dt &= \int_{\mathbb{R}_+} \lim_{n\to\infty} P\{X_n(t) = a\}dt \\ &\leq \liminf_{n\to\infty} \int_{\mathbb{R}_+} P\{X_n(t) = a\}dt. \end{aligned}$$

The last integral (of a piecewise constant function) is 0 since $X_n(t)$ has a normal distribution, and so $E[L_a] = 0$.

Proofs of the next two results are in [61, 64].

Theorem 56. (Dvoretzky, Erdös, and Kakutani 1961) *Almost every sample path of a Brownian motion B does not have a point of increase: for positive t and δ,*

$$P\{B(s) \le B(t) \le B(u) : (t-\delta)^+ \le s < t < u \le t+\delta\} = 0.$$

Analogously, every sample path of B does not have a point of decrease.

Theorem 57. (Paley, Wiener and Zygmund 1933) *Almost every sample path of a Brownian motion is nowhere differentiable.*

More insights into the wild behavior of a Brownian motion path are given by its linear and quadratic variations. The (linear) *variation* of a real-valued function f on an interval $[a,b]$ is

$$V_a^b(f) = \sup\Big\{\sum_{k=1}^{n} |f(t_k) - f(t_{k-1})| : a = t_0 < t_1 < \cdots < t_n = b\Big\}.$$

If this variation is finite, then f has the following properties:
- It can be expressed as the difference $f(t) = f_1(t) - f_2(t)$ of two increasing functions, where $f_1(t) = V_a^t(f)$.
- The f has a derivative at almost every point in $[a,b]$.
- Riemann-Stieltjes integrals of the form $\int_{[a,b]} g(t)df(t)$ exist.

In light of these observations, Theorem 57 implies that almost every sample path of a Brownian has an "unbounded" variation on any finite interval of positive length. Further insight into the behavior of Brownian paths in terms of their quadratic variation is in Exercise 33.

Because the sample paths of a Brownian motion B have unbounded variation a.s., a stochastic integral $\int_{[a,b]} X(t)dB(t)$ for almost every sample path cannot be defined as a classical Riemann-Stieltjes integral. Another approach is used for defining stochastic integrals with respect to a Brownian motion or with respect to a martingale. Such integrals are the basis of the theory of stochastic differential equations.

5.12 Brownian Bridge Process

We will now study a special Gaussian process called a Brownian bridge. Such a process is equal in distribution to a Brownian motion on $[0,1]$ that is restricted to hit 0 at time 1. An important application is its use in the nonparametric Kolmogorov-Smirnov statistical test that a random sample comes from a specified distribution. In particular, for large samples, the normalized difference between the empirical distribution and the true distribution is approximately the maximum of a Brownian bridge.

Throughout this section $\{X(t) : t \in [0,1]\}$ will denote a stochastic process on $\mathbb{R}$, and $B(t)$ will denote a standard Brownian motion. The process X is a *Brownian bridge* if it is a Gaussian process with mean 0 and covariance function

$$E[X(s)X(t)] = s(1-t), \qquad 0 \le s \le t \le 1.$$

Such a process is equal in distribution to the following Brownian motion "tied down" at 1.

Proposition 58. *The process $X(t) = B(t) - tB(1)$, $t \in [0,1]$, is a Brownian bridge.*

Proof. This follows since X is clearly a Gaussian process with zero mean and

$$\begin{aligned} E[X(s)X(t)] &= E[B(s)B(t) - tB(s)B(1)] - sE[B(1)B(t) - tB(1)^2] \\ &= s(1-t), \qquad s < t. \end{aligned}$$

The last equality uses $E[B(u)B(v)] = u$, for $u \le v$.

Because of its relation to Brownian motion, many basic properties of a Brownian bridge X can be related to those of Brownian motion. For instance, X has continuous paths that are not differentiable. Note that the negation $-X(t)$, and time reversal $X(1-t)$ are also Brownian bridges; related ideas are in Exercises 49 and 50.

We will now show how a Brownian bridge is a fundamental process related to empirical distributions. Suppose that $\xi_1, \xi_2, \ldots$ are i.i.d. random variables with distribution F. The *empirical distribution* associated with $\xi_1, \ldots, \xi_n$ is

$$F_n(t) = n^{-1} \sum_{k=1}^{n} \mathbf{1}(\xi_k \le x), \qquad x \in \mathbb{R},\ n \ge 1.$$

This function is an *estimator* of F based on n samples from it. The estimator is *unbiased* since clearly $E[F_n(x)] = F(x)$. It is also a *consistent* estimator since by the classical SLLN,

$$F_n(x) \to F(x) \qquad \text{a.s. as } n \to \infty. \tag{5.37}$$

This convergence is also uniform in x as follows.

Proposition 59. (Glivenko-Cantelli) *The empirical distributions satisfy*

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad \textit{a.s. as } n \to \infty.$$

Proof. Consider any $-\infty = x_1 < x_2 < \cdots < x_m = \infty$, and note that since F and F_n are nondecreasing, for $x \in [x_{k-1}, x_k]$,

$$F_n(x_{k-1}) - F(x_k) \le F_n(x) - F(x) \le F_n(x_k) - F(x_{k-1}).$$

Then

$$\sup_x |F_n(x) - F(x)| \le \max_k |F_n(x_{k-1}) - F(x_k)| + \max_k |F_n(x_k) - F(x_{k-1})|.$$

Letting $n \to \infty$ and letting the differences $x_k - x_{k-1}$ tend to 0, and then applying (5.37) to the preceding display proves the assertion for continuous F. Exercise 40 proves the assertion when F is not continuous.

An important application of empirical distributions concerns the following nonparametric text that a sample comes from a specified distribution.

Example 60. Kolmorogov-Smirnov Statistic. Suppose that $\xi_1, \xi_2, \ldots$ are i.i.d. random variables with a distribution F that is unknown. As mentioned above, the empirical distribution $F_n(x)$ is a handy unbiased, consistent estimator of F. Now, suppose we want to test the simple hypothesis H_0 that the sample is from a specified distribution F, versus the alternative hypothesis H_1 that the sample is not from this distribution. One approach is to use the classical chi-square test.

Another approach is to use a test based on the *Kolmogorov-Smirnov* statistic defined by

$$D_n = \sum_x |F_n(x) - F(x)|.$$

This is a measure of the distance between the empirical distribution F_n and F (which for simplicity we assume is continuous). The test would reject H_0 if $D_n > c$, and accept it otherwise. The c would be determined by the probability $P\{D_n > c|H_0\} = \alpha$, for a specified *level of significance* α. The conditioning on H_0 means conditioned that F is the true distribution.

When n is large, one can compute c by using the approximation

$$\begin{aligned} P\{n^{1/2} D_n \le x|H_0\} &\approx P\{\sup_{0\le t\le 1} |B(t) - tB(1)| \le x\} \\ &= 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2x^2}. \end{aligned}$$

This approximation follows from Theorem 61 below, and the summation formula is from [37].

We will now establish the limiting distribution of the Kolmogorov-Smirnov statistic.

Theorem 61. *The empirical distribution F_n associated with a sample from the distribution F satisfies*

$$n^{1/2} \sup_x |F_n(x) - F(x)| \xrightarrow{d} \sup_{0\le t\le 1} |X(t)|, \tag{5.38}$$

where X is a Brownian bridge.

Proof. From Exercise 40, we know that $F_n = G_n(F(\cdot))$ and

$$\sup_x |F_n(x) - F(x)| = \sup_{0 \le t \le 1} |G_n(t) - t|,$$

where $G_n(t) = n^{-1} \sum_{k=1}^n \mathbf{1}(U_k \le t)$ is the empirical distribution of the U_n, which are i.i.d. with a uniform distribution on $[0,1]$. The ξ_n and U_n are defined on the same probability space.

In light of this observation, assertion (5.38) is equivalent to

$$n^{-1/2}\|Y_n\| \xrightarrow{d} \|X\|,$$

where $Y_n(t) = \sum_{k=1}^n (\mathbf{1}(U_k \le t) - t)$, $0 \le t \le 1$, and $\|x\| = \sup_{t \le 1} |x(t)|$, for $x \in D$. To prove this convergence, it suffices by the continuous-mapping theorem to show that $n^{-1/2}Y_n \xrightarrow{d} X$ in D, since the map $x \to \|x\|$ from D to D is continuous (in the uniform topology).

Let κ_n be a Poisson random variable with mean n that is independent of the U_k. We will prove $n^{-1/2}Y_n \xrightarrow{d} X$ based on Exercise 53 by verifying

$$n^{-1/2}Y_{\kappa_n} \xrightarrow{d} X, \tag{5.39}$$

$$n^{-1/2}\|Y_n - Y_{\kappa_n}\| \xrightarrow{P} 0. \tag{5.40}$$

Letting $N_n(t) = \sum_{k=1}^{\kappa_n} \mathbf{1}(U_k \le t)$, where $N_n(1) = \kappa_n$, we can write

$$n^{-1/2}Y_{\kappa_n}(t) = n^{-1/2}(N_n(t) - nt) - tn^{-1/2}(N_n(1) - n).$$

Now N_n is a Poisson process on $[0,1]$ with rate n by the mixed-sample representation of Poisson processes in Theorem 26 of Chapter 3. Then from the functional central limit theorem for renewal processes in Corollary 51, the process $n^{-1/2}(N_n(t) - nt)$ converges in distribution in D to a Brownian motion B.

Applying this to the preceding display, it follows that the process $n^{-1/2}Y_{\kappa_n}(t)$ converges in distribution in D to the process $B(t) - tB(1)$, which is a Brownian bridge. This proves (5.39).

Next, note that

$$\begin{aligned} n^{-1/2}\|Y_n - Y_{\kappa_n}\| &\stackrel{d}{=} n^{-1/2} \sup_{0 \le t \le 1} \Big| \sum_{k=1}^{|n-\kappa_n|} (\mathbf{1}(U_k \le t) - t) \Big| \\ &= n^{-1/2}|\kappa_n - n| Z_n, \end{aligned} \tag{5.41}$$

where $Z_n = \sup_{0 \le t \le 1} |G_{|\kappa_n - n|}(t) - t|$. Since κ_n is the sum of n i.i.d. Poisson random variables with mean 1, it follows by the classical central limit theorem that $n^{-1/2}|\kappa_n - n| \xrightarrow{d} |B(1)|$. This convergence also implies $|\kappa_n - n| \xrightarrow{P} \infty$. Now $\sup_{0 \le t \le 1} |G_n(t) - t| \xrightarrow{P} 0$ by Proposition 59 and so this convergence

is also true with n replaced by $|\kappa_n - n|$; that is, $Z_n \xrightarrow{P} 0$. Applying these observations to (5.41) verifies (5.40), which completes the proof.

5.13 Geometric Brownian Motion

This section describes a geometric Brownian and related processes that are used for modeling stock prices or values of investments.

Let $X(t)$ denote the price of a stock (commodity or other financial instrument) at time t. Suppose the value of the stock has many small up and down movements due to continual trading. One possible model is a Brownian motion with drift $X(t) = x + \mu t + \sigma B(t)$. This might be appropriate as a crude model for local or short-time behavior. It is not very good, however, for medium or long term behavior, since the stationary increment property is not realistic (e.g., a change in price for the stock when it is \$50 should be different from the change when the value is \$200).

A more appropriate model for the stock price, which is used in practice, is

$$X(t) = xe^{\mu t + \sigma B(t)}. \tag{5.42}$$

Any process equal in distribution to X is a *geometric Brownian motion* with drift μ and volatility σ. Since $E[e^{\alpha B(t)}] = e^{\alpha^2 t/2}$, the moments of $X(t)$ are given by

$$E[X(t)^k] = x^k e^{k\mu t + k^2 t\sigma^2/2}, \qquad k \geq 1.$$

For instance,

$$E[X(t)] = xe^{\mu t + t\sigma^2/2} = x[1 + (\mu + \sigma^2/2)t] + o(t) \quad \text{as } t \downarrow 0.$$

The X is a diffusion process that satisfies the differential property

$$dX(t) = (\mu + \sigma^2/2)X(t)dt + \sigma X(t)dB(t).$$

We will not prove this characterization, but only note that by the moment formula above, it follows that the instantaneous drift and diffusion parameters for X are

$$\mu(x,t) = (\mu + \sigma^2/2)x, \qquad \sigma(x,t) = \sigma^2 x^2.$$

Although the geometric Brownian motion X does not have stationary independent increments, it does have a nice property of ratios of the increments. In particular, the ratio at the end and beginning of any time interval $[s, s+t]$ is

$$X(t+s)/X(s) = e^{\mu t + \sigma(B(s+t) - B(s))} \stackrel{d}{=} e^{\mu t + \sigma B(t)},$$

so its distribution is independent of s. Also, these ratios over disjoint equal-length time intervals are i.i.d. This means that as a model for a stock price,

one cannot anticipate any upward or downward movements in the price "ratios". So in this sense, the market is equitable (or not biased).

Does this also mean that the market is fair in the martingale sense that $X(t)$ is a martingale with respect to B? The answer is generally no.

However, X is a martingale if and only if $\mu + \sigma^2/2 = 0$ (a very special condition). This follows since $e^{-t(\mu+\sigma^2/2)}X(t)$ is a martingale with respect to B with mean x by Example 19 (and $E[X(t)] = x$ when $X(t)$ is such a martingale).

The geometric Brownian model (5.42) has continuous paths that do not account for large discrete-jumps in stock prices. To incorporate such jumps, another useful model is as follows.

Example 62. Prices with Jumps. Suppose the price of a stock at time t is given by $X(t) = e^{Y(t)}$, where $Y(t)$ is a real-valued stochastic process with stationary independent increments (e.g., a compound Poisson or Lévy process). These properties of Y also ensure that the price ratios are i.i.d. in disjoint, fixed-length intervals.

Assume as in Exercise 7 that the moment generating function $\psi(\alpha) = E[e^{\alpha Y(1)}]$ exists for α in a neighborhood of 0, and $E[e^{\alpha Y(t)}]$ is continuous at $t = 0$ for each α. Then it follows that

$$E[X(t)^k] = \psi(k)^t, \qquad k \geq 1.$$

In particular, if $Y(t)$ is a compound Poisson process with rate λ and its jumps have the moment generating function $G(\alpha)$, then $\psi(\alpha) = e^{-\lambda(1-G(\alpha))}$. Consequently,

$$E[X(t)^k] = e^{-\lambda t(1-G(k))}, \qquad k \geq 1.$$

Other possibilities are that Y is a sum of a Brownian motion plus an independent compound Poisson process, or that X is the sum of a geometric Brownian motion plus an independent compound Poisson process.

We will not get into advanced investment models using geometric Brownian motion such as Black-Scholes option pricing. However, the following illustrates an elementary computation for an option.

Example 63. Stock Option. Suppose that the price of one unit of a stock at time t is given by a geometric Brownian motion $X(t) = e^{B(t)}$. A customer has the option of buying one unit of the stock at a fixed time T at a price K, but the customer need not make the purchase. The value of the option to the customer is $(X(t) - K)^+$ since the customer will not buy the stock if $X(t) < K$. We will disregard any fee that the customer would pay in order to obtain the option.

The expectation of the option's value is

$$\begin{aligned} E[(X(T) - K)^+] &= \int_0^\infty P\{X(T) - K > x\}\, dx \\ &= \int_0^\infty P\{B(T) > \log(x + K)\}\, dx. \end{aligned}$$

This integral can be integrated numerically by using an approximation for the normal distribution of $B(T)$. A variation of this option is in Exercise 39.

5.14 Multidimensional Brownian Motion

Brownian motions in the plane and in multidimensional spaces are natural models for phenomena driven by several independent (or dependent) single-dimension Brownian motions. This section gives some insight into these multidimensional processes.

A stochastic process $B(t) = (B_1(t), \ldots, B_d(t))$, $t \geq 0$, in $\mathbb{R}^d$ is a *multidimensional Brownian motion* if $B_1, \ldots, B_d$ are independent Brownian motions on $\mathbb{R}$. Many basic properties of this process follow from results in one dimension. For instance, the multidimensional integral formula

$$\int_{\mathbb{R}^d} P\{x + B(t) \in A\} dx = |A|,$$

the Lebesgue measure of A, follows from the similar formula for $d = 1$ in Exercise 6. The preceding integral is used in Section 5.15 for particle systems.

Applications of Brownian motions in $\mathbb{R}^d$ typically involve intricate functions of the single-dimension components whose distributions determine system parameters (e.g., Exercise 54). Here is another classical application.

Example 64. Bessel Processes. Associated with a Brownian motion $B(t)$ in $\mathbb{R}^d$, consider its radial distance to the origin defined by

$$R(t) = \left(B_1(t)^2 + \cdots + B_d(t)^2\right)^{1/2}, \qquad t \geq 0.$$

Any process equal in distribution to R is a *Bessel process of order d*.

When $d = 1$, we have the familiar reflected Brownian motion process $R(t) = |B(t)|$. Exercise 19 mentioned that this is a Markov process and it specifies its distribution (also recall Theorem 11).

The $R(t)$ is also a Markov process on $\mathbb{R}$ for general d. Its transition probability $P\{R(t) \in A | R(0) = x\} = \int_A p^t(x, y)\, dy$ has the density

$$p^t(x, y) = t^{-1}(xy)^{1-d/2} y^{d-1} I_{d/2-1}(xy/t),$$

where I_β is the modified Bessel function of order $\beta > -1$ defined by

$$I_\beta(u) = \sum_{k=0}^{\infty} \frac{(u/2)^{2k+\beta}}{k! \Gamma(k + \beta + 1)}, \qquad u \in \mathbb{R}.$$

This is proved in [61]. We will only derive the density of $R(t)$ when $R(0) = 0$.

To this end, consider

$$R(t)^2/t = (B_1(t)^2 + \cdots + B_d(t)^2)/t \stackrel{d}{=} B_1(1)^2 + \cdots + B_d(1)^2.$$

The last sum of squares of d independent standard normal random variables is known to have a χ-squared density f with d degrees of freedom. This f is a gamma density with parameters $\alpha = d/2$ and $\lambda = 1/2$ (see the Appendix). Therefore, knowing that $R(0) = 0$,

$$P\{R(t) \le r\} = P\{R(t)^2/t \le r^2/t\} = \int_0^{r^2/t} f(x)\,dx. \tag{5.43}$$

The density of $R(t)$ is shown in Exercise 55.

Although the hitting times of $R(t)$ are complicated, we can evaluate their means. Consider the time $\tau_a = \inf\{t \ge 0 : R(t) = a\}$ to hit a value $a > 0$. This is a stopping time of $R(t)$ and $\tau_a \le \inf\{t \ge 0 : |B_1(t)| = a\} < \infty$ a.s. since the last stopping time is finite a.s. as noted in Theorem 11. Now, Exercise 56 shows that $R(t)^2 - t$ is a martingale with mean 0. Then the optional stopping result in Corollary 24 yields $E[R(\tau_a)^2 - \tau_a] = 0$. Therefore $E[\tau_a] = a^2$.

We will now consider a multidimensional process whose components are "dependent" one-dimensional Brownian motions with drift. Let $B(t)$ be a Brownian motion in $\mathbb{R}^d$, and let $C = \{c_{ij}\}$ be a $d \times d$ matrix of nonnegative real numbers that are symmetric ($c_{ij} = c_{ji}$) and nonnegative-definite ($\sum_i \sum_j u_i u_j c_{ij} \ge 0$, for $u \in \mathbb{R}^d$). As in the representation (5.8) of a multivariate normal vector, let A be a $k \times d$ matrix with transpose A^t and $k \le d$ such that $A^t A = C$. Consider the process $\{X(t) : t \ge 0\}$ in $\mathbb{R}^d$ defined by

$$X(t) = x + \mu t + B(t)A,$$

where x and μ are in $\mathbb{R}^d$.

Any process equal in distribution to X is a *generalized Brownian motion* in $\mathbb{R}^d$ with initial value x, drift μ and covariance matrix $C = A^t A$.

A major use for multidimensional Brownian motions is in approximating multidimensional random walks. The following result is an analogue of Donsker's Brownian motion approximation for one-dimensional random walks in Theorem 40.

Suppose that ξ_k, $k \ge 1$, are i.i.d. random vectors in $\mathbb{R}^d$ with mean vector $\mu = (\mu_1, \ldots, \mu_d)$ and covariances $c_{ij} = E[(\xi_{k,i} - \mu_i)(\xi_{k,j} - \mu_j)]$, $1 \le i, j \le d$. Define the processes $\{X_n(t) : t \ge 0\}$ in $\mathbb{R}^d$, for $n \ge 1$, by

$$X_n(t) = n^{-1/2} \sum_{k=1}^{\lfloor nt \rfloor} (\xi_k - \mu), \qquad t \ge 0.$$

Theorem 65. *Under the preceding assumptions, $X_n \stackrel{d}{\to} X$, as $n \to \infty$, where X is a generalized Brownian motion on $\mathbb{R}^d$ starting at 0, with no drift, and with covariance matrix $\{c_{ij}\}$.*

Sketch of Proof. Consider $X_{n,i}(t) = n^{-1/2} \sum_{k=1}^{\lfloor nt \rfloor} (\xi_{k,i} - \mu_i)$, which is the ith component of X_n. By Donsker's theorem, $X_{n,i} \xrightarrow{d} X_i$ for each i. Now, the Cramér-Wold theorem states that $(X_{n,1}, \ldots, X_{n,d}) \xrightarrow{d} (X_1, \ldots, X_d)$ in $\mathbb{R}^d$ if and only if $\sum_{i=1}^d a_i X_{n,i} \xrightarrow{d} \sum_{i=1}^d a_i X_i$ in $\mathbb{R}$ for any $a \in \mathbb{R}^d$. However, the latter holds by another application of Donsker's theorem. Therefore, the finite-dimensional distributions of X_n converge to those of X. To complete the proof that $X_n \xrightarrow{d} X$, it suffices to verify a certain tightness condition on the distributions of the processes X_n, which we omit.

5.15 Brownian/Poisson Particle System

This section describes a system in which particles occasionally enter an Euclidean space and move about independently as Brownian motions and eventually exit. The system data and dynamics are represented by a marked Poisson process like those in Chapter 2. The focus is on characterizing certain Poisson processes describing particle locations over time and departures as intricate functions of the arrival process and particle trajectories. The Brownian motion structure of the trajectories lead to tractable probabilities.

Consider a system of discrete particles that move about in the space $\mathbb{R}^d$ as follows. The locations and entry times of the particles are represented by the space-time Poisson process $N = \sum_n \delta_{(X_n, T_n)}$ on $\mathbb{R}^d \times \mathbb{R}$, where X_n is the location in $\mathbb{R}^d$ at which the nth particle enters at time T_n. This Poisson process is homogeneous in that

$$E[N(A \times I)] = \alpha |A| \lambda |I|,$$

where $|A|$ is the Lebesgue measure of the Borel set A. Here λ is the arrival rate of particles per unit time in any unit area, and α is the arrival rate per unit area in any unit time period. Note that, for bounded sets A and I, the $N(A \times I)$ is finite, but $N(\mathbb{R}^d \times I)$ and $N(A \times \mathbb{R})$ are infinite a.s. (because their Poisson means are infinite).

We assume that each particle moves in $\mathbb{R}^d$ independently as a d-dimensional Brownian motion $B(t)$, $t \geq 0$, for a length of time V with distribution G and then exits the space.

More precisely, let V_n, $n \in \mathbb{Z}$, be independent with $V_n \stackrel{d}{=} V$, and let B_n, $n \in \mathbb{Z}$, be independent with $B_n \stackrel{d}{=} B$. Assume $\{B_n\}$, $\{V_n\}$ are independent and independent of N. Viewing the B_n and V_n as independent marks of (X_n, T_n), the data for the entire system is defined formally by the marked Poisson process

$$M = \sum_n \delta(X_n, T_n, B_n, V_n), \qquad \text{on } S = \mathbb{R}^d \times \mathbb{R} \times C(\mathbb{R}, \mathbb{R}^d) \times \mathbb{R}_+.$$

Here $C(\mathbb{R}, \mathbb{R}^d)$ denotes the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}^d$. The mean measure of M is given by

$$E[M(A \times I \times F \times J)] = \alpha|A|\lambda|I|P\{B \in F\}P\{V \in J\}.$$

The interpretation is that the nth particle has a space-time entry at (X_n, T_n) and its location at time t is given by $X_n + B_n(t - T_n)$, where $t - T_n \le V_n$. At the end of its sojourn time V_n it exits the system at time $T_n + V_n$ from the location $X_n + B_n(V_n)$.

Let us see where the particles are at any time t. It is not feasible to account for all the particles that arrive up to time t, which is $N(\mathbb{R}^d \times (-\infty, t]) = \infty$. So we will consider particles that enter in a bounded time interval I prior to t, which is $t - I$ (e.g., $t - [a, b] = [t - b, t - a]$).

Now, the number of particles that enter $\mathbb{R}^d$ in a time interval I prior to t and are in A at time t is

$$N_t(I \times A) = \sum_n \delta_{(T_n, X_n + B_n(t - T_n))}(I \times A)\mathbf{1}(V_n > t - T_n).$$

The N_t is a point process on $\mathbb{R}_+ \times \mathbb{R}^d$.

Proposition 66. *The family of point processes $\{N_t : t \in \mathbb{R}\}$ is stationary in t, and each N_t is a Poisson process on $\mathbb{R}_+ \times \mathbb{R}^d$ with mean measure*

$$E[N_t(I \times A)] = \alpha\lambda|A| \int_I (1 - G(u))du. \tag{5.44}$$

Proof. By the form of its mean measure, the Poisson process M with its time axis shifted by an amount t is

$$\mathrm{S}^t M = \sum_n \delta_{(X_n, T_n + t, B_n, V_n)} \stackrel{d}{=} M, \qquad t \in \mathbb{R}.$$

Therefore, M is stationary in the time axis. To prove that N_t is stationary in t, it suffices by Proposition 104 in Chapter 3 to show that $N_t = f(\mathrm{S}^t M)$, for some function f.

Accordingly, for a locally-finite counting measure $\nu = \sum_n \delta_{(x_n, t_n, b_n, v_n)}$ on S, define the counting measure $f(\nu)$ on $\mathbb{R}_+ \times \mathbb{R}^d$ by

$$f(\nu) = \sum_n \delta_{(-t_n, x_n + b_n(-t_n))}\mathbf{1}(v_n > -t_n).$$

Then clearly, $N_t = f(\mathrm{S}^t M)$, which proves that N_t is stationary.

Next, note that N_t is a deterministic map of the Poisson process M restricted to the subspace $\{(x, s, b, v) \in S : s \le t, v > t - s\}$, in which any point (x, s, b, v) in the subspace is mapped to $(s, x + b(t - s))$. Then by Theorem 32 in Chapter 2, N_t is a Poisson process with mean measure given by

$$E[N_t(I \times A)] = \alpha\lambda \int_{t-I} \Big(\int_{\mathbb{R}^d} P\{x + B(t-s) \in A\}\, dx \Big) P\{V > t-s\} ds.$$

Because B is a Brownian motion, the integral in parentheses reduces to $|A|$ by Exercise 6. Therefore, using the change of variable $u = t - s$ in the last expression yields (5.44).

Next, let us consider departures from the system. The number of particles that enter $\mathbb{R}^d$ during the time set I and depart from A during the time set J is $\bar{N}(I \times A \times J)$, where $\bar{N}$ is a point process of the form

$$\bar{N} = \sum_n \delta_{(T_n, X_n + B_n(V_n), T_n + V_n)} \quad \text{on } \{(s,x,u) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} : s \le u\}.$$

Proposition 67. *The point process of departures $\bar{N}$ is a Poisson process with mean measure given by*

$$E[\bar{N}(I \times A \times J)] = \alpha\lambda |A| \int_{\mathbb{R}_+} |I \cap (J - v)| dG(v). \tag{5.45}$$

Proof. By its definition, $\bar{N}$ is a deterministic map g of the Poisson process M, where $g(X_n, T_n, B_n, V_n) = (T_n, X_n + B_n(V_n), T_n + V_n)$. Then by Theorem 32 in Chapter 2, $\bar{N}$ is a Poisson process with mean

$$E[\bar{N}(I \times A \times J)] = \alpha\lambda \int_I \int_{\mathbb{R}_+} \mathbf{1}(s+v \in J) \Big(\int_{\mathbb{R}^d} P\{x + B(v) \in A\}\, dx \Big) dG(v)\, ds.$$

The integral in parentheses reduces to $|A|$ by Exercise 6. Then an interchange of the order of integration in the last expression yields (5.45).

There are several natural generalizations of the preceding model with more dependencies among the marks and entry times and points, e.g., see Exercise 51. Although the processes N_t and $\bar{N}$ may still be Poisson, their mean values would be more complicated.

5.16 $G/G/1$ Queues in Heavy Traffic

Section 4.20 of Chapter 4 showed that the waiting times W_n for successive items in a $G/G/1$ queueing system are a function of a random walk. This suggests that the asymptotic behavior of these times can be characterized by the Donsker Brownian motion approximation of a random walk, and that is what we shall do now. We first describe the limit of W_n when the traffic intensity $\rho = 1$, and then present a more general FCLT for the W_n when the system is in *heavy traffic*: the traffic intensity is approximately 1.

Consider a $G/G/1$ queueing system, as in Section 4.20 of Chapter 4, in which items arrive at times that form a renewal process with inter-arrival

times U_n, and the service times are i.i.d. nonnegative random variables V_n that are independent of the arrival times. The service discipline is first-come-first-served with no preemptions. The inter-arrival and service times have finite means and variances, and the traffic intensity of the system is $\rho = E[V_1]/E[U_1]$. For simplicity, assume the system is empty at time 0.

Our interest is in the length of time W_n that the nth arrival waits in the queue before being processed. Section 4.20 of Chapter 4 showed that these waiting times satisfy the *Lindley* recursive equation

$$W_n = (W_{n-1} + V_{n-1} - U_n)^+, \qquad n \geq 1,$$

and consequently,

$$W_n = \max_{0 \leq m \leq n} \sum_{k=m+1}^{n} (V_{k-1} - U_k). \tag{5.46}$$

Under the assumptions on the U_n and V_n, it follows that

$$W_n \stackrel{d}{=} \max_{0 \leq m \leq n} S_m, \tag{5.47}$$

where $S_n = \sum_{m=1}^{n} \xi_m$ and $\xi_m = V_m - U_m$.

In case $\rho < 1$, Theorem 118 of Chapter 4 noted that

$$W_n \stackrel{d}{\to} \max_{0 \leq m < \infty} S_m.$$

In this section, we consider the limiting behavior of the waiting times W_n when ρ equals or approaches 1, meaning that the system is in *heavy traffic*.

We begin with the case $\rho = 1$, and describe the asymptotic behavior of the waiting times via the process

$$\widehat{W}_n(t) = \frac{W_{\lfloor nt \rfloor}}{\sigma\sqrt{n}}, \qquad t \geq 0.$$

Theorem 68. *Suppose the $G/G/1$ system defined above has $\rho = 1$ and $\sigma^2 = \mathrm{Var}(\xi_1) > 0$. Then*

$$\widehat{W}_n \stackrel{d}{\to} M \quad \text{in } D(\mathbb{R}_+) \text{ as } n \to \infty,$$

where $M(t) = \max_{s \leq t} B(s)$, the maximum process for a standard Brownian motion B. Hence

$$\frac{W_n}{\sigma\sqrt{n}} \stackrel{d}{\to} M(1) \stackrel{d}{=} |B(1)|.$$

Proof. Note that

$$\widehat{W}_n(t) \stackrel{d}{=} \frac{1}{\sigma\sqrt{n}} \max_{m \le \lfloor nt \rfloor} S_m = \frac{1}{\sigma\sqrt{n}} \sup_{s \le t} S_{[ns]}.$$

That is, $\widehat{W}_n(t) \stackrel{d}{=} f(X_n)(t)$, $t \ge 0$, where

$$X_n(t) = \frac{S_{\lfloor nt \rfloor}}{\sigma\sqrt{n}}, \qquad t \ge 0,$$

and $f : D(\mathbb{R}_+) \to D(\mathbb{R}_+)$ is the *supremum map* defined by

$$f(x)(t) = \sup_{0 \le s \le t} x(s), \qquad x \in D(\mathbb{R}_+).$$

Now the random walk S_n has steps with mean $E[\xi_1] = 0$, since $\rho = 1$; and $\sigma^2 = \mathrm{Var}(\xi_1)$. Then $X_n \stackrel{d}{\to} B$ by Donsker's theorem.

Next, it is clear that if $\|x_n - x\| \to 0$ in $D[0,T]$, then

$$\|f(x_n) - f(x)\| \le \|x_n - x\| \to 0 \qquad \text{in } D[0,T].$$

Then since $\|X_n - B\| \stackrel{P}{\to} 0$ in $D[0,T]$ for each T, it follows that

$$\|f(X_n) - f(B)\| \stackrel{P}{\to} 0 \quad \text{in } D(\mathbb{R}_+).$$

This along with $\widehat{W}_n = f(X_n)$ and $f(B) = M$ proves $\widehat{W}_n \stackrel{d}{\to} M$.

In particular, $\widehat{W}_n(1) \stackrel{d}{\to} M(1) \stackrel{d}{=} |B(1)|$, which proves the second assertion; that $M(1) \stackrel{d}{=} |B(1)|$ follows by Theorem 11.

The preceding result suggests that for any $G/G/1$ system in which $\rho \approx 1$, the approximation $\widehat{W}_n \stackrel{d}{\approx} M$ would be valid. A formal statement to this effect is as follows.

Consider a family of $G/G/1$ systems indexed by a parameter r with interarrival times U_n^r and service times V_n^r. Denote the other quantities by ρ_r, W_n^r, $S_n^r = \sum_{m=1}^n \xi_m^r$, etc., and consider the process

$$\widehat{W}_r(t) = \frac{W^r_{\lfloor rt \rfloor}}{\sigma\sqrt{r}}, \qquad t \ge 0.$$

Theorem 69. *Suppose the family of $G/G/1$ systems are such that $\rho_r \to 1$,*

$$\sup_r E[(\xi_1^r - E[\xi_1^r])^{2+\varepsilon}] < \infty, \quad \textit{for some } \varepsilon > 0,$$

$$r^{1/2} E[\xi_1^r] \to 0, \qquad \mathrm{Var}(\xi_1^r) \to \sigma^2 > 0, \qquad \textit{as } r \to \infty.$$

Then $\widehat{W}_r \stackrel{d}{\to} M$ in $D(\mathbb{R}_+)$ as $r \to \infty$.

Proof. As in the proof of Theorem 68, $\widehat{W}_r = f(X_r)$, where f is the supremum map and $X_r(t) = S_{\lfloor rt \rfloor}/\sigma\sqrt{r}$, $t \geq 0$. Then to prove the assertion, it suffices to show that $X_r \xrightarrow{d} B$ as $r \to \infty$.

Now, we can write

$$X_r(t) = Y_r(t) + (\lfloor rt \rfloor / r) r^{1/2} E[\xi_1^r],$$

where

$$Y_r(t) = \frac{1}{\sigma\sqrt{r}} \sum_{m=1}^{\lfloor rt \rfloor} (\xi_m^r - E[\xi_1^r]).$$

Under the hypotheses, $Y_r \xrightarrow{d} B$ by a theorem of Prokhorov 1956, and hence $X_r \xrightarrow{d} B$.

The preceding results are typical of many heavy-traffic limit theorems that one can obtain for queueing and related processes by the framework presented by Whitt [115]. In particular, when a system parameter, such as the waiting time above, can be expressed as a function of the system data (cumulative input and output processes), and that data under an appropriate normalization converges in distribution, then under further technical conditions, the system parameter also converges in distribution to the function of the limits of the data. Here is one of the general models in [115].

Example 70. Generalized $G/G/1$ System. Consider a generalization of the $G/G/1$ systems above in which the inter-arrival times U_n^r and service times V_n^r (the system data) are general random variables that may be dependent. Then the waiting times W_n^r can still be expressed as a function of the system data as in (5.46). In other words,

$$W_n^r = \tilde{S}_n - \min_{0 \leq m \leq n} \tilde{S}_m$$

where $\tilde{S}_n = \sum_{k=1}^n (V_{k-1} - U_k)$. As above, consider the processes

$$\widehat{W}_r(t) = \frac{W_{\lfloor rt \rfloor}^r}{\sigma\sqrt{r}}, \qquad X_r(t) = \frac{S_{\lfloor rt \rfloor}}{\sigma\sqrt{r}}, \qquad t \geq 0.$$

Then we can write $\widehat{W}_r = h(X_r)$, where $h : D(\mathbb{R}) \to D(\mathbb{R})$ is the *one-sided reflection* map defined by

$$h(x)(t) = x(t) - \inf_{0 \leq s \leq t} x(s), \qquad t \geq 0.$$

The reflection map h (like the supremum map above) is continuous in the uniform topology on $D[0, T]$ since

$$\|h(x) - h(y)\| \leq 2\|x - y\|, \qquad x, y \in D[0, T].$$

Then the continuous-mapping theorem yields the following result.

Convergence Criterion. If $X_r \xrightarrow{r} X$ in $D(\mathbb{R})$, then $\widehat{W}_r \xrightarrow{d} W$ in $D(\mathbb{R})$ as $r \to \infty$, where

$$W(t) = X(t) - \inf_{0 \le s \le t} X(s), \quad t \ge 0.$$

To apply this for a particular situation, one would use properties of the inter-arrival times and service times (as in Theorem 69) to verify $X_r \xrightarrow{r} X$. There are a variety of conditions under which the limit X is a Brownian motion, a process with stationary independent increments, or an infinitely divisible process; and other topologies on $D(\mathbb{R}_+)$ are often appropriate [115].

5.17 Brownian Motion in a Random Environment

Section 3.14 describes a Poisson process with a random intensity measure called a Cox process. The random intensity might represent a random environment or field that influences the locations of points. This section describes an analogous randomization for Brownian motions in which the time scale is determined by a stochastic process.

Let $\{X(t) : t \in \mathbb{R}_+\}$ and $\eta = \{\eta(t) : t \in \mathbb{R}_+\}$ be real-valued stochastic processes defined on the same probability space, such that $\eta(t)$ is a.s. nondecreasing with $\eta(0) = 0$ and $\eta(t) \to \infty$ a.s. as $t \to \infty$. The process X is a *Brownian motion directed by* η if the increments of X are conditionally independent given η, and, for any $s < t$, the increment $X(t) - X(s)$ has a conditional normal distribution with variance $\tau_t - \tau_s$. These conditions, in terms of the moment generating function for the increments of X, say that, for $0 = t_0 < t_1 < \cdots < t_n$ and $u_1, \ldots, u_n$ in $\mathbb{R}_+$,

$$E\Big[\exp\Big\{\sum_{i=1}^{n} u_i[X(t_i) - X(t_{i-1})]\Big\} \,\Big|\, \eta\Big] \tag{5.48}$$
$$= \exp\Big\{\frac{1}{2}\sum_{i=1}^{n} u_i^2[\eta(t_i) - \eta(t_{i-1})]\Big\} \quad \text{a.s.}$$

A directed Brownian motion is equal in distribution to a standard Brownian motion with random time parameter as follows.

Remark 71. A process X is a Brownian motion directed by η if and only if $X \stackrel{d}{=} B \circ \eta'$, where B and η' are defined on a common probability space such that B is a standard Brownian motion independent of η', and $\eta' \stackrel{d}{=} \eta$. This follows by the definition above and consideration of the moment generating function of the increments of the processes. The process $B \circ \eta'$ is a

Brownian motion subordinated to η (like a Markov chain subordinated to a Poisson process, which we saw in Chapter 4). In case η is strictly increasing, Exercise 43 shows that $X = B \circ \eta$ a.s., where B is defined on the same probability space as X and η.

A Brownian motion X directed by η inherits many properties of standard Brownian motions. The proofs usually follow by conditioning on η and using properties of this process. Here are some examples.

Example 72. $E[X(t)] = 0$, and $\mathrm{Var}[X(t)] = E[\eta(t)]$.

Example 73. Consider $\tau_a^X = \inf\{t : X(t) \geq a\}$. Then

$$P\{\tau_a^X \leq t\} = \int_{\mathbb{R}_+} P\{\eta(t) \leq u\} P\{\tau_a \in du\},$$

where $\tau_a = \inf\{t : B(t) = a\}$.

Example 74. Suppose that $X_1, \ldots, X_m$ are Brownian motions directed by $\eta_1, \ldots, \eta_m$, respectively, and $(X_1, \eta_1), \ldots, (X_m, \eta_m)$ are independent. Then $X(t) = X_1(t) + \cdots + X_m(t)$ is a Brownian motion directed by $\eta(t) = \eta_1(t) + \cdots \eta_m(t)$.

Example 75. FCLT. For a Brownian motion X directed by η, define

$$X_r(t) = b_r^{-1/2} X(rt) \qquad t \geq 0,$$

where $b_r \to \infty$ are constants. By Remark 71 and the scaling $b_r^{-1/2} B \stackrel{d}{=} B(b_r \cdot)$, we can write $X_r \stackrel{d}{=} B \circ Y_r$, where $Y_r(t) = \eta'(rt)/b_r$ and the Brownian motion B and η' are independent. Then by the property of the composition mapping in Proposition 48, we obtain the following result, where I is the identity function: If $Y_r \stackrel{d}{\to} I$ in $D(\mathbb{R}_+)$, then $X_r \stackrel{d}{\to} B$ in $D(\mathbb{R}_+)$.

5.18 Exercises

For the following exercises, B will denote a standard Brownian motion.

Exercise 1. Show that each of the following processes is also a Brownian motion.

(a) $B(t+s) - B(s)$ *Translated* process for a fixed time s.
(b) $-B(t)$ *Reflected* process
(c) $c^{-1/2} B(ct)$ *Scaling*, for $c > 0$
(d) $B(T) - B(T-t)$, $t \in [0, T]$, *Time-reversal* on $[0, T]$ for fixed T.

Exercise 2. The *time-inversion* of a Brownian motion B is the process $X(0) = 0$,

$$X(t) = tB(1/t), \qquad t > 0.$$

Prove that X is a Brownian motion. First show that $X(t) \to 0$ a.s. as $t \downarrow 0$.

Exercise 3. Suppose that $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous, strictly increasing function with $h(0) = 0$, and $h(t) \uparrow \infty$. Find the mean and covariance functions for the process $B(h(t))$. Show that $B(h(t)) \stackrel{d}{=} X(t)$ for each t, where $X(t) = (h(t)/t)^{-1/2}B(t)$. Are the processes $B(h(\cdot))$ and X equal in distribution?

Exercise 4. For $0 < s < t$, show that the conditional density of $B(s)$ given $B(t) = b$ is normal with conditional mean and variance

$$E[B(s)|B(t) = b] = bs/t, \quad \text{Var}[B(s)|B(t) = b] = s(t-s)/t.$$

For $t_1 < s < t_2$, show that

$$P\{B(s) \le x|B(t_1) = a,\ B(t_2) = b\} = P\{B(s-t_1) \le x-a|B(t_2-t_1) = b-a\}.$$

Using these properties prove that the conditional density of $B(s)$ given $B(t_1) = a$, $B(t_2) = b$ is normal with conditional mean and variance

$$\frac{a + (b-a)(s-t_1)}{(t_2-t_1)} \quad \text{and} \quad \frac{(s-t_1)(t_2-s)}{(t_2-t_1)}.$$

Exercise 5. Consider the process $X(t) = ae^{-\alpha t} + \sigma^2 B(t)$, $t \ge 0$, where $a, \alpha \in \mathbb{R}$ and $\sigma > 0$. Find the mean and covariance functions for this process. Show that X is a Gaussian process by applying Theorem 5. Does X have independent increments, and are these increments stationary? Is X a martingale, submartingale or supermartingale with respect to B?

Exercise 6. For a density f on $\mathbb{R}$ that is symmetric ($f(x) = f(-x)$, $x \in \mathbb{R}$), show that

$$\int_{\mathbb{R}} \Big(\int_{A-x} f(y)dy \Big) dx = |A| \qquad \text{(The Lebesgue measure of } A\text{)}.$$

Use this to verify that, for a Brownian motion $B(t)$,

$$\int_{\mathbb{R}} P\{x + \mu t + \sigma B(t) \in A\} dx = |A|,$$

independent of t, μ and σ.

Exercise 7. Let $Y(t)$ be a real-valued process with stationary independent increments. Assume that the moment generating function $\psi(\alpha) = E[e^{\alpha Y(1)}]$ exists for α in a neighborhood of 0, and $g(t) = E[e^{\alpha Y(t)}]$ is continuous at $t = 0$ for each α. Show that $g(t)$ is continuous in t for each α, and that

$g(t) = \psi(\alpha)^t$. Use the fact that $g(t+u) = g(t)g(u)$, $t, u \geq 0$, and that the only continuous solution to this equation has the form $g(t) = e^{tc}$, for some c that depends on α (because $g(t)$ depends on α). Show that

$$E[Y(t)] = tE[Y(1)], \qquad \text{Var}[Y(t)] = t\text{Var}[Y(1)].$$

Exercise 8. Let $X(t) = \mu t + \sigma B(t)$ be a Brownian motion with drift, where $\mu > 0$. Suppose that a signal is triggered whenever $M(t) = \sup_{s \leq t} X(s)$ reaches the levels $0, 1, 2, \ldots$. So the nth signal is triggered at time $\tau_n = \inf\{t : M(t) = n\}$, $n \geq 0$. Show that these times of the signals form a renewal process and find the mean, variance and Laplace transform of the times between signals.

In addition, obtain this information under the variation in which a signal is triggered whenever $M(t)$ reaches the levels $0 = L_0 < L_1 < L_2 < \ldots$, where $L_n - L_{n-1}$ are independent exponential random variables with rate λ.

Exercise 9. *When is a Gaussian Process Stationary?* Recall that a stochastic process is stationary if its finite-dimensional distributions are invariant under shifts in time. This is sometimes called *strong stationarity*. A related notion is that a real-valued process $\{X(t) : t \geq 0\}$ is *weakly stationary* if its mean function $E[X(t)]$ is a constant and its covariance function $\text{Cov}(X(s), X(t))$ depends only on $|t-s|$. Weak stationarity does not imply strong stationarity. However, if a real-valued process is strongly stationary and its mean and covariance functions are finite, then the process is weakly stationary. Show that a Gaussian process is strongly stationary if and only if it is weakly stationary.

Exercise 10. Suppose that Y_n, $n \in \mathbb{Z}$, are independent normally distributed random variables with mean μ and variance σ^2. Consider the *moving average process*

$$X_n = a_0 Y_n + a_1 Y_{n-1} + \cdots + a_m Y_{n-m}, \qquad n \in \mathbb{Z},$$

where $a_0, \ldots, a_m$ are real numbers. Show that $\{X_n : n \in \mathbb{Z}\}$ is a Gaussian process that is stationary and specify its mean and covariance functions. Justify that this process is not Markovian and does not have stationary independent increments.

Exercise 11. Derive the mean and variance formulas in (5.14) for $M(t)$.

Exercise 12. *Equal Cruise-Control Settings.* Two autos side by side on a highway moving at 65 mph attempt to move together at the same speed by setting their cruise-control devises at 65 mph. As in many instances, nature does not always correspond to one's wishes and the actual cruise-control settings are independent normally distributed random variables V_1 and V_2 with mean $\mu = 65$ and standard deviation $\sigma = 0.4$. Find the probability that the autos move at the same speed. Find $P\{|V_1 - V_2| < .3\}$.

Exercise 13. Letting $\tau_a = \inf\{t : B(t) = a\}$, find the probability that B hits 0 in the time interval (τ_a, τ_b), where $0 < a < b$.

Exercise 14. *Arc sine Distribution.* Let $U = sin^2\theta$, where θ has a uniform distribution on $[0, 2\pi]$. Verify that $P\{U \le u\} = \arcsin\sqrt{u}$, $u \in [0, 1]$, which is the arc sine distribution.

Let X_1, X_2 be independent normally distributed random variables with mean 0 and variance 1. Show that $X_1^2/(X_1^2 + X_2^2) \stackrel{d}{=} U$. Hint: In the integral representation for $P\{X_1^2/(X_1^2 + X_2^2) \le u\}$, use polar coordinates where (x_1, x_2) is mapped to $r = (x_1^2 + x_2^2)^{1/2}$ and $\theta = \arctan x_2/x_1$.

Is it true that $U \stackrel{d}{=} 1 - U$?

Exercise 15. Prove Theorem 15 for $u \neq 1$. Using this result, find the distribution of $\eta = \sup\{t \in [0, u] : B(t) = 0\}$ for $u \neq 1$.

Exercise 16. Suppose that B and $\tilde{B}$ are independent Brownian motions. Find the moment generating function of $\tilde{B}(\tau_a)$ at the time when B hits a, which is $\tau_a = \inf\{t : B(t) = a\}$. Show that $\{\tilde{B}(\tau_a) : a \in \mathbb{R}_+\}$ considered as a stochastic process has stationary independent increments.

Exercise 17. For the hitting time $\tau = \inf\{t > 0 : B(t) \notin (-a, a)\}$, where $a > 0$, prove that its Laplace transform is

$$E[e^{-\lambda\tau}] = 1/\arccos(a\sqrt{2\lambda}).$$

Mimic the proof of Theorem 32 using the facts that $B(\tau)$ is independent of τ, and $P\{B(\tau) = -a\} = P\{B(\tau) = a\} = 1/2$.

Exercise 18. *Continuation.* In the context of the preceding exercise, verify that $E[\tau] = a^2$, and $E[\tau^2] = 5a^4/3$.

Exercise 19. Let $M(t) = \sup_{s \le t} B(s)$, and consider the process $X(t) = M(t) - B(t)$, $t \ge 0$. Show that

$$X(t) \stackrel{d}{=} M(t) \stackrel{d}{=} |B(t)|, \qquad t \ge 0.$$

(The processes X and $|B(\cdot)|$ are Markov processes on $\mathbb{R}_+$ with the same transition probabilities, and hence they are equal in distribution. However, they are not equal in distribution to M, since the latter is nondecreasing.)

Show that

$$P\{X(t) \le z | X(s) = x\} = \int_{-\infty}^{z} [\varphi(y - x; t - s) + \varphi(-y - x; t - s)]\, dy,$$

where $\varphi(x; t) = e^{-x^2/2t}/\sqrt{2\pi t}$. In addition, verify that

$$P\{M(t) > a | X(t) = 0\} = e^{-a^2/2t}.$$

Exercise 20. *Reflection Principle for Processes.* Suppose that τ is an a.s. finite stopping time for a Brownian motion B, and define

$$X(t) = B(t \wedge \tau) - (B(t) - B(t \wedge \tau)), \qquad t \geq 0.$$

Prove that X is a Brownian motion. Hint: Show that $X \stackrel{d}{=} B$ by using the strong Markov property along with the process $B'(t) = B(\tau + t) - B(\tau)$, $t \geq 0$, and the representations

$$B(t) - B(t \wedge \tau) = B'((t - \tau)^+), \qquad B(t) = B(t \wedge \tau) + B'((t - \tau)^+).$$

Exercise 21. *Continuation.* For the hitting time $\tau_a = \inf\{t > 0 : B(t) = a\}$, show that the reflected process

$$X(t) = B(t)\mathbf{1}(\tau_a \leq t) + (2a - B(t))\mathbf{1}(\tau_a > t)$$

is a Brownian motion. Use the result in the preceding exercise.

Exercise 22. Use the reflection principle to find an expression for

$$P\{B(t) > y, \ \min_{s \leq t} B(s) > 0\}.$$

Exercise 23. The value of an investment is modeled as a Brownian motion with drift $X(t) = x + \mu t + \sigma B(t)$, with an upward drift $\mu > 0$. Find the distribution of $\overline{M}(t) = \min_{s \leq t} X(s)$. Use this to find the distribution of the lowest value $\overline{M}(\infty) = \inf_{t \in \mathbb{R}_+} X(t)$ when $x = 0$. In addition, find

$$P\{X(t) - \overline{M}(t) > a\}, \qquad a > 0.$$

Exercise 24. The values of two stocks evolve as independent Brownian motions X_1 and X_2 with drifts, where $X_i(t) = x_i + \mu_i t + \sigma_i B_i(t)$, and $x_1 < x_2$. Find the probability that X_2 will stay above X_1 for at least s time units. Let τ denote the first time that the two values are equal. Find $E[\tau]$ when $\mu_1 < \mu_2$ and when $\mu_1 > \mu_2$.

Exercise 25. Show that

$$P\{B(1) \leq x | B(s) \geq 0, s \in [0, 1]\} = 1 - e^{-x^2/2}.$$

Hint: Consider $\tilde{B}(t) = B(1) - B(1 - t)$ and show that the conditional probability is equal to $P\{\tilde{B}(1) \leq x | \tilde{M}(1) = \tilde{B}(1)\}$, where $\tilde{M}(t) = \sup_{s \leq t} \tilde{B}(s)$.

Exercise 26. Consider a compound Poisson process $Y(t) = \sum_{n=1}^{N(t)} \xi_n$, where $N(t)$ is a Poisson process with rate λ and the ξ_n are i.i.d. and independent of N. Suppose ξ_1 has a mean μ, variance σ^2 and moment generating function $\psi(\alpha) = E[e^{\alpha \xi_1}]$. Show that the following are martingales with respect to Y:

$$X_1(t) = Y(t) - \lambda\mu t, \qquad X_2(t) = (Y(t) - \lambda\mu t)^2 - t\lambda(\mu^2 + \sigma^2),$$
$$X_3(t) = e^{\alpha Y(t) - \lambda t(1-\psi(\alpha))}, \qquad t \geq 0.$$

Find the mean $E[X_i(t)]$, for each i.

Exercise 27. Suppose $X(t)$ denotes the stock level of a certain product at time t and the holding cost up to time t is $Y(t) = h\int_0^t X(s)\,ds$, where h is the cost per unit time of holding one unit in inventory. Show that if X is a Brownian motion B, then the mean and covariance functions of Z are

$$E[Y(t)] = 0, \qquad \mathrm{Cov}(Y(s), Y(t)) = h^2 s^2(t/2 - s/6), \qquad s \leq t.$$

Find the mean and covariance functions of Y if $X(t) = x + \mu t + \sigma B(t)$, a Brownian motion with drift; or if X is a compound Poisson process as in the preceding problem.

Exercise 28. Prove that $Y(t) = \int_0^t B(s)\,ds$, $t \geq 0$, is a Gaussian process with mean 0 and $E[Y(t)^2] = t^3/3$.

Hint: Show that $Z = \sum_{i=1}^n u_i X(t_i)$ has a normal distribution for any $t_1, \ldots, t_n$ in $\mathbb{R}_+$, and $u_1, \ldots, u_n$ in $\mathbb{R}$. Since a Riemann integral is the limit of sums of rectangles, we know that $Z = \lim_{n\to\infty} Z_n$, where

$$Z_n = \sum_{i=1}^n u_i \sum_{k=1}^n (t_i/n) B(kt_i/n).$$

Justify that each Z_n is normally distributed, and that its limit (using moment generating functions) must also be normally distributed.

Exercise 29. *Continuation.* Suppose $X(t) = \exp\{\int_0^t B(s)\,ds\}$, $t \geq 0$. Verify that $E[X(t)] = e^{t^6/6}$.

Exercise 30. Let $Y = \{Y_n, n \geq 0\}$ be independent random variables (that need not be identically distributed) with finite means. Suppose X_0 is a deterministic function of Y_0 with finite mean. Define

$$X_n = X_0 + \sum_{i=1}^n Y_i, \qquad X_n' = X_0 \prod_{i=1}^n Y_i, \qquad n \geq 1.$$

Show that X_n is a discrete-time martingale with respect to Y if $E[Y_i] = 0$. How about if $E[Y_i] \geq 0$? Is X_n a martingale with respect to itself? What can you say about X_n' if the Y_i are positive with $E[Y_i] = 1$? or ≥ 1?

Exercise 31. *Wald Equation for Discounted Sums.* Suppose that $\xi_0, \xi_1, \ldots$ are costs incurred at discrete times and they are i.i.d. with mean μ. Consider the discounted cost process $Z_n = \sum_{m=0}^n \alpha^m \xi_m$, where $\alpha \in (0,1)$ is a discount

factor. Suppose that τ is a stopping time of the process ξ_n such that $E[\tau] < \infty$ and $E[\alpha^\tau]$ exists for some $0 < \alpha < 1$. Prove that

$$E[Z_\tau] = \frac{\mu(1 - \alpha E[\alpha^\tau])}{(1 - \alpha)}.$$

Do this by finding a convenient martingale and applying the optional stopping theorem; there is also a direct proof without the use of martingales.

Next, consider the process $S_n = \sum_{m=0}^{n} \xi_m$, $n \geq 0$, and show that

$$E[\sum_{m=0}^{\tau} \alpha^m S_m] = \frac{\mu E[\tau]}{1 - \alpha} - \frac{\alpha\mu(1 - \alpha E[\alpha^\tau])}{(1 - \alpha)^2}.$$

Exercise 32. *Continuation.* In the preceding problem, are the results true under the weaker assumption that $\xi_0, \xi_1, \ldots$ are such that $E[\xi_0] = \mu$ and $E[\xi_n | \xi_0, \ldots, \xi_{n-1}] = \mu$, $n \geq 1$?

Exercise 33. *Quadratic Variation.* Consider the quadratic increments

$$V(t) = \sum_i (B(t_i) - B(t_{i-1}))^2$$

over a partition $0 = t_0 < t_1 < \cdots < t_k = t$ of $[0, t]$. Verify $E[V(t)] = t$ and

$$\mathrm{Var}[V(t)] = \sum_i (t_i - t_{i-1})^2 \mathrm{Var}[B(1)^2].$$

Next, for each $n \geq 1$, let $V_n(t)$ denote a similar quadratic increment sum for a partition $0 = t_{n0} < t_{n1} < \cdots < t_{nk_n} = t$, where $\max_k(t_{nk} - t_{n,k-1}) \to 0$. Show that $E[V_n(t)^2 - t] \to 0$ (which says $V_n(t)$ converges in mean square distance to t). The function t is the *quadratic variation* of B in that it is the unique function (called a *compensator*) such that $B(t) - t$ is a martingale.

One can also show that $V_n \to t$ a.s. when the partitions are nested.

Exercise 34. *Random Time Change of a Martingale.* Suppose that X is a martingale with respect to $\mathcal{F}_t$ and that $\{\tau_t : t \geq 0\}$ is a nondecreasing process of stopping times of $\mathcal{F}_t$ that are bounded a.s. Verify that $X(\tau_t)$ is a martingale with respect to $\mathcal{F}_t$, and that its mean is $E[X(\tau_t)] = E[X(0)]$.

Exercise 35. *Optional Switching.* Suppose that X_n and Y_n are two martingales with respect to $\mathcal{F}_n$ that represent values of an investment in a fair market that evolve under two different investment strategies. Suppose an investor begins with the X-strategy and then switches to the Y-strategy at an a.s. finite stopping time τ of $\mathcal{F}_n$, and that $X_\tau = Y_\tau$. Then the investment value would be

$$Z_n = X_n \mathbf{1}(n < \tau) + Y_n \mathbf{1}(n \geq \tau),$$

where X_τ is the value carried forward at time τ. Show that there is no benefit for the investor to switch at τ by showing that Z_n is a martingale. Use the representation

$$Z_{n+1} = X_{n+1}\mathbf{1}(n < \tau) + Y_{n+1}\mathbf{1}(n \geq \tau) - (X_\tau - Y_\tau)\mathbf{1}(\tau = n+1).$$

Exercise 36. Prove that if σ and τ are stopping times of $\mathcal{F}_t$, then so are $\sigma \wedge \tau$ and $\sigma + \tau$.

Exercise 37. Prove that if $X(t)$ and $Y(t)$ are submartingales with respect to $\mathcal{F}_t$, then so is $X(t) \vee Y(t)$.

Exercise 38. Consider a geometric Brownian motion $X(t) = xe^{B(t)}$. Find the mean and distribution of $\tau_a = \inf\{t : X(t) = a\}$.

Exercise 39. Recall the investment option in Example 63 in which a customer may purchase a unit of a stock at a price K at time T. Consider this option with the additional stipulation that the customer "must" purchase a unit of the stock before time T if its price reaches a prescribed level a, and consequently the other purchase at time T is not allowed. In this setting, the customer must purchase the stock at the price a prior to time t, if $\max_{x \leq t} X(s) = e^{M(t)} > a$, where $M(t) = \max_{s \leq t} B(s)$. Otherwise, the option of a purchase at the price K is still available at time T. In this case, the value of the option is

$$Z = (1-a)\mathbf{1}(M(T) > \log a) + (X(T) - K)^+ \mathbf{1}(M(T) \leq \log a).$$

Prove that

$$E[Z] = 2(1-a)[1 - \Phi(\log a/\sqrt{T})] + \int_0^{\log a} \int_0^y (e^y - K) f_T(x,y)\, dx\, dy,$$

where $f_t(x,y)$ is the joint density of $B(t), M(t)$ and Φ is the standard normal distribution. Verify that

$$f_t(x,y) = \frac{2(2y-x)}{\sqrt{2\pi t^3}} e^{-(2y-x)^2/2t}, \qquad x \leq y,\ y \geq 0.$$

Verify that $E[Z]$ is minimized at the value a at which the integral term equals the preceding term. This would be the worst scenario for the customer.

Exercise 40. Prove Proposition 59 when F is not continuous. Use the fact from Exercise 11 in Chapter 1 that $\xi_n \stackrel{d}{=} F^{-1}(U_n)$, where U_n are i.i.d. with the uniform distribution on $[0,1]$. By Theorem 16 in the Appendix, you can assume the ξ_n and U_n are on the same probability space. Then the empirical distribution $G_n(t) = n^{-1} \sum_{k=1}^n \mathbf{1}(U_k \leq t)$ of the U_n satisfies $F_n = G_n(F(\cdot))$. Conclude by verifying that

$$\sup_x |F_n(x) - F(x)| = \sup_{t \le 1} |G_n(t) - t| \to 0 \qquad \text{a.s. as } n \to \infty,$$

where the limit is due to Proposition 59 for a continuous distribution.

Exercise 41. Let X be a Brownian motion directed by η. Suppose the process η has stationary independent nonnegative increments and $E[e^{-\alpha\eta(t)}] = \psi(\alpha)^t$, where $\psi(\alpha) = E[e^{-\alpha\eta(1)}]$. Determine the moment generating function of $X(1)$ (as a function of ψ) and show that X has stationary independent increments.

Exercise 42. Show that if X_n is a nonnegative supermartingale, then the limit $X = \lim_{n\to\infty} X_n$ exists a.s. and $E[X] \le E[X_0]$. Use the submartingale convergence theorem and Fatou's lemma.

Exercise 43. Let X be a Brownian motion directed by η, where the paths of η are strictly increasing a.s. Show that $X(t) = B(\eta(t))$, $t \in \mathbb{R}_+$, where B is a Brownian motion (on the same probability space as X and η) that is independent of η.

Hint: Define $B(t) = X(\hat{\eta}(t))$, where $\hat{\eta}(t) = \inf\{s \ge 0 : \eta(s) = t\}$. Argue that $\hat{\eta}(\eta(t)) = t$ and $X(t) = B(\eta(t))$, for each t, and that

$$E\Big[\exp\Big\{\sum_{i=1}^n u_i[B(t_i) - B(t_{i-1})]\Big\} \,\Big|\, \eta\Big] = \exp\{\frac{1}{2}\sum_{i=1}^n u_i^2(t_i - t_{i-1})\} \quad \text{a.s.}$$

Thus B is a Brownian motion and it is independent of η since the last expression is not random.

Exercise 44. As a variation of the model in Section 5.17, a real-valued process X is a *Brownian motion with drift μ and variability σ directed by η* if, for $0 = t_0 < t_1 < \cdots < t_n$ and $u_1, \ldots, u_n$ in $\mathbb{R}_+$,

$$E\Big[\exp\Big\{\sum_{i=1}^n u_i[X(t_i) - X(t_{i-1})]\Big\} \,\Big|\, \eta\Big]$$
$$= \exp\Big\{\sum_{i=1}^n u_i\mu[\eta(t_i) - \eta(t_{i-1})]\ \frac{1}{2}\sum_{i=1}^n u_i^2\sigma^2[\eta(t_i) - \eta(t_{i-1})]\Big\} \quad \text{a.s.}$$

Show that if $t^{-1}\eta(t) \to c$ a.s. for some $c > 0$, then

$$t^{-1}X(t) \to c\mu, \quad \text{a.s.} \qquad t^{-1}\max_{s\le t} X(s) \to c\mu \quad \text{a.s.}$$

Exercise 45. Prove Theorem 39 on continuous mappings for "separable" metric spaces by applying the coupling result for the a.s. representation of convergence in distribution (Theorem 16 in the Appendix).

Exercise 46. Use Donsker's theorem to prove that

$$P\{n^{-1/2}(S_n - \min_{k\le n} S_k) > x\} \to e^{-x^2/2}.$$

Exercise 47. In the context of Donsker's theorem, consider the range

$$Y_n = \max_{k\le n} S_k - \min_{k\le n} S_k$$

of the random walk. Show that $n^{-1/2}Y_n \xrightarrow{d} Y$, where $E[Y] = 2\sqrt{2/\pi}$. Express Y as a functional of a Brownian motion.

Exercise 48. *FCLT for Markov Chains.* Let Y_n be an ergodic Markov chain on a countable state space S with stationary distribution π. For a function $f : S \to \mathbb{R}$, consider the process

$$X_n(t) = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{\lfloor nt \rfloor} [f(Y_k) - a], \qquad t \in [0,1].$$

Specify assumptions (and a, σ) under which $X_n \xrightarrow{d} B$ as $n \to \infty$, and prove it.

Exercise 49. Show that if B is a Brownian motion, then $(1-t)B(t/(1-t))$ and $tB(1-t)/t)$ are Brownian bridges. In addition, show that if X is a Brownian bridge, then $(1+t)X(t/(1+t))$ and $(1+t)X(1/(1+t))$ are Brownian motions. Hint: Take advantage of the Gaussian property.

Exercise 50. For a Brownian bridge X, find expressions for the distribution of $\overline{M}(1) = \min_{t\le 1} X(t)$ and $M(1) = \max_{t\le 1} X(t)$.

Exercise 51. Consider the Brownian/Poisson model in Section 5.15 with the difference that the Poisson input process N is no longer time-homogeneous and its mean measure is

$$E[N(A \times I)] = \alpha|A|\Lambda(I),$$

where Λ is a measure on the time axis $\mathbb{R}$. As in Section 5.15, let $N_t(I \times A)$ denote the number of particles that enter S in the time interval $t - I$ and are in A at time t. Verify that each N_t is a Poisson process with

$$E[N_t(I \times A)] = \alpha|A| \int_{t-I} P\{V > t - s\}\Lambda(ds).$$

Is the family $\{N_t : t \in \mathbb{R}\}$ stationary as it is in Section 5.15?

Exercise 52. *Continuity of Addition in $D \times D$.* Assume that $(X_n, Y_n) \xrightarrow{d} (X, Y)$ in $D \times D$ and $\mathrm{Disc}(X) \cap \mathrm{Disc}(Y)$ is empty a.s., where $\mathrm{Disc}(x)$ denotes the discontinuity set of x. Prove that $X_n + Y_n \xrightarrow{d} X + Y$.

Exercise 53. Show that $X_n \xrightarrow{d} X$ in D if $\hat{X}_n \xrightarrow{d} X$ in D and $X_n - \hat{X}_n \xrightarrow{d} 0$. Do this by proving and applying the property that if $X_n \xrightarrow{d} X$ in D and $Y_n \xrightarrow{d} y$ in D for non-random y, then $(X_n, Y_n) \xrightarrow{d} (X, y)$ in D^2 and $X_n + Y_n \xrightarrow{d} X + y$ in D, when X has continuous paths a.s.

Exercise 54. Suppose that $(B_1(t), B_2(t))$ is a Brownian motion in $\mathbb{R}^2$, and define $\tau_a = \inf\{t : B_1(t) = a\}$. Then $X(a) = B_2(\tau_a)$ is the value of B_2 when B_1 hits a. The process $\{X(a) : a \geq 0\}$ is, of course, a Brownian motion directed by $\{\tau_a : a \geq 0\}$. Show that X has stationary independent increments and that $X(a)$ has a Cauchy distribution with density

$$f(x) = \frac{1}{a\pi(1 + (x/a)^2)}, \qquad x \in \mathbb{R}.$$

Hint: Find the characteristic function of $X(a)$.

Exercise 55. Consider the Bessel process $R(t) = \left(B_1(t)^2 \cdots B_d(t)^2\right)^{1/2}$ as in (5.43). Show that its density is

$$f_{R(t)} = \frac{2}{(2t)^{n/2}\Gamma(n/2)} r^{d-1} e^{-r^2/2t}.$$

Evaluate this for $d = 3$ by using the fact that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$. Show that $\Gamma(1/2) = \sqrt{\pi}$ by its definition $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ and the property of the normal distribution that $\sqrt{2}\int_0^\infty e^{-t^2/2}\, dt = \sqrt{\pi}$.

Exercise 56. *Continuation.* For the Bessel process $R(t)$ in the preceding exercise, show that $R(t)^2 - t$ is a martingale with mean 0.

Exercise 57. Suppose that $X(t)$ is a Brownian bridge. Find an expression in terms of normal distributions for $p_t = P\{|X(1/2) - X(t)| > 0\}$. Is p_t strictly increasing to 1 on $[1/2, 1]$?

Exercise 58. Let $X(t)$ denote a standard Brownian motion in $\mathbb{R}^3$ and let A denote the unit ball. Find the distribution of the hitting time $\tau = \inf\{t : X(t) \in A^c\}$. Is $\tau \stackrel{d}{=} \inf\{t : |B(t)| > 1\}$?

Exercise 59. For the $G/G/1$ system described in Section 5.16, consider the waiting times

$$W_n = \max_{0 \leq m \leq n} \sum_{\ell=m+1}^{n} (V_{\ell-1} - U_\ell).$$

Show that if $\rho > 1$, then $n^{-1} W_n \to E[V_1 - U_1]$ a.s. as $n \to \infty$.

Chapter 6
Appendix

This appendix covers background material from probability theory and real analysis. Included is a review of elementary notation and concepts of probability as well as theorems from measure theory, which are major tools of applied probability. More details can be found in the following textbooks:
Probability Theory — Billingsley 1968, Breiman 1992, Chung 1974, Durrett 2005, Feller 1972, Grimmett and Stirzaker 2001, Kallenberg 2004, Shiryaev 1995.
Real Analysis — Ash and Doléans-Dade 2000, Bauer 1972, Hewitt and Stromberg 1965.

6.1 Probability Spaces and Random Variables

The underlying frame of reference for random variables or a stochastic process is a probability space. A *probability space* is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set of *outcomes*, $\mathcal{F}$ is a family of subsets of Ω called *events*, and P is a *probability measure* defined on these events. The family $\mathcal{F}$ is a *σ-field* (or σ-algebra): If $A \in \mathcal{F}$, then so is its complement A^c, and if a sequence A_n is in $\mathcal{F}$, then so is its union $\cup_n A_n$.

The probability measure P satisfies the properties of being a *measure*: It is a function $P : \mathcal{F} \to [0,1]$ such that, for any finite or countably infinite collection of disjoint sets A_n in $\mathcal{F}$,

$$P(\cup_n A_n) = \sum_n P(A_n).$$

Furthermore, P satisfies $P(\Omega) = 1$.

Under this definition, $P(A) \leq 1$, $P(A^c) = 1 - P(A)$, where $A^c = \Omega \backslash A$ (the complement of A), and

R. Serfozo, *Basics of Applied Stochastic Processes*,
Probability and its Applications.

$$P(A) \le P(B), \qquad A \subset B,$$
$$P(A_n) \to P(A), \qquad \text{if } A_n \uparrow A \text{ or } A_n \downarrow A.$$

The definition of a random variable involves the notion of a measurable function. Suppose $(S, \mathcal{S})$ and $(S', \mathcal{S}')$ are measurable spaces (sets with associated σ-fields). A function $f : S \to S'$ is *measurable* if

$$f^{-1}(A) = \{x \in S : f(x) \in A\} \in \mathcal{E}, \qquad \text{for each } A \in \mathcal{E}'.$$

That is, the set of all x's that f maps into A is in $\mathcal{E}$.

Typically, S will be the outcome space Ω, the real line $\mathbb{R}$, the d-dimensional Euclidean space $\mathbb{R}^d$, or a metric space. We adopt the standard convention that the σ-field $\mathcal{S}$ for S is its *Borel* σ-field — the smallest σ-field containing all open sets in S (or the σ-field consisting of countable unions of open sets and all complements of these). We sometimes write $\mathcal{B}$ and $\mathcal{B}_+$ for the Borel σ-fields of $\mathbb{R}$ and $\mathbb{R}_+$. A useful property is that if $f : S \to S'$ and $g : S' \to S''$ are measurable, then the *composition function* $g \circ f(x) = g(f(x))$ is measurable.

The rest of this sections concerns classical real-valued random variables. We discuss random variables on metric spaces in Section 6.3. A *random variable* X on a probability space $(\Omega, \mathcal{F}, P)$ is a measurable mapping from Ω to $\mathbb{R}$. The measurability of X ensures that $\mathcal{F}$ contains all sets of the form

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}, \qquad B \in \mathbb{B}_+. \tag{6.1}$$

These are the types of events for which P is defined. One usually constructs (or assumes) the σ-field $\mathcal{F}$ is large enough such that the random variables of interest are measurable. For instance, if X, Y and Z are of interest, one can let $\mathcal{F} = \sigma(X, Y, Z)$, the "smallest σ-field" containing all sets of the form (6.1) for X, Y and Z, so that they are measurable.

A statement about events or random variables is said to hold *almost surely* (a.s.) if the statement holds with probability one (some say the statement is true almost everywhere (a.e.) on Ω with respect to P). For instance $X + Y \le Z$ a.s. Also, we sometimes omit a.s. from elementary statements like $X = Y$ and $X \le Y$ that hold a.s.

All of the probability information of X in "isolation" (not associated with other random quantities on the probability space) is contained in its *distribution function*

$$F(x) = P\{X \le x\}, \qquad x \in \mathbb{R}.$$

Here $P\{X \le x\} = P(\{\omega : X(\omega) \le x\})$. A distribution function has at most a finite or countable number of discontinuities (which may be a dense set); this is a well-known property of any increasing function. We sometimes write the distribution as $F_X(x)$. Several standard distribution functions are in the next section.

The random variable X is *discrete* if the range of X is a countable set S in $\mathbb{R}$. In this case, the *probability function of* X is $P\{X = x\}$, $x \in S$; and

$$P\{X \in A\} = \sum_{x \in A} P\{X = x\}, \qquad A \subset S.$$

The mean (or expectation) of X is

$$E[X] = \sum_{x \in S} xP\{X = x\},$$

provided the sum exists (it is absolutely convergent, meaning the sum of its absolute values is finite).

The random variable X is *continuous* if there is a (measurable) *density function* $f : \mathbb{R} \to \mathbb{R}_+$ such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$ and $P\{X \in A\} = \int_A f(x)\,dx$, $A \subset \mathbb{R}$. Then the distribution of X is $F(x) = \int_{-\infty}^{x} f(y)\,dy$, and so $f(x) = F'(x)$, the derivative of F. The mean (or expectation) of X is

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

provided the integral exists (the integral of its absolute value is finite).

There are random variables that are not discrete or continuous. Regardless of whether the random variable X is discrete, continuous or general, its mean will be denoted by

$$E[X] = \int_{\mathbb{R}} xdF(x),$$

a Riemann-Stieltjes integral defined in Section 6.4. In addition to its mean, other summary measures of a random variable X are as follows. The *nth moment* of X is $E[X^n]$, and the *nth moment about its mean* $\mu = E[X]$ is $E[(X - \mu)^n]$. The *variance* of X is

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

Whenever we refer to these moments, we assume they are finite.

6.2 Table of Distributions

The following are tables of some standard distributions and their means, variances, and moment generating functions (which we discuss shortly).

Discrete Random Variables

Random Variable	$P\{X = x\}$	$E[X]$	$\mathrm{Var}[X]$	$E[e^{sX}]$
Binomial $n \geq 1,\ p \in (0,1)$	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	np	$np(1-p)$	$(pe^s + (1-p))^n$
Poisson $\lambda > 0$	$e^{-\lambda}\lambda^x / n!$ $x = 0, 1, \ldots$	λ	λ	$e^{-\lambda(1-e^{-s})}$
Geometric $p \in (0,1)$	$p(1-p)^{x-1}$ $x = 1, 2, \ldots,$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^s}{1-(1-p)e^s}$
Negative Binomial $r \geq 1,\ p \in (0,1)$	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$ $x = r, r+1, \ldots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$(\frac{pe^s}{1-(1-p)e^s})^r$

Continuous Random Variables

Random Variable	Density $f(x)$	$E[X]$	$\mathrm{Var}[X]$	$E[e^{sX}]$
Normal $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}, \quad x \in \mathbb{R}$	μ	σ^2	$e^{\mu s + \sigma^2 s^2/2}$
Exponential $\lambda > 0$	$\lambda e^{-\lambda x}, \quad x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - s}$
Gamma* $\alpha \geq 0,\ \lambda > 0$	$\frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$(\frac{\lambda}{\lambda - s})^\alpha$
Uniform on $[a, b]$	$\frac{1}{b-a}, \quad a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bs} - e^{as}}{s(b-a)}$
Beta $a, b > 0$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ $0 \leq x \leq 1$	$\frac{a}{(a+b)}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\cdots$

*The gamma density with integer $\alpha = n \geq 1$ is $\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$; it is called an *Erlang density*. The gamma density with $\alpha = n/2$ and $\lambda = 1/2$ is a *ξ-squared density* with n degrees of freedom.

6.3 Random Elements and Stochastic Processes

A unified way of discussing random vectors, stochastic processes and other random quantities is in terms of random elements. Suppose one is interested in a random element that takes values in a space S with a σ-field $\mathcal{S}$. A *random element in* S, defined on a probability space $(\Omega, \mathcal{F}, P)$, is a measurable mapping X from Ω to S. The X is also called an S*-valued random variable*.

For our purposes, the space S will be a countable set, a Euclidean space $\mathbb{R}^d$, or a function space with a distance metric (for representing a stochastic process). To accommodate these and other spaces as well, we adopt the standard convention that $(S, \mathcal{S})$ is a *Polish space*. That is, S is a metric space that is complete (each Cauchy sequence is convergent) and separable (there is a countable dense set in S); and $\mathcal{S}$ is the Borel σ-field generated by the open sets. A *metric* on S is a map $d : S \times S \to \mathbb{R}_+$ such that

$$\begin{aligned} &d(x, y) = d(y, x), \qquad && d(x, y) = 0 \quad \text{if and only if } x = y, \\ &d(x, z) \le d(x, y) + d(y, z), && x, y, z \in S. \end{aligned}$$

Our discussion of functions, integrals, convergence, etc. on S does not require a familiarity of Polish spaces, since these concepts are understandable by interpreting them as being on $\mathbb{R}^d$. We use terminology involving Polish spaces and random elements in this appendix because it allows for a rigorous and unified presentation of background material, but this terminology is not used throughout the book.

The *probability distribution* of a random element X in S is the probability measure

$$F_X(B) = P\{X \in B\} = P \circ X^{-1}(B), \qquad B \in \mathcal{S}.$$

If X and Y are random elements whose distributions are equal, we say that X is *equal in distribution* to Y and denote this by $X \stackrel{d}{=} Y$. The underlying probability spaces for X and Y need not be the same.

Loosely speaking, a stochastic process is a collection of random variables (or random elements) defined on a single probability space. Hereafter, we will simply use the term "random elements" (which includes random variables), and let $(S, \mathcal{S})$ denote the Polish space where they reside.

A *discrete-time stochastic process* (or random sequence) is a collection of random elements $X = \{X_n : n \ge 0\}$ in S defined on a probability space $(\Omega, \mathcal{F}, P)$. The nonnegative integer n is a *time parameter* and S is the *state space* of the process. The value $X_n(\omega) \in S$ is the *state* of the process at time n associated with the outcome ω.

Note that X is also a random element in the infinite product space S^∞ with the product σ-field $\mathcal{S}^\infty$: the smallest σ-field generated by sets $B_1 \times \cdots \times B_n$, B_j's $\in \mathcal{S}$. Its distribution $P\{X \in B\}$, for $B \in \mathcal{S}^\infty$, is uniquely defined in terms of its *finite-dimensional distributions*

$$P\{X_1 \in B_1, \ldots, X_n \in B_n\}, \qquad B_j \in \mathcal{S},\ n \geq 1.$$

Consequently, if Y is another random element in S^∞ whose finite-dimensional distributions are equal to those of X, then $X \stackrel{d}{=} Y$. We sometimes refer to the process $X = \{X_n : n \geq 0\}$ by X_n.

Stochastic processes in continuous time are defined similarly to those in discrete time, but their evolutions over time are typically more complicated. A *continuous-time stochastic process* is a collection of random elements $\{X(t) : t \geq 0\}$ in S defined on a probability space, where $X(t, \omega)$ is the state at time t associated with the outcome ω. The function $t \to X(t, \omega)$ from $\mathbb{R}_+$ to S, for a fixed ω, is the *sample path* or *trajectory* associated with the outcome ω. Accordingly, $X(t)$ is a random function from $\mathbb{R}_+$ to S. More precisely, the entire process $X = \{X(t) : t \geq 0\}$ is a random element in a space of functions from $\mathbb{R}_+$ to S. We sometimes refer to the process by $X(t)$.

A standard example is when the sample paths of X are in the set $D(R_+)$ of functions from $\mathbb{R}_+$ to S that are right-continuous with left hand limits — often called *cadlag* functions (from the French *continu à droite, limites à gauche*). Then X is a random element in $D(R_+)$, with an appropriate metric depending on one's application (e.g., a uniform metric or a metric for the Skorohod topology [11, 60, 113]), and $P\{X \in B\}$ is for a Borel set $B \subset D(R_+)$ of sample paths. The distribution of X is uniquely determined by its *finite-dimensional distributions*

$$P\{X(t_1) \in B_1, \ldots, X(t_n) \in B_n\}, \qquad t_1 < \cdots < t_n,\ B_j \in \mathcal{S},\ n \geq 1.$$

In summary, a stochastic process is a family of random variables or random elements defined on a probability space that contains all the probability information about the process. We will use the standard convention of suppressing the ω in random elements such as X_n or $X(t)$, and not displaying the underlying probability space $(\Omega, \mathcal{F}, P)$, unless it is essential for the exposition. Also, all the functions appearing in this book are measurable, and we will mention this property only when it is needed.

6.4 Expectations as Integrals

We defined the mean of discrete and continuous random variables above. These are special cases of the following definition for any real-valued random variable.

Definition 1. Let X be a random variable with distribution function $F(x) = P\{X \leq x\}$. The *mean* (or expectation) of X is defined by the Riemann-Stieltjes integral

$$E[X] = \int_{\mathbb{R}} x dF(x),$$

provided the integral exists.

This general definition is needed for random variables that are not discrete or continuous. For instance, if X has positive probabilities at points in a countable set S, and also has a sub-density $f(x)$ elsewhere on $\mathbb{R}$, then

$$E[X] = \sum_{x \in S} xP\{X = x\} + \int_{\mathbb{R}} xf(x)\, dx.$$

Riemann-Stieltjes integrals are similar to Riemann integrals in calculus. A Riemann-Stieltjes integral of a function $g : [a, b] \to \mathbb{R}$ with respect to F is constructed by limits of the upper and lower Darboux sums $\mathcal{D}^\chi$ and $\mathcal{D}_\chi$ defined on the set of points $\chi = \{a = x_0 < x_1 < \ldots < x_n = b\}$ by

$$\mathcal{D}^\chi = \sum_{j=1}^{n} \sup\{g(x) : x_{j-1} \le x \le x_j\}[F(x_j) - F(x_{j-1})],$$

and $\mathcal{D}_\chi$ is defined similarly with sup replaced by inf. The Riemann-Stieltjes integral of g exists if, for any $\varepsilon > 0$, there is a set χ depending on g and ε such that $\mathcal{D}^\chi - \mathcal{D}_\chi < \varepsilon$. When it exists, the integral has the form (e.g., see [53])

$$\int_{[a,b]} g(x) dF(x) = \inf_\chi \mathcal{D}^\chi = \sup_\chi \mathcal{D}_\chi.$$

This integral is a Riemann integral $\int_{[a,b]} g(x)f(x)dx$ (as in calculus), when $dF(x) = f(x)dx$ and dx is the *Lebesgue* measure on $\mathbb{R}$.

Riemann-Stieltjes integrals on infinite intervals are defined similar to Riemann integrals. For instance, for $g : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}} g(x) dF(x) = \lim_{a,b \to \infty} \int_{[-a,b]} g(x) dF(x),$$

provided the limit exists and is finite.

For a function $g : S \to \mathbb{R}$ on a metric space S, its integral with respect to a measure μ on S is defined as a Lebesgue integral [53] denoted by

$$\int_S g(x)\mu(dx).$$

We denote the differential by $\mu(dx)$ instead of $d\mu(x)$ to emphasize that it is not a Riemann-Stieltjes integral.

Another equivalent expression for the expectation of X in terms of the probability P is the Lebesgue integral

$$E[X] = \int_\Omega X(\omega) P(d\omega).$$

A few properties of the expectation operator are $E[a] = a$, for $a \in \mathbb{R}$,

$$E[X+Y] = E[X] + E[Y], \qquad E[X] \le E[Y] \quad \text{if } X \le Y,$$

$$E[\sum_{j=1}^{n} a_j X_j] = \sum_{j=1}^{n} a_j E[X_j].$$

6.5 Functions of Stochastic Processes

Many features of a stochastic process, or related quantities of interest, are expressed as functions of the process. This section contains several examples and a formula for evaluating expectations of real-valued functions of random elements and processes.

Suppose that X is a random element in a Polish space S, such as a discrete- or continuous-time process $X = \{X_n : n \ge 0\}$ or $X = \{X(t) : t \ge 0\}$, and denote its distribution by $F_X(B) = P\{X \in B\}$, $B \in \mathcal{S}$. Consider a measurable function $g : S \to S'$ and define $Y = g(X)$. This Y is a random element in S' since it is a composition of X and g, which are measurable. When X is a continuous-time process, $g(x)$ is a function on the space of sample paths $x = \{x(t) : t \ge 0\}$. The distribution of Y is

$$P\{Y \in B\} = P\{g(X) \in B\} = P\{X \in g^{-1}(B)\},$$

where $g^{-1}(B) = \{x \in S : g(x) \in B\}$. Then the distribution of Y as a function of F_X is the probability measure

$$F_Y(B) = F_X \circ g^{-1}(B) = F_X(g^{-1}(B)), \qquad B \in \mathcal{S}'.$$

In some cases, the function g is a standard measurable operation on real numbers such as addition, subtraction, maximum, etc. For instance, if $X = \{X_n : n \ge 0\}$ is a family of random variables, then $Y = X_1 + \cdots + X_n$ is a random variable for fixed $n < \infty$, since the addition function $g(x) = x_1 + \cdots + x_n$ from $\mathbb{R}^\infty$ to $\mathbb{R}$ is measurable. Other standard examples of random variables that are measurable functions of X include

$$\prod_{j=1}^{n} X_j, \qquad \max_{1 \le j \le n} X_j, \qquad \sum_{j=1}^{n} a_n (X_n - X_j).$$

Examples based on multiple compositions of functions are

$$\sup_{n\geq 0} X_n - \inf_{n\geq 0} X_n, \qquad \sup_{n\geq 0}\Big[e^{-aX_n}\sum_{j=1}^{n}(Y_j - Y_{j-1})\Big],$$

provided they exist.

Examples of $g(X)$ for a continuous-time process X include analogues of those above as well as

$$\int_0^t X(t-s)\,ds, \qquad \int_{\mathbb{R}_+} e^{-a(t)} \inf_{s\leq t} X(s)\,dt.$$

In modeling a stochastic system, the state of the system, or a performance measure for it, are often of the form $Y(t) = g(t, X)$, where the process X represents the time-dependent system data, and the function $g(t, x)$ represents the dynamics of the system.

We will now describe a useful formula for the mean of a real-valued function $Y = g(X)$ of the random element X. The mean of Y in terms of the distribution F_X of X is the Lebesgue integral

$$E[g(X)] = \int_S g(x)F_X(dx), \tag{6.2}$$

provided it exists. This follows since $F_Y = F_X \circ g^{-1}$ and, by the change-of-variable formula below, we have

$$E[Y] = \int_{\mathbb{R}} yF_X \circ g^{-1}(dy) = \int_S g(x)F_X(dx).$$

Change-of-variable Formula for Lebesgue integrals. Suppose F is a measure on S, and $g : S \to S'$ and $h : S' \to \mathbb{R}$ are measurable. Then

$$\int_S h \circ g(x)F(dx) = \int_{S'} h(y)F \circ g^{-1}(dy), \tag{6.3}$$

provided both integrals exist (one exists if and only if the other one does).

Important functions of random variables are generating functions and transforms. They are tools for characterizing distributions and evaluating their moments. The *moment generating function* of a random variable X is

$$m_X(s) = E[e^{sX}] = \int_{\mathbb{R}} e^{sx} dF_X(x),$$

provided the integral exists for s in some interval $[0, a]$, where $a > 0$. A major property is that a moment generating function uniquely determines a distribution and vice versa ($m_X = m_Y$ if and only if $F_X = F_Y$). Also, the nth moment of X, when it exists, has the representation

$$E[X^n] = m_X^{(n)}(0),$$

which is the nth derivative of m_X at 0. Moment generating functions of some standard distributions are given in Section 6.2 below.

Two variations of moment generating functions for special types of random variables are as follows. For a "nonnegative" random variable X, its *Laplace transform* (or the Laplace-Stieltjes transform of F_X) is

$$E[e^{-sX}] = \int_{\mathbb{R}_+} e^{-sx} dF_X(x), \qquad s \geq 0.$$

For a "discrete" random variable X whose range is contained in the nonnegative integers, its *generating function* is

$$E[s^X] = \sum_{n=0}^{\infty} s^n P\{X = n\}, \qquad -1 \leq s \leq 1.$$

Laplace transforms and generating functions play the same role as moment generating functions in that they uniquely determine distribution functions and yield moments of random variables. Laplace transforms are also defined for increasing functions F that need not be bounded, such as renewal functions, and they can also be extended to the complex plane. A similar statement applies to generating functions.

A generalization of a moment generating function is a characteristic function. The *characteristic function* of a random variable X (or the *Fourier-Stieltjes transform* of F_X) is defined by

$$E[e^{isX}] = \int_{\mathbb{R}} e^{isx} dF_X(x) \qquad s \in \mathbb{R},$$

where $i = \sqrt{-1}$ and $e^{isx} = \cos sx + i \sin sx$. A characteristic function, which is complex-valued, exists for "any" random variable (or distribution function). In contrast, a moment generating function is real-valued, but it only exists for a random variable whose moments exist. There is a one-to-one correspondence between distribution functions and characteristic functions, and moments are expressible by derivatives of characteristic functions at 0. A characteristic function is typically used when the more elementary moment generating function, Laplace transform, or generating function are not applicable.

The following are useful inequalities involving expectations of functions of random variables. For a random variable X and increasing $g : \mathbb{R} \to \mathbb{R}_+$,

$$P\{X \geq x\} \leq E[g(X)]/g(x), \qquad \textit{Markov's Inequality.}$$

An example is

$$P\{|X - E[X]| \geq x\} \leq \mathrm{Var}[X]/x^2, \qquad \textit{Chebyshev's Inequality.}$$

For random variables X, Y with finite second moments,

$$E|XY| \le \sqrt{E[X^2]E[Y^2]}, \qquad \textit{Cauchy-Buniakovsky-Schwarz.}$$

For random variables $X_1, \ldots, X_n$ and convex $g : \mathbb{R}^n \to \mathbb{R}$,

$$E[g(X_1, \ldots, X_n)] \ge g(E[X_1], \ldots, E[X_n]), \qquad \textit{Jensen's Inequality.}$$

6.6 Independence

In this section, we define independent random variables and random elements, and describe several functions of them including summations.

Random variables $X_1, \ldots, X_n$ are *independent* if, for Borel sets $B_1, \ldots, B_n$ in $\mathbb{R}$,

$$P\{X_1 \in B_1, \ldots, X_n \in B_n\} = \prod_{j=1}^{n} P\{X_j \in B_j\} = \prod_{j=1}^{n} F_{X_j}(B_j).$$

An infinite family of random variables are independent if any finite collection of the random variables are independent. The same definitions apply to independence of random elements such as the independence of stochastic processes.

Many properties of functions of independent random elements can be analyzed in terms of their distributions. Here is an important formula for expectations. Suppose X and Y are independent random elements in S and S', respectively, and $g : S \times S' \to \mathbb{R}$ is measurable. Then by (6.2) and Fubini's theorem stated below, we have

$$\begin{aligned} E[g(X,Y)] &= \int_{S\times S'} g(x,y) F_X(dx) F_Y(dy) \\ &= \int_{S'} \Big[\int_S g(x,y) F_X(dx) \Big] F_Y(dy), \end{aligned} \tag{6.4}$$

provided the integral exists.

For the next result, we use the notion that a measure space $(S, \mathcal{S}, \mu)$ is σ-*finite* if there is a partition $B_1, B_2, \ldots$ of S such that $\mu(B_n) < \infty$, for each n. A Polish space has this property.

Theorem 2. (Fubini) *Suppose μ and η are measures on the respective σ-finite spaces $(S, \mathcal{S})$ and $(S', \mathcal{S}')$, and $g : S \times S' \to \mathbb{R}$ is measurable. If g is nonnegative or $\int_{S\times S'} |g(x,y)|\mu(dx)\eta(dy)$ is finite, then*

$$\int_{S\times S'} g(x,y)\mu(dx)\eta(dy) = \int_{S'} \Big[\int_S g(x,y)\mu(dx)\Big]\eta(dy)$$
$$= \int_S \Big[\int_{S'} g(x,y)\eta(dy)\Big]\mu(dx).$$

This says that if the integral exists on the product space, then it equals the single-space integrals done separately (in either order).

A special case of (6.4) is

$$E[g(X)h(Y)] = \int_S g(x)F_X(dx)\int_{S'} h(y)F_Y(dy) = E[g(X)]E[h(Y)].$$

This generalizes, for independent $X_1,\ldots,X_n$ in S and $g_j : S \to \mathbb{R}$, to

$$E[\prod_{j=1}^n g_j(X_j)] = \prod_{j=1}^n E[g_j(X_j)].$$

provided the expectations exist.

Example 3. Suppose $X_1,\ldots,X_n$ are independent nonnegative random variables. Then the moment generating function of $Y = \sum_{j=1}^n X_j$ is

$$E[e^{sY}] = \prod_{j=1}^n E[e^{sX_j}].$$

Now, assume each X_j has an exponential distribution with rate λ, and so $E[e^{sX_j}] = \lambda/(\lambda - s)$, $0 \le s < \lambda$. Consequently, $E[e^{sY}] = [\lambda/(\lambda - s)]^n$, which is the moment generating function of a gamma (or Erlang) distribution with parameters λ and n (see Section 6.2). Hence Y has this gamma distribution.

One can also determine distributions of sums of independent random variables by convolutions of their distributions. Specifically, if X and Y are independent random variables, then

$$P\{X + Y \le z\} = \int_{\mathbb{R}} F_Y(z - x)dF_X(x). \tag{6.5}$$

That is, $F_{X+Y}(z) = F_X \star F_Y(z)$, where $\star$ is the convolution operator defined below. To prove (6.5), first note that by (6.2) (even for dependent X and Y), we have

$$P\{X + Y \le z\} = E[\mathbf{1}(X + Y \le z)] = \int_{x+y\le z} F_{X,Y}(dx \times dy).$$

Then applying $F_{X,Y}(dx \times dy) = dF_X(x)dF_Y(y)$ (from the independence) and Fubini's theorem to this double integral yields (6.5).

Properties of convolutions are as follows. The *convolution* of two distributions F and G is defined by

$$F \star G(z) = \int_{\mathbb{R}} G(z-x)dF(x). \tag{6.6}$$

Note that $F \star G = G \star F$. Also, if $F(0-) = G(0-) = 0$, then

$$F \star G(z) = \int_{(0,z]} G(z-x)dF(x).$$

If F and G have respective densities f and g, then the derivative of (6.6) yields the formula

$$f \star g(z) = \int_{\mathbb{R}} g(z-x)f(x)dx.$$

These formulas reduce to sums when F and G are discrete distributions.

Convolutions of several distributions are defined in the obvious way, for instance $F \star G \star H = F \star (G \star H)$, and if X, Y and Z are independent random variables, then $F_{X+Y+Z} = F_X \star F_Y \star F_Z$. The nth fold convolutions $F^{n\star}(x)$ of a distribution F, for $n \geq 0$, are defined recursively by $F^{0\star}(x) = \mathbf{1}(x \geq 0)$ and, for $n \geq 1$,

$$F^{n\star}(x) = F^{(n-1)\star} \star F(x) = F \star \cdots \star F(x), \qquad n \text{ convolutions.}$$

If $T_n = X_1 + \cdots + X_n$ where the X_j are independent with distribution F, then $F_{T_n} = F^{n\star}$.

The definition (6.6) of a convolution also extends to more general functions $\mu \star h$, where μ is a measure on $\mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ is such that the integral exists. For example, renewal theory involves convolutions of the form

$$U \star h(t) = \int_{[0,t]} h(t-s)dU(s),$$

where $U(t) = \sum_{n=0}^{\infty} F^{n\star}(t)$, $F(0) = 0$ and $h(t)$ is bounded on compact sets and is 0 for $t < 0$.

6.7 Conditional Probabilities and Expectations

Section 1.22 in Chapter 1 reviewed fundamentals of conditional probabilities and expectations for discrete random variables. This section continues the review for random elements as well as non-discrete random variables.

Suppose X and Y, defined on a common probability space $(\Omega, \mathcal{F}, P)$, are random elements in Polish spaces S and S', respectively. A *probability kernel* from S' to S is a function $p : S' \times \mathcal{S} \to [0,1]$ such that $p(y, \cdot)$ is a probability measure on $\mathcal{S}$ for each $y \in S'$, and $p(\cdot, B)$ is a measurable function for each $B \in \mathcal{S}$. There exists a probability kernel $p(y, B)$ from S' to S such that

$$P\{X \in B\} = \int_{S'} p(y, B) F_Y(dy), \qquad B \in \mathcal{S}. \tag{6.7}$$

The kernel p is unique in the sense that if p' is another such kernel, then $p(Y, \cdot) = p'(Y, \cdot)$ a.s. (e.g., see [60]).

Definition 4. Using the preceding notation, the (random) probability measure

$$P\{X \in B|Y\} = p(Y, B), \qquad B \in \mathcal{S}$$

is the *conditional probability measure of X given Y*. When X is a random variable with a finite mean, the *conditional expectation* of X given Y is

$$E[X|Y] = \int_{\mathbb{R}} xp(Y, dx). \tag{6.8}$$

Conditional probabilities and expectations — which are random quantities — are sometimes written as non-random quantities

$$P\{X \in B|Y = y\} = p(y, B), \qquad E[X|Y = y] = \int_{\mathbb{R}} xp(y, dx), \qquad y \in S'.$$

An important property of conditional expectations, which follows from the definition, is

$$E[X] = E[E[X|Y]] = \int_{S'} E[X|Y = y] F_Y(dy).$$

Similarly, $P\{X \in B\} = E[P\{X \in B|Y\}]$. These formulas are useful tools for evaluating the mean or distribution of X in terms of the conditional means or distributions.

Another important property is that $E[X|Y]$ is a measurable function of Y, because $E[X|Y] = h(Y)$, where $h(y) = E[g(X, y)] = \int_S xp(y, dx)$ is measurable. Since Definition 4 is for random elements, Y may represent several random elements. For instance $E[X|Y, Z]$ is a measurable function of Y, Z and

$$E[X|Z] = E\Big[E[X|Y, Z]\Big|Z\Big].$$

Definition 4 is consistent with the definition used for elementary random variables. For instance, in case Y is a discrete random variable, the probability kernel that satisfies (6.7) is

$$p(y, B) = P\{X \in B|Y = y\} = P\{X \in B, Y = y\}/P\{Y = y\}.$$

Next suppose X and Y are continuous random variables such that

$$P\{X \in A, Y \in B\} = \int_{A \times B} f(x, y)\, dx\, dy,$$

where $f(x, y)$ is their joint density. Then the probability kernel satisfying (6.7) is

$$p(y, B) = \int_B f(x, y)/f(y)\, dx.$$

The preceding notions extend to conditional probabilities $P\{X \in B|\mathcal{F}\}$ and expectations $E[X|\mathcal{F}]$ for conditioning on a σ-field $\mathcal{F}$ instead of a random element. Definition 4 includes these cases when $\mathcal{F} = \sigma(Y)$, the smallest σ-field generated by Y. We will only use this general notation only in Chapter 5.

Here are some more properties of conditional expectations (assuming they exist) for measurable $f : S' \to \mathbb{R}$ and $g : S \times S' \to \mathbb{R}$:

$$E[Xf(Y)|Y] = f(Y)E[X|Y],$$
$$E[g(X, Y)|Y = y] = E[g(X, y)|Y = y].$$

Furthermore, if X and Y are independent, then

$$E[g(X, Y)|Y = y] = E[g(X, y)],$$

or equivalently $E[g(X, Y)|Y] = E[h(Y)]$, where $h(y) = E[g(X, y)]$. Most of the standard properties of expectations extend to conditional expectations. For instance, $P\{X \in B|Y\} = E[\mathbf{1}(X \in B)|Y]$,

$$E[X|Y] \le E[Z|Y] \qquad \text{if } X \le Z,$$
$$E[f(X) + g(Z)|Y] = E[f(X)|Y] + E[g(Z)|Y].$$

6.8 Existence of Stochastic Processes

A stochastic process is commonly defined by designating a set of properties that its distribution and sample paths must satisfy. The existence of the process amounts to showing that there exist a probability space and functions on it (the sample paths) that satisfy the designated properties of the process. This section describes Kolmogorov's theorem that is used for such a task.

For a stochastic process $\{X(t) : t \in \mathbb{R}_+\}$ on a Polish space S, we mentioned that any probability for it can be expressed in terms of its finite-dimensional probability measures μ_I on S^I. For any finite set I in $\mathbb{R}_+$, and $A_t \in \mathrm{S}$, $t \in I$,

$$\mu_I(\times_{t\in I} A_t) = P\{X(t) \in A_t, t \in I\}. \tag{6.9}$$

Here S is the Borel σ-field on S, and the product σ-field S^I is used on S^I.

These probability measures satisfy the *consistency condition*

$$\mu_J(\cdot \times S^{J\setminus I}) = \mu_I(\cdot), \qquad I \subset J. \tag{6.10}$$

That is, for $A_t \in \mathrm{S}$, $t \in J$,

$$\mu_J(\times_{t\in J} A_t) = \mu_I(\times_{t\in I} A_t), \qquad \text{if } A_t = \mathrm{S} \text{ for } t \in J\backslash I.$$

Theorem 5. (Kolmogorov Extension Theorem) *For probability measures μ_I that satisfy the consistency condition (6.10), there exists a stochastic process $\{X(t) : t \in \mathbb{R}_+\}$ on S that is defined on a probability space $(\Omega, \mathcal{F}, P)$ such that μ_I are its finite-dimensional distributions as in (6.9).*

The proof of this result in [60] defines $(\Omega, \mathcal{F}) = (S^{\mathbb{R}_+}, \mathrm{S}^{\mathbb{R}_+})$, and defines P by

$$P(\ \cdot \times S^{J^c}) = \mu_J(\cdot).$$

Here μ_J are probability measures for countable sets J that are extensions, defined by (6.10), of the probability measures μ_I for finite sets I. Also, for each t, the function $X(t)$ from Ω to S is defined as the projection map

$$X(t) = X(t, \omega) = \omega_t, \qquad \omega = \{\omega_t : t \in \mathbb{R}_+\} \in \Omega.$$

The preceding theorem also applies to processes in which the time parameter t takes values in other time sets such as $\mathbb{R}$ or the nonnegative integers. Here is an important example.

Corollary 6. (Existence of Independent Random Elements) *For probability measures $\mu_1, \mu_2, \ldots$ on a Polish space S, there exist independent S-valued random elements $X_1, X_2, \ldots$ defined on a probability space $(\Omega, \mathcal{F}, P)$ such that $P\{X_j \in \cdot\} = \mu_j(\cdot)$, $j \geq 1$.*

Recall that processes X and X' are equal in distribution, denoted by $X \stackrel{d}{=} X'$, if their finite-dimensional probability measures are equal. However, their sample paths need not be the same. For instance, suppose

$$X(t) = 0, \qquad X'(t) = \mathbf{1}(\xi = t), \qquad t \in \mathbb{R}_+,$$

where ξ is exponentially distributed with rate λ. Clearly $X \stackrel{d}{=} X'$ since $P\{\xi = t\} = 0$ for any t. On the other hand, their sample paths are not the same: $P\{X(t) = 0,\ t \in \mathbb{R}_+\} = 1$, but $P\{X'(t) = 0,\ t \in \mathbb{R}_+\} = 0$.

Although the consistency condition on finite-dimensional probability measures μ_I guarantees the existence of a stochastic process, additional conditions on the μ_I are needed to infer that the sample paths of the process have certain properties (e.g., that the paths are continuous, or right-continuous).

6.9 Convergence Concepts

Many properties of stochastic processes are expressed in terms of convergence of random variables and elements. There are several types of convergence, but for our purposes we primarily use convergence with probability one and convergence in distribution, which we now describe.

We begin with a review of convergence of real numbers. A sequence of real numbers x_n *converges* to some $x \in \mathbb{R}$, denoted by $\lim_{n\to\infty} x_n = x$, if, for any $\varepsilon > 0$, there exists a number N such that

$$|x_n - x| < \varepsilon, \qquad n \geq N.$$

We sometimes refer to this convergence as $x_n \to x$.

One often establishes convergence with the quantities

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{k\geq n} x_k, \qquad \limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{k\geq n} x_k. \tag{6.11}$$

This *limit inferior* and *limit superior* clearly satisfy

$$-\infty \leq \liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n \leq \infty.$$

If both of these quantities are equal to a finite x, then $x_n \to x$.

For insight into this result, let a and b denote the lim inf and lim sup in (6.11) and assume they are finite. By its definition, a is the lower "cluster value" of the x_n's in that x_n is in the interval $[a, a+\varepsilon]$ infinitely often (i.o.), for fixed $\varepsilon > 0$. Similarly, x_n is in the interval $[b-\varepsilon, b]$ i.o. Now x_n does not converge to a limit if $a < b$ (since x_n is arbitrarily close to both a and b i.o.). On the other hand, $x_n \to x$ if and only if $a = b = x$.

The preceding properties of real numbers readily extend to random variables. Suppose $X, X_1, X_2, \ldots$ are random variables on a probability space. The sequence X_n *converges with probability one* to X if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega), \qquad \omega \in \Omega' \subset \Omega,$$

where $P(\Omega') = 1$. We denote this convergence by

$$X_n \to X, \quad \text{a.s.} \quad \text{as } n \to \infty.$$

Now, the quantities

$$\liminf_{n\to\infty} X_n, \qquad \limsup_{n\to\infty} X_n$$

are random variables, since the functions $\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$ are measurable. If there is a random variable X such that

$$\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n = X \quad a.s.,$$

then $X_n \to X$, a.s. as $n \to \infty$. This follows by the analogous property for a sequence of real numbers.

Next, we consider the convergence in probability as well as convergence a.s. of random elements.

Definition 7. Let $X, X_1, X_2, \ldots$ be random elements in a metric space S, where $d(x, y)$ denotes the metric distance between x and y. The sequence X_n *converges a.s. to* X in S, denoted by $X_n \to X$ a.s., if $d(X_n, X) \to 0$ a.s. The X_n *converges in probability* to X, denoted by $X_n \xrightarrow{P} X$, if

$$\lim_{n\to\infty} P\{d(X_n, X) > \varepsilon\} = 0, \qquad \varepsilon > 0.$$

Many applications involve establishing the convergence of a function $f(X_n)$, when X_n converges. A useful tool for this is the following.

Proposition 8. (Continuous-mapping Property) *Suppose $X, X_1, X_2, \ldots$ are random elements in a metric space S, and $f : S \to S'$ where S' is a metric space. Assume f is continuous on S, or on a set B such that $X \in B$ a.s. If $X_n \to X$ a.s. in S, then $f(X_n) \to f(X)$ a.s. The same statement is true for convergence in probability.*

For instance, if $(X_n, Y_n) \to (X, Y)$ a.s. in $\mathbb{R}^2$, then

$$X_n + Y_n \to X + Y, \qquad X_n Y_n \to XY \qquad \text{a.s. in } \mathbb{R} .$$

This follows since the addition and multiplication functions from $\mathbb{R}^2$ to $\mathbb{R}$ are continuous. Similarly, $X_n/Y_n \to X/Y$ if $Y \neq 0$ a.s. These statements also hold for convergence in probability.

Another important mode of convergence of random variables and random elements is convergence in distribution or weak convergence. A sequence of distributions F_n on $\mathbb{R}$ *converges weakly* to a distribution F, denoted by $F_n \xrightarrow{w} F$, if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for all continuity points x of F. A sequence of random variables X_n *converges in distribution* to a random variable X, denoted by $X_n \xrightarrow{d} X$, if $F_{X_n} \xrightarrow{w} F_X$, as $n \to \infty$. This notion extends to metric spaces as follows.

Definition 9. A sequence of probability measures P_n on a metric space S *converge weakly* to a probability measure P, denoted by[1] $P_n \xrightarrow{w} P$, if

$$\int_S f(x) P_n(dx) \to \int_S f(x) P(dx),$$

[1] Some authors use $P_n \Rightarrow P$ instead of $P_n \xrightarrow{w} P$, and $X_n \Rightarrow X$ instead of $X_n \xrightarrow{d} X$.

for any bounded continuous $f : S \to \mathbb{R}$. Suppose that X and X_n are random elements in S, possibly defined on different probability spaces. The X_n *converges in distribution* to X, denoted by $X_n \xrightarrow{d} X$, if $F_{X_n} \xrightarrow{w} F_X$, or equivalently,

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

for any bounded continuous $f : S \to \mathbb{R}$. This definition is for random elements in the same space S; we do not need the more general convergence for random elements in different spaces.

The preceding modes of convergence for random elements in a metric space S have the hierarchy

$$X_n \to X \text{ a.s.} \quad \Longrightarrow \quad X_n \xrightarrow{P} X \quad \Longrightarrow \quad X_n \xrightarrow{d} X.$$

Here are several characterizations of convergence in distribution.

Theorem 10. (Portmanteau theorem) *For random elements $X, X_1, X_2, \ldots$ in a metric space S, the following statements are equivalent.*
(i) $X_n \xrightarrow{d} X$.
(ii) $\liminf_{n\to\infty} P\{X_n \in G\} \geq P\{X \in G\}$, *for any open set G.*
(iii) $\limsup_{n\to\infty} P\{X_n \in F\} \leq P\{X \in F\}$, *for any closed set F.*
(iv) $P\{X_n \in B\} \to P\{X \in B\}$, *for any Borel set B with $X \neq \partial B$ a.s.*[2]

The convergence in distribution of random variables is equivalent to the convergence of their characteristic functions. Specifically, $X_n \xrightarrow{d} X$ in $\mathbb{R}$ if and only if $E[e^{isX_n}] \to E[e^{isX}]$, $s \in \mathbb{R}$. Similar statements hold for moment generating functions, Laplace transforms or generating functions. The continuous-mapping property in Proposition 8 extends to convergence in distribution as follows.

Theorem 11. (Continuous Mappings; Mann and Wald, Prohorov, Rubin) *Suppose $X_n \xrightarrow{d} X$ in a metric space S, and $B \in \mathcal{S}$ is such that $X \in B$ a.s. Let $f, f_1, f_2, \ldots$ be measurable functions from S to a metric space S'.*
(a) *If f is continuous on B, then $f(X_n) \xrightarrow{d} f(X)$.*
(b) *If $f_n(x_n) \to f(x)$, for any $x_n \to x \in B$, then $f_n(X_n) \xrightarrow{d} f(X)$.*

As an example, for random variables X_n and Y_n, suppose $(X_n, Y_n) \xrightarrow{d} (X, Y)$ as $n \to \infty$. Then

$$X_n + Y_n \xrightarrow{d} X + Y, \qquad X_n Y_n \xrightarrow{d} XY.$$

This follows by the continuous-mapping theorem since addition and multiplication are continuous functions from $\mathbb{R}^2$ to $\mathbb{R}$. Caution: Such results may not be true, however, when X_n and Y_n are real-valued stochastic processes.

[2] ∂B denotes the *boundary* of B.

The next results address the following question for random variables X_n. If $X_n \to X$ a.s. (or in probability or distribution), what are the additional conditions under which $E[X_n] \to E[X]$ or $E|X_n - X| \to 0$?

Lemma 12. (Fatou) *If X_n are nonnegative random variables (or are bounded from below), then* $\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n]$.

Theorem 13. (Monotone Convergence) *If X_n are nonnegative random variables (or are bounded from below) and $X_n \uparrow X$ a.s., then $E[X_n] \uparrow E[X]$ as $n \to \infty$, where $E[X] = \infty$ is possible.*

Theorem 14. (Dominated Convergence) *If X_n are random variables such that $X_n \to X$ a.s., where $|X_n| \leq Y$ and $E[Y] < \infty$, then $E|X|$ exists and $E[X_n] \to E[X]$ as $n \to \infty$.*

These results describing the convergence of $E[X_n] = \int_\Omega X_n(\omega)P(d\omega)$, are random-variable versions of basic theorems for Lebesgue integrals (and for summations as well). We will use the results a few times for real-valued functions f, f_n on a measurable space $(S, \mathcal{S})$ with a measure μ. For instance, the monotone convergence says that if $0 \leq f_n \uparrow f$, then

$$\int_S f_n(x)\mu(dx) \to \int_S f(x)\mu(dx).$$

The next results address the convergence of $E[X_n]$ when X_n converges in probability.

Theorem 15. (Convergence in mean or in L^1) *Suppose $X_n \xrightarrow{P} X$ in $\mathbb{R}$, and $E|X_n|$ and $E|X|$ are finite. Then the following statements are equivalent.*
(i) $E|X_n - X| \to 0$ *(X_n converges to X in L^1).*
(ii) $E|X_n| \to E|X|$.
(iii) *The X_n are* uniformly integrable *in the sense that*

$$\lim_{x\to\infty} \sup_{n\geq 0} E\Big[|X_n|\mathbf{1}(|X_n| \geq x)\Big] \to 0.$$

Some convergence theorems such as the preceding result and the dominated convergence theorem are also true when the assumption that $X_n \to X$ in probability or a.s. is replaced by the weaker assumption $X_n \xrightarrow{d} X$. This is due to the following a.s. representation for convergence in distribution of random elements.

Theorem 16. (Coupling; Skorohod, Dudley) *Suppose $X_n \xrightarrow{d} X$ in a Polish space S. Then there exist random elements Y_n and Y in S, defined on a single probability space, such that $Y_n \stackrel{d}{=} X_n$, $Y \stackrel{d}{=} X$, and $Y_n \to Y$ a.s.*

Loosely speaking, *coupling* is a method for comparing random elements on different probability spaces, usually to prove convergence theorems or

stochastic ordering properties. For instance, suppose X_n is a random element in S_n defined on a probability space $(\Omega_n, \mathcal{F}_n, P_n)$, for $n \geq 0$. Random elements Y_n in S_n defined on a single probability space $(\Omega, \mathcal{F}, P)$ form a *coupling of* X_n if $Y_n \stackrel{d}{=} X_n$, $n \geq 0$.

Theorem 16 and the classical dominated convergence yield the following dominated convergence for convergence in distribution.

Theorem 17. *Suppose* $X_n \stackrel{d}{\to} X$ *in* $\mathbb{R}$ *and there exists a random variable* Y *with finite mean such that* $|X_n| \leq_d Y$*, meaning*

$$P\{|X_n| > x\} \leq P\{Y > x\}, \qquad x \geq 0.$$

Then $E|X|$ *exists and* $E[X_n] \to E[X]$ *as* $n \to \infty$.

Bibliographical Notes

The history of theoretical stochastic processes can be found in books on theoretical probability or stochastic processes. Extensive notes on this are in [46, 61], and miscellaneous references are in the miscellaneous books [10, 18, 26, 37, 38, 41, 42, 62, 63, 84, 96, 102, 106, 114].

The history of applied stochastic processes parallels that of theoretical work. The developments in applied stochastic processes stem from advances in theoretical probability as well as from problems that cry out for solutions. Standard texts on applied stochastic processes are [10, 27, 42, 46, 62, 63, 76, 78, 93, 94, 98], and those with a more specialized focus are [2, 54, 59, 109, 116]. Much of the material on actual applications has been presented in technical reports or in specialized journals that is not conducive to a unified review. So I will only comment on some of the main themes and representative references while discussing the chapters.

Most of Chapter 1 on Markov chains is standard. Exceptions are the reflected random walk (a framework for several models), which is a discrete-time version of the Skorohod equation for reflected Brownian motion [64], and the network models are discrete-time versions of the continuous-time Jackson network models in Chapter 4. Further background and more intricate branching process models are discussed in [49, 58, 62, 63]. References on general Markov processes are the early work [75] and the more current books [38, 37, 41, 42, 61, 84, 96, 106]. Phase-type distributions and computational algorithms, which were not covered, are discussed in [85].

Chapter 2 on Renewal and Regenerative Processes is a fairly thorough coverage of the literature in the 1950s and 1960s, including the works [3, 42, 81, 107] and ending with [69] (only a few articles have appeared after this one). Regenerative processes were popularized with the work of Smith [107]; many applications continue to be based on analyzing processes at embedded times that may or may not be regeneration times. The examples here and in other chapters on reliability and maintenance, production systems, and insurance are indicative of those fields; see for instance [4, 5, 8] and [22, 48, 94].

The richness of the applications of Poisson processes led me to devote an entire chapter to it. Chapter 3 covers classical Poisson processes, but goes further to show the variety of Poisson processes in space and time and their transformations that yield tractable models for systems. Standard references on point processes are the early work of Poisson [87] and later works are [16, 30, 60, 66]; further examples are in [6, 40, 44, 73, 101, 103]. The section on batch-service queueing systems describes a classic Markov decision model [33] (or stochastic dynamic programming model). Dynamic programming is described in [77, 90, 109]. The Markov/Poisson particle model is an elementary example of independent particles [34]; interacting particle systems are discussed in [80]. The Grigelionis theorem [45] showing that Poisson processes are natural limits for sparse point processes, is analogous to the central limit theory for summation processes converging to Brownian motion covered in Chapter 5.

Queueing processes, input-output systems, and stochastic networks have been a main part of applied stochastic processes. This is reflected in Chapter 4 on Continuous-Time Markov Chains. In addition to covering the basics of CTMCs, the chapter provides numerous queueing and network models. Sample references on early work on queueing processes are [20, 28, 40, 44, 57, 68, 70, 71, 72, 82, 91], books on queueing are [2, 6, 12, 13, 15, 28, 29, 39, 47, 59, 74, 88, 108, 116], and works on stochastic networks are [7, 9, 23, 25, 67, 92, 95, 99, 101, 110, 111]. Palm probabilities, that began with Palm [86], are introduced to describe certain PASTA (Poisson actions see time averages) properties [116] of queues and other processes.

The final chapter on Brownian motion covers many of its properties, without getting into more advanced stochastic integration. Books that discuss Brownian motion include [10, 18, 46, 61, 64]. Several key works on Donsker's functional central limit theorem [35, 36] and based on ideas of Prohorov and Skorohod, followed by Billingsley, are [11, 35, 36, 104, 105]. Major applications in queueing and elsewhere are nicely reviewed in Whitt [55, 113]; also see the early work of Kingman 1961 and Iglehart and Whitt 1970 on heavy-traffic queues. During the last three decades, there has been a substantial stream of articles on heavy-traffic systems, Brownian approximations and fluid queues. Much of this research has been done by Harrison and his colleagues Bramson, Chen, Dai, Mandelbaum, Reiman, and Williams [14, 16, 24, 25, 30, 31, 50, 51, 52, 77, 83, 92, 115].

References

1. Ash, R. B. and C. A. Doléans-Dade (2000). *Probability and Measure Theory*, 2nd Ed. Academic Press, San Diego.
2. Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York.
3. Athreya, K., McDonald, D. and P. Ney (1972). Coupling and the renewal theorem. *Amer. Math. Monthly*, **85**, 809–814.
4. Arjas, E. and I. Norros (1989). Change of life distribution via hazard transformation: An inequality with application to minimal repair. *Math Oper. Res*, **14**, 355–361.
5. Aven, R. and U. Jensen (1999). *Mathematical Theory of Reliability*. Springer, New York.
6. Baccelli, F. and P. Brémaud (1994). *Elements of Queueing Theory*. Springer, New York.
7. Baccelli, F. and S. Foss (1994). Ergodicity of Jackson-type queueing networks. *Queueing Systems*, **17**, 5–72.
8. Barlow, R. and F. Proschan (1995). *Mathematical Theory of Reliability*. John Wiley & Sons, New York.
9. Baskett, F., Chandy, K. M., Muntz, R. R. and F. G. Palacios (1975). Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.*, **22**, 248–260.
10. Bhattacharya, R. N. and E. C. Waymire (1990). *Stochastic Processes with Applications*. John Wiley & Sons, New York.
11. Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, New York.
12. Borovkov, A. A. (1976). *Stochastic Processes in Queueing theory*. Springer, New York.
13. Borovkov, A. A. (1984). *Asymptotic Methods in Queueing theory*. John Wiley & Sons, New York.
14. Boxma, O. J. and J. W. Cohen (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems*, **33**, 177–204.
15. Brandt, A., Franken, P. and B. Lisek (1990). *Stationary Stochastic Models*. John Wiley & Sons, New York.
16. Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, **30**, 89–148.
17. Brandt, A. and G. Last (1995). *Marked Point Processes on the Real Line: The Dynamic Approach*. Springer, New York.
18. Breiman, L. (1968). *Probability*. Addison-Wesley, Reading, Mass.
19. Brémaud, P. (1999). *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York.
20. Burke, P. J. (1956). The output of a queuing system. *Operations Res.*, **4**, 699–704.

21. Burman, D. Y., Lehoczky, J. P. and Y. Lim (1984). Insensitivity of blocking probabilities in a circuit-switched network. *J. Appl. Prob.*, **21**, 853–859.
22. Buzzacott, J. A. and J. G. Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*. Prentice–Hall, Englewood Cliffs, New Jersey.
23. Chao, X., Pinedo, M. and M. Miyazawa (1999). *Queueing Networks: Negative Customers, Signals and Product Form*. John Wiley & Sons, New York.
24. Chen, H. and A. Mandelbaum (1994). Heirarchical modeling of stochastic networks, part II: strong approximations. *Stochastic Modeling and Analysis of Manufacturing Systems*. D. D. Yao ed. Springer, New York, 107–131.
25. Chen, H. and D. D. Yao (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimzation*. Springer, New York.
26. Chung, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic Press, New York.
27. Çinlar, E. (1975). *Introduction to Stochastic Processes*. Prentice–Hall, Englewood Cliffs, New Jersey.
28. Cohen, J. W. (1982). *The Single Server Queue*. revised ed. North-Holland, Amsterdam.
29. Cooper, R. B. (1982). *Introduction to Queueing Theory*. 2nd ed. North-Holland, Amsterdam.
30. Dai, J. G. (1994). On positive recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Applied Prob.*, **4**, 49–77.
31. Dai, J. G. and J. M. Harrison (1991). Steady-state analysis of RBM in a rectangle: numerical methods and a queueing application. *Ann. Applied Prob.*, **1**, 16–35.
32. Daley, D. J. and D. Vere-Jones (1988). *An Introduction to the Theory of Point Processes*. Springer, New York.
33. Deb, R. K. and R. Serfozo (1973). Optimal control of batch service queues. *Adv. Appl. Prob.*, **5**, 340–361.
34. Derman, C. (1955). Some contributions to the theory of denumerable Markov chains. *Trans. Amer. Math. Soc.*, **79**, 541–555.
35. Donsker, M. (1951). An invariance principle for certain probability limit theorems. *Mem. Amer. Math. Soc.* **6**.
36. Donsker, M. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.*, **23**, 277–281.
37. Durrett, R. (2005). *Probability: Theory and Examples*, 2nd ed. Duxbury Press, Belmont, CA.
38. Dynkin, E. B. (1965). *Markov Processes, I II*. Springer, New York.
39. El-Taha, M. and S. Stidham Jr. (1999). *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishers, Boston.
40. Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt. Tidskr. Mat. B*, **20**, 33–41.
41. Ethier, S. N. and T. G. Kurtz (1986). *Markov Processes: Characterization and convergence*. John Wiley & Sons, New York.
42. Feller, W. (1968, 1971). *An Introduction to Probability Theory and its Applications, Vol I (3rd ed.), Vol. II (2nd ed.)* John Wiley & Sons, New York.
43. Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.
44. Franken, P., Köenig, D., Arndt, U. and V. Schmidt (1981). *Queues and Point Processes*. Akademie Verlag, Berlin.
45. Grigelionis, B. (1963). On the convergence of sums of random step processes to a Poisson process. *Theory Probab. Appl.*, **8**, 172–182.
46. Grimmett, G. R. and D. R. Stirzaker (2001). *Probability and Random Processes*, 3rd ed. Oxford University Press, Oxford.
47. Gross, D. and C. M. Harris (1985). *Fundamentals of Queueing Theory*. John Wiley & Sons, New York.

48. Hackman, S. T. (2007). *Production Economics*. Springer, New York.
49. Harris, T. E. (1963). *The Theory of Branching Processes*. Springer, New York.
50. Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. John Wiley & Sons, New York.
51. Harrison, J. M. (2000). Brownian models of queueing networks: canonical representation of workload. *Ann. Appl. Prob.*, **10**, 75–103.
52. Harrison, J. M. and R. J. Williams (1996). A multiclass closed queueing network with unconventional heavy traffic behavior. *Ann. Appl. Prob.*, **6**, 1–47.
53. Hewitt, E. and K. Stromberg (1965). *Real and Abstract Analysis*. Springer, New York.
54. Heyman, D. P. and M. J. Sobel (1982, 1984). *Stochastic models in Operations Research, Vol I, Vol II*. McGraw-Hill, New York.
55. Iglehart, D. L. and W. Whitt (1970). Multiple channel queues in heavy traffic, I, II. *Adv. Appl. Prob.*, **2**, 150–177, 355–369.
56. Jacod, J. and A. N. Shiryaev (1987). *Limit Theorems for Stochastic processes*. Springer, New York.
57. Jackson, J. R. (1957). Networks of waiting lines. *Operations Res.*, **5**, 518–552.
58. Jagers, P. (1975). *Branching Processes with Biological Applications*. John Wiley & Sons, New York.
59. Kalashnikov, V. V. (1994). *Mathematical Methods in Queueing Theory*. Kluwer, Dordrecht.
60. Kallenberg, O. (1986). *Random Measures*, 4th ed. Academic Press, London.
61. Kallenberg, O. (2004). *Foundations of Modern Probability*, 2nd ed. Springer, New York.
62. Karlin, S. and H. M. Taylor (1975). *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.
63. Karlin, S. and H. M. Taylor (1981). *A Second Course in Stochastic Processes*, Academic Press, New York.
64. Karatzas, I. and S. Shreve (1991). *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer, New York.
65. Karatzas, I. and S. Shreve (1998). *Methods of Mathematical Finance*, Springer, New York.
66. Karr, A. (1991). *Point Processes and Their Statistical Inference*. 2nd ed. Marcel Dekker, New York.
67. Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley & Sons, New York.
68. Kendall, D. G. (1953). Stochastic processes occuring in the theory of queues and their analysis by means of the imbedded Markov chain. *Ann. Math. Statist.*, **24**, 338–354.
69. Kesten, H. (1974). Renewal theory for Markov chains. *Ann. Probab.*, **3**, 355–387.
70. Khintchine, A. Y. (1960). *Mathmatical Methods in the Theory of Queueing*. Griffin, London.
71. Kiefer, J. and J. Wolfowitz (1956). On the characteristics of the general queueing process with applications to random walks. *Ann. Math. Statist.*, **27**, 147–161.
72. Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Proc. camb. Phil. Soc.*, **57**, 902–904.
73. Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press, Oxford.
74. Kleinrock, L. (1975, 1976). *Queueing Systems. Volume 1 Theory, Volume 2 Computer Applications*. John Wiley & Sons, New York.
75. Kolmogorov, A. N. (1936). Zur Theorie der Markoffschen Ketten. *Math. Ann.*, **112**, 155–160.
76. Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. Chapman and Hall, London.
77. Kushner, H. J. (2001). *Heany Traffic Analysis of Controlled Queueing and Communication Networks*. Springer, New York.

78. Lefebvre, M. (2007). *Applied Stochastic Processes*. Springer, New York.
79. Lewis, P. A. W. and G. S. Shedler (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Navak Res, Losust. Quart.*, **29**, 403–413.
80. Liggett, T. (1972). *Interacting Particle Systems*. Springer, New York.
81. Lindvall, T. (1992). *Lectures on the Coupling Method*. John Wiley & Sons, New York.
82. Little, J. D. C. (1961). A proof for the queueing formula: $L = \lambda W$. *Operations Res.*, **9**, 383–387.
83. Mandelbaum, A. and W. A. Massey (1995). Strong approximation for time-dependent queues. *Math. Oper. Res.*, **20**, 33–64.
84. Meyn, S. and R. L. Tweedie (1993). *Stationary Markov Chains and Stochastic Stability*. Springer, New York.
85. Neuts, M. F. (1994). *Matrix-Geometric Solutions in Stochastic Models*. An algorithmic approach. Corrected reprint of the 1981 original. Dover Publications, Inc., New York.
86. Palm, C. (1943). Intensitätsschwankungen in Fernsprechverkehr. *Ericsson Technics*, **44**, 1-189.
87. Poisson, S. D. (1837). *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, Prédédées des Régles Générales du Calcul des Probabilitiés.* Bachelier, Paris.
88. Prabhu, N. U. (1998). *Queues, Insurance, Dams, and Data*, 2nd ed. Springer, New York.
89. Prokhorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability. *Theory Probab. Appl.*, **1**, 157–214.
90. Putterman, M. L. (1994). *Markov decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.
91. Reich, E. (1957). Waiting times when queues are in tandem. *Ann. Math. Statist.*, **28**, 768–773.
92. Reiman, M. I. (1984). Open queuing networks in heavy traffic. *Math. Oper. Res.*, **9**, 441–458.
93. Resnick, S. I. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston.
94. Rolski, T., Schmidli, H., Schmidt, V. and J. Teugels. (1999). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons, New York.
95. Robert, P. (2003). *Stochastic Networks and Queues*. Springer-Verlag, Berlin.
96. Rogers, L. C. G. and D. Williams (1979-94; 1987). *Diffusions, Markov Processes, and Martingales, Vol. 1 (2dn ed.), and Vol. 2*. John Wiley & Sons, Chichester.
97. Rolski, T. (1981). *Stationary Random Processes Associated with Point Processes*. Lecture Notes in Statistics, **5**. Springer, New York.
98. Ross, S. (1996). *Stochastic Processes*, 2nd ed. John Wiley & Sons, New York.
99. Roussas, G. (1973). *A First Course in Mathematical Statistics*, Addison-Wesley, Reading, MA.
100. Ryll-Nardzewski, C. (1961). Remarks on process of calls. *Proc. 4th Berkeley Symposium, vol* 2. University of California Press, Berkeley, 455–465.
101. Serfozo, R. F. (1999). *Introduction to Stochastic Networks*. Springer, New York.
102. Shiryaev, A. N. (1995). *Probability*, 2nd ed. Springer, New York.
103. Sigman, K. (1995). *Stationary Marked Point Processes: An Intuitive Approach*. Chapman and Hall, New York.
104. Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theory Probab. Appl.*, **1**, 261–290.
105. Skorohod, A. V. (1957). Limit theorems for stochastic processes with independent increments. *Theory Probab. Appl.*, **2**, 122–142.
106. Stroock, D. W. (2005). *An Introduction to Markov Processes*. Graduate Texts in Mathematics 230, Springer, Berlin.
107. Smith, W. L. (1955). Regenerative stochastic processes. *Proc. Roy. Soc., Ser. A*, **232**, 6–31.

108. Takács, L. (1967). *Combinatorial Methods in the theory of Stochastic Processes*. John Wiley & Sons, New York.
109. Tijms, H. (1986). *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley & Sons, New York.
110. Walrand, J. (1988). *Introduction to Queueing Networks*. Prentice–Hall, Englewood Cliffs, New Jersey.
111. Whittle, P. (1986). *Systems in Stochastic Equilibrium*. John Wiley & Sons, New York.
112. Whitt, W. (1983). The queueing network analyzer. *AT&T Bell Labs. Tech. J.*, **62**, 2779–2815.
113. Whitt, W. (2002). *Stochastic Process Limits*. Springer, New York.
114. Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.
115. Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems*, **30**, 27–88.
116. Wolff, R. W. (1989). *Stochastic Modelling and the Theory of Queues*. Prentice-Hall, New York.

Notation

$\mathbf{1}(\cdot)$	The indicator function that is 1 or 0 when $(\cdot)$ is true or false
$A(t) = t - T_{N(t)}$	Backward recurrence time
$B(t) = T_{N(t)+1} - t$	Forward recurrence time
$B,\ B(t)$	Standard Brownian motion process
$C_K^+(S)$	Set of continuous $f : S \to \mathbb{R}_+$ with compact support
$\delta_x(A) = \mathbf{1}(x \in A)$	Dirac measure with unit mass at 1
$\mathbb{D}$	Set of right-continuous, piece-wise constant functions $x : \mathbb{R} \to S$ with left-hand limits, and finite number of jumps in finite intervals
$D = D[0,1],\ D[0,T],\ D(\mathbb{R})$	Set of real valued functions on $[0,1]$, $[0,T]$, $\mathbb{R}$ that are right continuous with left-hand limits
DRI	Directly Riemann integrable
E	Expectation operator
$E[X\|Y],\ E[X\|A]$	Conditional expectations
$E[e^{sX}]$	Moment generating function of X
$\hat{F}(\alpha) = \int_{\mathbb{R}_+} e^{-\alpha t} dF(t)$	Laplace transform of F
$e_i = (0, \ldots, 0, 1, 0 \ldots, 0)$	ith unit vector
$\mathcal{F},\ \mathcal{F}_t,\ \mathcal{F}_n$	σ-field of events
$\mathcal{F}_\tau$	σ-field of events up to stopping time τ
$\mathcal{F}_t^Y$	σ-field of events generated by process Y
$f(t),\ g(t),\ h(t),\ H(t)$	Functions
$f(t) = f^+(t) - f^-(t)$	f equals its positive part minus its negative part
$F(t),\ G(t)$	Distribution functions
$F_e(x) = \frac{1}{\mu}\int_0^x [1 - F(s)]ds$	Equilibrium distribution of F
γ_i	Invariant measure of a CTMC
$H(t) = h(t) + F \star H(t)$	Renewal equation
$\lambda,\ \lambda(A)$	Arrival rate, and rate of entering A
$\mu,\ \mu(A)$	Service rate or mean, and measure of A
$M(t),\ M(A \times B)$	Counting process, or Poisson process

$M(t) = \max_{s \le t} B(s)$	Maxima of Brownian motion B
$N(t)$	Counting process, or renewal process
$N_{\mathcal{T}}$	Point process of $\mathcal{T}$-transitions of a CTMC
$N(\mu, \sigma^2)$	Normal random variable with mean μ and variance σ^2
$\mathrm{P} = (p_{ij})$	Matrix of Markov chain transition probabilities
Random element X in S	X is measurable map from a probability space to S
$P_{\mathcal{T}}(\cdot)$	Palm probability of a $\mathcal{T}$-transition of a CTMC
$p_{ij}(t)$	Transition probability of a CTMC
$p_i,\ p(x)$	Stationary distributions of a CTMC
q_i	Exponential sojourn rate in state i of a CTMC
$q_{ij},\ q(x,y)$	Transition rates of a CTMC
$Q = (q_{ij})$	Transition rate matrix of a CTMC
$\mathbb{R},\ \mathbb{R}_+$	The real numbers and nonnegative real numbers
S	A countable state space for Markov chains in Chapter 1
$S,\ \mathcal{S},\ \hat{\mathcal{S}}$	Polish space and its Borel sets and bounded Borel sets
$\tau,\ \tau_i,\ \tau_i(n)$	Stopping time, and entrance times to state i
$\mathcal{T}$	Subset of sample paths in D
$\mathcal{T}$-transition	A jump time t of a CTMC X with $S^t X \in \mathcal{T}$
T_n	Time of nth event occurrence, or nth renewal time, or time of nth jump in a CTMC
$U(t) = \sum_{n=0}^{\infty} F^{n\star}(t)$	Renewal function
$X_n,\ Y_n$	Markov chains or sequences of random variables
$X(t),\ Y(t),\ Z(t)$	Continuous-time stochastic processes
$\xi_n = T_n - T_{n-1}$	nth inter-renewal time, or time between event occurrences
$\xi_n = T_{n+1} - T_n$	Sojourn time in a CTMC
$W(t),\ W(A)$	Waiting time process or sojourn time in set A
$\mathbb{Z},\ \mathbb{Z}_+$	The integers and nonnegative integers
a.s.	Almost surely, meaning with probability one
$x \vee y = \max\{x, y\}$	Maximum of x and y
$x \wedge y = \min\{x, y\}$	Minimum of x and y
$X(t) \xrightarrow{d} Y$	$X(t)$ converges in distribution to Y as $t \to \infty$
$X \stackrel{d}{=} Y$	The distributions of X and Y are equal
$x^+ = \max\{0, x\}$	Positive part of x
$x^- = -\min\{0, x\}$	Negative part of x and $x = x^+ - x^-$
$\sum_{n=1}^{N(t)}(\cdot) = 0$	When $N(t) = 0$
$\prod_{k=1}^{x}(\cdot) = 1$	When $x = 0$
$\lfloor x \rfloor$	Largest integer $\le x$, or the integer part of x
$\lceil x \rceil$	Smallest integer $\ge x$
$f(t) = o(g(t))$ as $t \to t_0$	$\lim_{t \to t_0} f(t)/g(t) = 0$
$f(t) = O(g(t))$ as $t \to t_0$	$\limsup_{t \to t_0} \lvert f(t)\rvert / \lvert g(t)\rvert < \infty$
$a \Rightarrow b,\ a \Leftrightarrow a$	a implies b, and a is equivalent to b

Index