APPLICATIONS OF MATHEMATICS

STOCHASTIC MODELLING AND APPLIED PROBABILITY

51

Søren Asmussen

Applied Probability and Queues

Second Edition

Springer

Stochastic Mechanics
Random Media
Signal Processing and Image Synthesis
Mathematical Economics and Finance
Stochastic Optimization
Stochastic Control
Stochastic Models in Life Sciences

Applications of Mathematics

Stochastic Modelling and Applied Probability

51

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Applications of Mathematics

1 Fleming/Rishel, **Deterministic and Stochastic Optimal Control** (1975)

2 Marchuk, **Methods of Numerical Mathematics,** Second Ed. (1982)

3 Balakrishnan, **Applied Functional Analysis,** Second Ed. (1981)

4 Borovkov, **Stochastic Processes in Queueing Theory** (1976)

5 Liptser/Shiryayev, **Statistics of Random Processes I: General Theory,** Second Ed. (1977)

6 Liptser/Shiryayev, **Statistics of Random Processes II: Applications,** Second Ed. (1978)

7 Vorob'ev, **Game Theory: Lectures for Economists and Systems Scientists** (1977)

8 Shiryayev, **Optimal Stopping Rules** (1978)

9 Ibragimov/Rozanov, **Gaussian Random Processes** (1978)

10 Wonham, **Linear Multivariable Control: A Geometric Approach,** Third Ed. (1985)

11 Hida, **Brownian Motion** (1980)

12 Hestenes, **Conjugate Direction Methods in Optimization** (1980)

13 Kallianpur, **Stochastic Filtering Theory** (1980)

14 Krylov, **Controlled Diffusion Processes** (1980)

15 Prabhu, **Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication,** Second Ed. (1998)

16 Ibragimov/Has'minskii, **Statistical Estimation: Asymptotic Theory** (1981)

17 Cesari, **Optimization: Theory and Applications** (1982)

18 Elliott, **Stochastic Calculus and Applications** (1982)

19 Marchuk/Shaidourov, **Difference Methods and Their Extrapolations** (1983)

20 Hijab, **Stabilization of Control Systems** (1986)

21 Protter, **Stochastic Integration and Differential Equations** (1990)

22 Benveniste/Métivier/Priouret, **Adaptive Algorithms and Stochastic Approximations** (1990)

23 Kloeden/Platen, **Numerical Solution of Stochastic Differential Equations** (1992)

24 Kushner/Dupuis, **Numerical Methods for Stochastic Control Problems in Continuous Time,** Second Ed. (2001)

25 Fleming/Soner, **Controlled Markov Processes and Viscosity Solutions** (1993)

26 Baccelli/Brémaud, **Elements of Queueing Theory** (1994)

27 Winkler, **Image Analysis, Random Fields, and Dynamic Monte Carlo Methods: An Introduction to Mathematical Aspects** (1994)

28 Kalpazidou, **Cycle Representations of Markov Processes** (1995)

29 Elliott/Aggoun/Moore, **Hidden Markov Models: Estimation and Control** (1995)

30 Hernández-Lerma/Lasserre, **Discrete-Time Markov Control Processes: Basic Optimality Criteria** (1996)

31 Devroye/Györfi/Lugosi, **A Probabilistic Theory of Pattern Recognition** (1996)

32 Maitra/Sudderth, **Discrete Gambling and Stochastic Games** (1996)

33 Embrechts/Klüppelberg/Mikosch, **Modelling Extremal Events** (1997)

34 Duflo, **Random Iterative Models** (1997)

(continued after index)

Søren Asmussen

Applied Probability and Queues

Second Edition

With 46 Illustrations

Springer

Søren Asmussen
Department of Theoretical Statistics
Department of Mathematical Sciences
University of Aarhus
Ny Munkegade
DK-8000 Aarhus, Denmark
asmus@imf.au.dk

Cover illustration: Cover pattern by courtesy of Rick Durrett, Cornell University, Ithaca, New York

Mathematics Subject Classification (2000): 60K25, 60K05

Library of Congress Cataloging-in-Publication Data
Asmussen, Søren.
Applied probability and queues / Søren Asmussen. — 2nd ed.
p. cm. (Applications of mathematics)
Includes bibliographical references and index.
ISBN 0-387-00211-1 (alk. paper)
1. Stochastic processes. 2. Queuing theory. 3. Markov processes. I. Title. II. Series.
QA274 .A825 2003
519.2—dc21 2002042737

ISBN 0-387-00211-1 Printed on acid-free paper.

The first edition of Applied Probability and Queues was published by John Wiley & Sons, Inc.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10901212

Typesetting: Pages created by the author using the Springer LaTEX 2e svsing6.cls macro package.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface

This book treats the mathematics of queueing theory and some related areas, as well as the basic mathematical tools for the study of such models. It thus aims to serve as an introduction to queueing theory, to provide a thorough treatment of tools such as Markov processes, renewal theory, random walks, Lévy processes, matrix–analytic methods and change of measure, and to treat in some detail basic structures such as the $GI/G/1$ and $GI/G/s$ queues, Markov–modulated models, queueing networks, and models within the areas of storage, inventory and insurance risk. Within this framework the choice of topics is, however, rather traditional. The aim has been to present what I consider the basic knowledge in the area, not to advocate special directions in which the area is at present developing.

The first edition was published in 1987. This second edition incorporates about 100 extra pages containing an extended treatment of queueing networks and matrix–analytic methods as well as a number of additional topics, in particular Poisson's equation, the fundamental matrix, insensitivity, rare events and extreme values for regenerative processes, Palm theory, rate conservation, Lévy processes, reflection, Skorokhod problems, Loynes's lemma, Siegmund duality, light traffic, heavy tails, the Ross conjecture and ordering, and finite buffer problems.

Also, the references, typically given in the Notes following the separate sections, have been thoroughly updated. It should be noted, however, that these Notes are mainly intended as a first guidance for further reading, not as a bibliography or history of the subject. When a textbook or a survey paper dealing with a topic is available, this is the preferred reference rather than the original papers. Thus, details of priority are treated rather sporadically. The principle has been to cite only the most important milestones and classical texts, but otherwise to make the references as up–to–date as possible. Thus, compared to the first edition, many older references have been removed.

The reader should be familiar with probability theory at the level of Breiman (1968), Chung (1974), Durrett (1991) or Shiryaev (1996). Most readers are likely to know large parts of Chapters I–II, which therefore may serve mainly as a refresher or reference part. However, one should note that I.5–8 has much material not usually included in introductory texts. How to read the rest of the book is a question of particular interests. The reader oriented towards queueing theory may want to concentrate first on Chapters III–IV and next on X–XII after having skimmed Chapters V, VI and VIII for needed background; the reader with more general interests will find Chapters V–IX and XIII more relevant.

The writing of both the first and the second editions of this book has been an immense pleasure to me. This is due not least to the interest shared by friends, collegues and students. Their impact cannot be overestimated, and the list of people who in some way have influenced the book would be huge. Let me just mention and thank a few who have contributed with detailed comments on the second edition: Niels Hansen, Masakiyo Miyazawa, Mats Pihlsgård, Tomasz Rolski, Volker Schmidt, Karl Sigman and Anders Tolver Jensen. Most figures were done by Jane Bjørn Vedel (supported by MaPhySto, Aarhus) and my mother, Hanna Asmussen, typed much of the material that is close to the first edition.

Finally, I gratefully acknowledge the permission of World Scientific Publishing Co., Singapore, to incorporate some parts (XI.2 and XIII.3) which are close to the exposition in Asmussen (2000).

Søren Asmussen
Aarhus
February 2003

Contents

Preface **v**

Notation and Conventions **xi**

Part A: Simple Markovian Models **1**

I Markov Chains **3**

1 Preliminaries 3

2 Aspects of Renewal Theory in Discrete Time 7

3 Stationarity 11

4 Limit Theory 16

5 Harmonic Functions, Martingales and Test Functions . 20

6 Nonnegative Matrices 25

7 The Fundamental Matrix, Poisson's Equation and the CLT 29

8 Foundations of the General Theory of Markov Processes 32

II Markov Jump Processes **39**

1 Basic Structure 39

2 The Minimal Construction 41

3 The Intensity Matrix 44

4 Stationarity and Limit Results 50

5 Time Reversibility 56

III Queueing Theory at the Markovian Level 60
1 Generalities 60
2 General Birth–Death Processes 71
3 Birth–Death Processes as Queueing Models 75
4 The Phase Method 80
5 Renewal Theory for Phase–Type Distributions 88
6 Lindley Processes 92
7 A First Look at Reflected Lévy Processes 96
8 Time–Dependent Properties of $M/M/1$ 98
9 Waiting Times and Queue Disciplines in $M/M/1$ . . . 108

IV Queueing Networks and Insensitivity 114
1 Poisson Departure Processes and Series of Queues . . . 114
2 Jackson Networks 117
3 Insensitivity in Erlang's Loss System 123
4 Quasi–Reversibility and Single–Node Symmetric Queues 125
5 Quasi–Reversibility in Networks 131
6 The Arrival Theorem 133

Part B: Some General Tools and Methods 137

V Renewal Theory 138
1 Renewal Processes 138
2 Renewal Equations and the Renewal Measure 143
3 Stationary Renewal Processes 150
4 The Renewal Theorem in Its Equivalent Versions . . . 153
5 Proof of the Renewal Theorem 158
6 Second–Moment Results 159
7 Excessive and Defective Renewal Equations 162

VI Regenerative Processes 168
1 Basic Limit Theory 168
2 First Examples and Applications 172
3 Time–Average Properties 177
4 Rare Events and Extreme Values 179

VII Further Topics in Renewal Theory and Regenerative Processes 186
1 Spread–Out Distributions 186
2 The Coupling Method 189
3 Markov Processes: Regeneration and Harris Recurrence 198
4 Markov Renewal Theory 206
5 Semi–Regenerative Processes 211
6 Palm Theory, Rate Conservation and PASTA 213

VIII Random Walks **220**
1 Basic Definitions 220
2 Ladder Processes and Classification 223
3 Wiener–Hopf Factorization 227
4 The Spitzer–Baxter Identities 229
5 Explicit Examples. $M/G/1$, $GI/M/1$, $GI/PH/1$ 233

IX Lévy Processes, Reflection and Duality **244**
1 Lévy Processes 244
2 Reflection and Loynes's Lemma 250
3 Martingales and Transforms for Reflected Lévy Processes 255
4 A More General Duality 260

Part C: Special Models and Methods **265**

X Steady-State Properties of $GI/G/1$ **266**
1 Notation. The Actual Waiting Time 266
2 The Moments of the Waiting Time 269
3 The Workload 272
4 Queue Length Processes 276
5 $M/G/1$ and $GI/M/1$ 279
6 Continuity of the Waiting Time 284
7 Heavy Traffic Limit Theorems 286
8 Light Traffic 290
9 Heavy–Tailed Asymptotics 295

XI Markov Additive Models **302**
1 Some Basic Examples 302
2 Markov Additive Processes 309
3 The Matrix Paradigms $GI/M/1$ and $M/G/1$ 316
4 Solution Methods 328
5 The Ross Conjecture and Other Ordering Results . . . 336

XII Many–Server Queues **340**
1 Comparisons with $GI/G/1$ 340
2 Regeneration and Existence of Limits 344
3 The $GI/M/s$ Queue 348

XIII Exponential Change of Measure **352**
1 Exponential Families 352
2 Large Deviations, Saddlepoints and the Relaxation Time 355
3 Change of Measure: General Theory 358
4 First Applications 362

5 Cramér–Lundberg Theory 365
6 Siegmund's Corrected Heavy Traffic Approximations . 369
7 Rare Events Simulation 373
8 Markov Additive Processes 376

XIV Dams, Inventories and Insurance Risk 380
1 Compound Poisson Dams with General Release Rule 380
2 Some Examples 387
3 Finite Buffer Capacity Models 389
4 Some Simple Inventory Models 396
5 Dual Insurance Risk Models 399
6 The Time to Ruin 401

Appendix 407
A1 Polish Spaces and Weak Convergence 407
A2 Right–Continuity and the Space *D* 408
A3 Point Processes 410
A4 Stochastical Ordering 411
A5 Heavy Tails 412
A6 Geometric Trials 412
A7 Semigroups of Positive Numbers 413
A8 Total Variation Convergence 413
A9 Transforms 414
A10 Stopping Times and Wald's Identity 414
A11 Discrete Skeletons 415

Bibliography 416

Index 431

Notation and Conventions

The basic principle for references within the book is to specify the chapter number only when it is not the current one. Thus, say, Proposition 1.3, formula (2.7) or Section 5 of Chapter IV are referred to as IV.1.3, IV.(2.7) and IV.5, respectively, in all chapters other than IV where we write Proposition 1.3, (2.7) and Section 5.

Symbols such as say A, η, etc. do not of course have the same meaning throughout the book and may be used interchangeably for real numbers, measures and so on. For queueing processes, some effort has been made to make the notation (introduced in III.1) reasonably consistent throughout the book. One inconvenience is that the associated random walk becomes $S_n = X_0 + \cdots + X_{n-1}$ and not $X_1 + \cdots + X_n$ as in Chapter VIII. Of course, similar (hopefully minor) incidents occur at a number of other places.

The expression $\mathbb{E}[X;\, A]$ means $\mathbb{E}XI(A)$, where $I(A)$ is the indicator of A (if say $A = \{X > 0\}$, we write $\mathbb{E}[X;\, X > 0]$). By $X \overset{\mathscr{D}}{=} Y$ we mean equality in distribution and by $X_n \xrightarrow{\mathscr{D}} X$ convergence in distribution (weak convergence). The relation $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$ (other limits may also occur), whereas $a_n \approx b_n$ indicates various different types of asymptotics, often just at the heuristical level. We use occassionally $\overline{\lim}$ instead of $\limsup$, and similarly for $\underline{\lim}$, $\liminf$. Ends of proofs, examples or remarks are marked by the symbol $\square$.

The typeface $\mathbb{P}, \mathbb{E}$ is used for probability and expectation; $\mathbb{P}_e, \mathbb{E}_e$ have a special meaning by referring to stationarity (equilibrium or steady state, cf. III.1). Matrices and vectors are in boldface $\boldsymbol{A}, \boldsymbol{t}, \boldsymbol{\pi}$, etc.; usually, matrices have uppercase Roman letters (occasionally Greek), column vectors lowercase Roman letters and row vectors lowercase Greek letters. The column vector with all entries equal to 1 is denoted $\mathbf{1}$, the ith unit vector $\mathbf{1}_i$. The transpose of $\boldsymbol{A}$ is written $\boldsymbol{A}^{\mathsf{T}}$.

The standard sets are denoted as follows:

$\mathbb{N} = \{0, 1, 2, \ldots\}$	the natural numbers
$\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$	the integers
$\mathbb{Q} = \bigl\{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\backslash\{0\}\bigr\}$	the rationals
$\mathbb{R} = (-\infty, \infty)$	the real numbers
$\mathbb{C} = \{x + \mathrm{i}y : x, y \in \mathbb{R}\}$	the complex numbers

(no special notation like $\mathbb{R}_+$ is used for $(0,\infty)$ or $[0,\infty)$). The index set for the time parameter of a stochastic process, usually $\mathbb{N}$, $\mathbb{Z}$, $[0,\infty)$ or $(-\infty,\infty)$, is denoted by $\mathbb{T}$ if more than one possibility may occur.

The set D of functions $\{x_t\}$ which are right–continuous ($x_s \to x_t$, $s \downarrow t$) and have left–hand limits $x_{t-} = \lim_{s\uparrow t} x_s$ is frequently encountered. If, say, t varies in $[0,1]$ and x_t is E–valued, we may specify this by writing $D([0,1],E)$. Most often D stands for $D[0,\infty) = D\bigl([0,\infty),\mathbb{R}\bigr)$. D_0 is the set of D–functions with finite lifelength; see A2.

Some main abbreviations are given in the following list (others occur locally):

LLN	law of large numbers
CLT	central limit theorem
LIL	law of the iterated logarithm
l.h.s.	left–hand side
r.h.s.	right–hand side
a.s.	almost surely
i.i.d.	independent identically distributed
i.o.	infinitely often
r.v.	random variable
t.v.	total variation
w.l.o.g.	without loss of generality
w.p.	with probability
w.r.t.	with respect to
d.R.i.	directly Riemann integrable
ch.f.	characteristic function
m.g.f.	moment generating function
c.g.f.	cumulant generating function
g.c.d.	greatest common divisor
supp	support
spr	spectral radius

The notation $\widehat{F}$ for the transform of a probability distribution may denote either of the probability generating function, the m.g.f. or the ch.f.; see A9.

The delta function is $\delta_{ij} = I(i = j)$, whereas δ_x often denotes the measure degenerate at x.

Part A:
Simple Markovian Models

I
Markov Chains

1 Preliminaries

We consider a Markov chain $X_0, X_1, \ldots$ with discrete (i.e. finite or countable) state space $E = \{i, j, k, \ldots\}$ and specified by the transition matrix $\boldsymbol{P} = (p_{ij})_{i,j \in E}$. By this we mean that $\boldsymbol{P}$ is a given $E \times E$ matrix such that $\boldsymbol{p}_{i\cdot} = (p_{ij})_{j \in E}$ is a probability (vector) for each i, and that we study $\{X_n\}$ subject to exactly those governing probability laws $\mathbb{P} = \mathbb{P}_{\boldsymbol{\mu}}$ (*Markov probabilities*) for which

$$\mathbb{P}(X_0 = i_0,\, X_1 = i_1, \ldots, X_n = i_n) \;=\; \mu_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} \tag{1.1}$$

where $\mu_i = \mathbb{P}(X_0 = i)$. The particular value of the initial distribution $\boldsymbol{\mu}$ is unimportant in most cases and is therefore suppressed in the notation. An important exception is the case where X_0 is degenerate, say at i, and we write then $\mathbb{P}_i$ so that $\mathbb{P}_i(X_0 = i) = 1$.

Given $\boldsymbol{\mu}$, it is readily checked that (1.1) uniquely determines a probability distribution on $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$. Appealing to basic facts from the foundational theory of Markov processes (to be discussed in Section 8), this set of probabilities can be uniquely extended to a probability law $\mathbb{P}_{\boldsymbol{\mu}}$ governing the whole chain. Thus, since the transition matrix $\boldsymbol{P}$ is fixed here and in the following, the Markov probabilities are in one–to–one correspondence with the set of initial distributions.

If $\mathbb{P}$ is a Markov probability, then (with the usual a.s. interpretation of conditional probabilities and expectations)

$$p_{ij} \;=\; \mathbb{P}_i(X_1 = j) \;=\; \mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i\big), \tag{1.2}$$

$$\begin{aligned}\mathbb{P}(X_{n+1}=j \,|\, \mathscr{F}_n) &= p_{X_n j} = \mathbb{P}_{X_n}(X_1=j), &(1.3)\\ \mathbb{E}\big[h(X_n,X_{n+1},\ldots)\,\big|\,\mathscr{F}_n\big] &= \mathbb{E}_{X_n}h(X_0,X_1,\ldots)\,. &(1.4)\end{aligned}$$

Conversely[1], either (1.3) or (1.4) is sufficient for $\mathbb{P}$ to be a Markov probability. The formal proof of these facts is an easy (though in part lengthy) exercise in conditioning arguments and will not be given here. However, equations (1.2), (1.3), (1.4) have important intuitive contents. Thus (1.4) means that at time n, the chain is restarted with the new initial value X_n. Equivalently, the post–n–chain $X_n, X_{n+1}, \ldots$ evolves as the Markov chain itself, started at X_n but otherwise independent of the past. Similarly, in simulation terminology (1.3) means that the chain can be stepwise constructed by at step n drawing X_{n+1} according to $\boldsymbol{p}_{X_n\cdot}$ (to get started, draw X_0 according to $\boldsymbol{\mu}$).

Recall from A10 (the Appendix) that a *stopping time* σ is a r.v. with values in $\mathbb{N}\cup\{\infty\}$ and satisfying $\{\sigma=n\}\in\mathscr{F}_n$ for all n, that $\mathscr{F}_\sigma$ denotes the σ–algebra which consists of all disjoint unions of the form $\cup_0^\infty A_n$ with $A_n\in\mathscr{F}_n$, $A_n\subseteq\{\sigma=n\}$ (here $n=\infty$ is included with the convention $\mathscr{F}_\infty = \sigma(X_0,X_1,\ldots)$), and that σ and X_σ are measurable w.r.t. to $\mathscr{F}_\sigma$. The important *strong Markov property* states that for the sake of predicting the future development of the chain a stopping time may be treated as a fixed deterministic point of time. For example, we have the following extension of (1.4):

Theorem 1.1 (STRONG MARKOV PROPERTY) *Let σ be a stopping time. Then a.s. on $\{\sigma<\infty\}$ it holds that*

$$\mathbb{E}\big[h(X_\sigma,X_{\sigma+1},\ldots)\,\big|\,\mathscr{F}_\sigma\big] = \mathbb{E}_{X_\sigma}h(X_0,X_1,\ldots). \tag{1.5}$$

Proof. We must show that for $A\in\mathscr{F}_\sigma$, $A\subseteq\{\sigma<\infty\}$ we have

$$\mathbb{E}\big[h(X_\sigma,X_{\sigma+1},\ldots);\,A\big] = \mathbb{E}\big[\mathbb{E}_{X_\sigma}h(X_0,X_1,\ldots);\,A\big].$$

However, if $A\in\mathscr{F}_n$ and $\sigma=n$ on A, this is immediate from (1.4). Replace A by $A\cap\{\sigma=n\}$ and sum over n. □

The mth power (iterate) of the transition matrix is denoted by $\boldsymbol{P}^m = (p_{ij}^m)$. An easy calculation (e.g. let $n=nm$ in (1.1) and sum over the i_k with $k\notin\{0,m,\ldots,nm\}$) shows that $X_0,X_m,X_{2m},\ldots$ is a Markov chain and that its transition matrix is simply $\boldsymbol{P}^m$.

Associated with each state is the hitting time

$$\tau(i) = \inf\{n\ge 1:\, X_n=i\}$$

(with the usual convention $\tau(i)=\infty$ if no such n exists) and the number of visits $N_i=\sum_1^\infty I(X_n=i)$ to i. Clearly, $\{\tau(i)<\infty\}=\{N_i>0\}$ and we

[1] The meaning of (1.4) is that this should hold for any $h:\,E\times E\times\cdots\to\mathbb{R}$ for which (1.4) makes sense, say h is bounded or nonnegative; similarly, (1.5) should hold for all n and j. In (1.3), $\mathbb{P}_{X_n}(X_1=j)$ means $g(x)=\mathbb{P}_x(X_1=j)$ evaulated at $x=X_n$.

call i *recurrent* if the recurrence time distribution $\mathbb{P}_i(\tau(i)=k)$ is proper, i.e. if $\mathbb{P}_i(\tau(i)<\infty)=1$, and *transient* otherwise. The chain itself is recurrent (transient) if all states are so.

Proposition 1.2 *Let i be some fixed state. Then:*
(i) *The following assertions* (a), (b), (c) *are equivalent:* (a) *i is recurrent;* (b) *$N_i=\infty$ $\mathbb{P}_i$–a.s.;* (c) *$\mathbb{E}_iN_i=\sum_1^\infty p_{ii}^m=\infty$;*
(ii) *the following assertions* (a′), (b′), (c′) *are equivalent as well:* (a′) *i is transient;* (b′) *$N_i<\infty$ $\mathbb{P}_i$–a.s.;* (c′) *$\mathbb{E}_iN_i=\sum_1^\infty p_{ii}^m<\infty$.*

Proof. Define $\tau(i;1)=\tau(i)$,

$$\tau(i;k+1) \;=\; \inf\{n>\tau(i;k):\,X_n=i\}, \quad \theta \;=\; \mathbb{P}_i(\tau(i;1)<\infty).$$

Then N_i is simply the number of k with $\tau(i;k)<\infty$, and by the strong Markov property and $X_{\tau(i;k)}=i$,

$$\begin{aligned}\mathbb{P}_i(\tau(i;k+1)<\infty) &= \mathbb{E}_i\mathbb{P}\big(\tau(i;k+1)<\infty,\tau(i;k)<\infty\,\big|\,\mathscr{F}_{\tau(i;k)}\big)\\ &= \mathbb{E}_i\big[\mathbb{P}\big(\tau(i;k+1)<\infty\,\big|\,\mathscr{F}_{\tau(i;k)}\big);\;\tau(i;k)<\infty\big]\\ &= \mathbb{E}_i\big[\mathbb{P}_{X_{\tau(i;k)}}(\tau(i;1)<\infty);\;\tau(i;k)<\infty\big]\\ &= \theta\mathbb{P}_i(\tau(i,k)<\infty) \;=\; \cdots \;=\; \theta^{k+1}. \end{aligned}\tag{1.6}$$

If (a) holds, then $\theta=1$ so that it follows that all $\tau(i;k)<\infty$ $\mathbb{P}_i$–a.s., and (b) also holds. Clearly, (b)⇒(c) so that for part (i) it remains to prove (c)⇒(a) or equivalently (a′)⇒(c′). But if $\theta<1$, then

$$\mathbb{E}_iN_i \;=\; \sum_{k=0}^\infty \mathbb{P}_i(N_i>k) \;=\; \sum_{k=1}^\infty \mathbb{P}_i(\tau(i;k)<\infty) \;=\; \sum_{k=1}^\infty \theta^k \;<\; \infty.$$

For part (ii), it follows by negation that (a′) $\Longleftrightarrow$ (c′) $\Longleftrightarrow$ (b″) $\mathbb{P}_i(N_i<\infty)>0$. However, clearly (b′)⇒(b″) and from (1.6) it is seen that if (b″) holds, then $\theta<1$. Thus (b″) ⇒ (a′). □

It should be noted that though Proposition 1.2 gives necessary and sufficient conditions for recurrence/transience, the criteria are almost always difficult to check: even for extremely simple transition matrices $\boldsymbol{P}$, it is usually impossible to find closed expressions for the p_{ii}^m. Some alternative general approaches are discussed in Section 5, but in many cases the recurrence/transience classification leads into arguments particular for the specific model.

Our emphasis in the following is on the recurrent case and we shall briefly discuss some aspects of the set–up. Two states i,j are said to communicate, written $i\leftrightarrow j$, if i can be reached from j (i.e. $p_{ji}^m>0$ for some m) and vice versa. Clearly, the relation is transitive and symmetric. Now suppose i is recurrent and that j can be reached from i. Then also i can be reached from j. In fact even $\tau(i)<\infty$ $\mathbb{P}_j$–a.s. since otherwise $\mathbb{P}_i(\tau(i)=\infty)>0$.

Furthermore, j is recurrent since

$$\sum_{m=1}^{\infty} p_{jj}^{m} \geq \sum_{m=1}^{\infty} p_{ji}^{m_1} p_{ii}^{m} p_{ij}^{m_2} = \infty$$

if m_1, m_2 are chosen with $p_{ji}^{m_1} > 0$, $p_{ij}^{m_2} > 0$. Obviously $i \leftrightarrow i$ by recurrence, and it follows that $\leftrightarrow$ is an equivalence relation on the recurrent states so that we may write

$$E = T \cup R_1 \cup R_2 \cdots, \tag{1.7}$$

where $R_1, R_2, \ldots$ are the equivalence classes (*recurrent classes*) and T the set of transient states. It is basic to note that the recurrent classes are *closed* (or *absorbing*), i.e.

$$\mathbb{P}_i(X_n \in R_k \text{ for all } n) = 1 \quad \text{when} \quad i \in R_k$$

(this follows from the above characteriztion of R_k as the set of all states that can be reached from i). When started at $i \in R_k$ the chain therefore evolves within R_k only, and the state space may be reduced to R_k. If, on the other hand, $X_0 = i$ is transient, two types of paths may occur: either $X_n \in T$ for all n or at some stage the chain enters a recurrent class R_k and is *absorbed*, i.e. evolves from then on in R_k.

Most often one can restrict attention to *irreducible* chains, defined by the requirement that all states in E communicate. Such a chain is either transient or E consists of exactly one recurrent class. In fact, if a recurrent state, say i, exists at all, it follows from the above that any other state j is in the same recurrence class as i.

A recurrent state is called *positive recurrent* if the mean recurrence time $\mathbb{E}_i\tau(i)$ is finite. Otherwise i is *null recurrent.* The *period* $d = d(i)$ is the period of the recurrence–time distribution, i.e. the greatest integer d such that $\mathbb{P}_i(\tau(i) \in L_d) = 1$ where $L_d = \{d, 2d, 3d, \ldots\}$. If $d = 1$, i is *aperiodic.*

Proposition 1.3 *Let R be a recurrent class. Then the states in R* (i) *are either all positive recurrent or all null recurrent;* (ii) *have all the same period.*

Proof. (i) is deferred to Section 3. Let $i, j \in R$ and choose r, s with $p_{ij}^{r} > 0$, $p_{ji}^{s} > 0$. Then $p_{ii}^{r+s} > 0$, i.e. $r+s \in L_{d(i)}$, and whenever $p_{jj}^{n} > 0$, $p_{ii}^{r+s+n} > 0$ also, i.e. $r + s + n \in L_{d(i)}$ so that $n \in L_{d(i)}$ also. It follows that $\mathbb{P}_j(\tau(j) \in L_{d(i)}) = 1$, i.e. $d(j) \geq d(i)$. By symmetry, $d(i) \geq d(j)$. □

Proposition 1.4 *Let i be aperiodic and recurrent. Then:* (a) *there exists n_i such that $p_{ii}^{m} > 0$ for all $m \geq n_i$;* (b) *if j can be reached from i, then there exists n_j such that $p_{ij}^{m} > 0$ for all $m \geq n_j$.*

Proof. For (a), see A7.1(a). For (b), choose k_j with $p_{ij}^{k_j} > 0$ and let $n_j = n_i + k_j$. □

Problems

1.1 Explain that $\mathbb{P}_{\boldsymbol{\mu}} = \sum_{i\in E} \mu_i \mathbb{P}_i$.
1.2 Show that (1.2) implies (1.4).
1.3 Show that if $\theta = p_{ii} > 0$, then the exit time $\eta(i) = \inf\{n \geq 1 : X_n \neq i\}$ has a geometric distribution, $\mathbb{P}_i(\eta(i) = n) = (1-\theta)\theta^{n-1}$, $n = 1, 2, \ldots$.
1.4 In a number of population processes one encounters Markov chains with $E = \mathbb{N}$, X_n representing the population size at time n, state 0 absorbing and $\mathbb{P}_i(\tau(0) < \infty) > 0$ for all i. Explain why it is reasonable to denote $\{\tau(0) < \infty\}$ as the event of *extinction*. Show that any state $i \geq 1$ is transient and that $X_n \to \infty$ a.s. on the event $\{\tau(0) = \infty\}$ of nonextinction.

Notes In this book, we use the terminology that a Markov chain has discrete time and a Markov process has continuous time (the state space may be discrete as here or general as in Section 8). However, one should note that it is equally common to let "chain" refer to a discrete state space and "process" to a general one (time may be discrete or continuous).

One more convention: the bold typeface for say the initial distribution $\boldsymbol{\mu}$ indicates a representation as a (row) vector, but in many contexts it is more convenient to think of the measure interpretation, and we then write μ. Similarly, a function on the state space may be written either as a (column) vector $\boldsymbol{f} = (f_i)_{i\in E}$ or just as f (with value $f(i)$ at i) We will change freely between these notations; say we use whichever of $\nu(f), \boldsymbol{\nu f}$ which in a given context is convenient to represent $\sum \nu_i f(i)$. Accordingly, we can think of the transition matrix $\boldsymbol{P}$ as an operator acting on measures to the left and on functions to the right, and we sometimes write $\boldsymbol{\nu P}$ as νP and $\boldsymbol{Pf}$ as Pf. A particularly important function is the constant 1 which we write as $\mathbf{1}$ in vector notation.

Markov chains and processes with a discrete state space form in many ways a natural starting point of applied probability: when considering a specific phenomenon, the first attempt to formulate and solve a stochastic model is usually performed within the Markovian set–up, and also the mathematical question arising in connection with Markov chains are to a large extent the same as for more general models (in particular, this is so in queueing theory). The present text therefore starts with a treatment of the relevant features of discrete Markov chains and (in Chapter II) processes. The exposition is in principle self–contained, but the novice will miss examples, and thus the aim is more to provide a refresher and reference, covering also some topics that are not in all textbooks.

We will not list the many textbooks containing introductory chapters on Markov chains and processes. More advanced treatments of discrete Markov chains are in Brémaud (1999), Chung (1967), Freedman (1971), Kemeny *et al.* (1976) and Orey (1971), and of discrete Markov processes in Chung (1967) and Anderson (1991).

2 Aspects of Renewal Theory in Discrete Time

Let $f_1, f_2, \ldots$ be the point probabilities of a distribution on $\{1, 2, \ldots\}$. Then by a (discrete time) *renewal process* governed by $\{f_n\}$ we understand a point process (see A3 for the terminology) on $\mathbb{N}$ with epochs $S_0 = 0$, $S_n = Y_1 + \cdots + Y_n$, where the Y_i are i.i.d. with common distribution $\{f_n\}$. Instead of epochs, we usually speak of *renewals*. The associated *renewal sequence* $u_0, u_1, \ldots$ is defined by $u_k = \mathbb{P}(S_n = k \text{ for some } k \geq 0)$, i.e. the probability of a renewal at k.

A renewal occurs at $k > 0$ if either $Y_1 = k$ which happens w.p. $f_k = f_k u_0$, or if $Y_1 = \ell < k$ and $Y_2 + \cdots + Y_n = k - \ell$ for some n. The probability of this is $f_\ell u_{k-\ell}$, and so

$$u_k \;=\; f_k u_0 + f_{k-1}u_1 + \cdots + f_1 u_{k-1}, \quad k \geq 1, \tag{2.1}$$

i.e. in convolution equation $u = \delta_0 + u * f$ where $\delta_{0i} = I(i = 0)$. In conjunction with $u_0 = 1$, (2.1) clearly uniquely determines $\{u_n\}$.

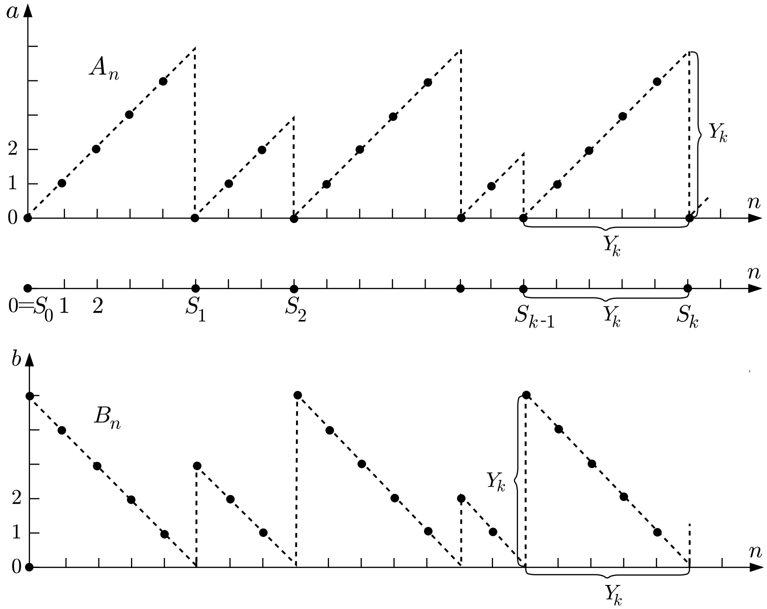

Figure 2.1

These concepts are intimately related to Markov chains. Consider some fixed recurrent state i, let $Y_1 = \tau(i)$ and more generally let Y_k be the inter–occurence time between the $(k-1)$th and kth visit to i. Then $Y_1, Y_2, \ldots$ are i.i.d. w.r.t. $\mathbb{P}_i$ according to the strong Markov property, the common distribution $\{f_n\}$ is the recurrence time distribution of i and the renewals are the visits to i so that $u_n = p_{ii}^n$. Conversely, *any* renewal processs can be constructed in this way from a Markov chain which we shall denote by $\{A_n\}$. Indeed, define $A_n = n - \sup\{S_k : S_k \leq n\}$ as the *backward recurrence time* at n, i.e. the time passed since the last renewal; see Fig. 2.1. Then the paths of $\{A_n\}$ are at 0 exactly at the renewals, i.e. the renewals are the recurrence times of 0, and the Markov property follows by noting that $\{A_n\}$ moves from i to either $i+1$ or 0, the probability of $i+1$ being $\mathbb{P}(Y_k > i+1 \,|\, Y_k > i)$

independently of $A_0, \ldots, A_{n-1}$. The state space E is $\mathbb{N}$ if $\{f_k\}$ has infinite support and $\{0, 1, \ldots, K-1\}$ with $K = \inf\{k : f_1 + \cdots + f_k = 1\}$ otherwise. A closely related important Markov chain is the *forward recurrence time chain* $\{B_n\}$, i.e. B_n is the waiting time until the next renewal after n; see again Fig. 2.1. The Markov property is even more immediate since the paths decrease deterministically from i to $i-1$ if $i > 1$, whereas the value of B_{n+1}, following $B_n = 1$, is chosen according to $\{f_k\}$ independently of the past. The state space is $\{1, 2, \ldots\}$ in the infinite support case and $\{1, \ldots, K\}$ otherwise, and a renewal occurs at n if and only if $B_{n-1} = 1$.

Lemma 2.1 *$\{u_n\}$ and $\{f_n\}$ have the same period d.*

Proof. Since $u_n \geq f_n$, it is clear that the period d_f of $\{f_n\}$ is at least that d_u of $\{u_n\}$. Conversely, it is only possible that $\mathbb{P}(S_k = n) > 0$ and hence $u_n > 0$ if n is a multiple of d_f. Hence $d_u \geq d_f$. □

If $d = 1$ in Lemma 2.1, we will call the renewal sequence (process) *aperiodic.*

Renewal processes with the Y_k having a possible continuous distribution will play a major role in later parts of the book. We shall here exploit the connection between (discrete) renewal processes and Markov chains in the limit theory. Within the framework of renewal processes, the main result is as follows (to be translated to Markov chains in Section 4):

Theorem 2.2 *Let $\{u_n\}$ be an aperiodic renewal sequence governed by $\{f_n\}$ and define $\mu = \sum_1^\infty n f_n = \mathbb{E}Y_1$. Then $u_n \to 1/\mu$ as $n \to \infty$* (here $1/\infty = 0$).

Proof. Define $r_n = f_{n+1} + f_{n+2} + \cdots = \mathbb{P}(Y_1 > n)$ and let L be the index of the last renewal in $\{0, \ldots, n\}$. Then $L = \ell$ if there is a renewal at ℓ and the next Y is $> n - \ell$, i.e. the probability is $u_\ell r_{n-\ell}$ so that

$$1 \;=\; \mathbb{P}(L \leq n) \;=\; r_0 u_n + r_1 u_{n-1} + \cdots + r_n u_0. \tag{2.2}$$

Now let $\lambda = \limsup u_n$ and choose $n(k)$ such that $u_{n(k)} \to \lambda$. Let i satisfy $f_i > 0$. Choosing N such that $r_N < \epsilon$, we obtain from (2.1) and $u_n \leq 1$ that for k sufficiently large

$$\lambda - \epsilon \;\leq\; u_{n(k)} \;\leq\; r_N + \sum_{j=1}^{N} f_j u_{n(k)-j} \tag{2.3}$$

$$\leq\; \epsilon + (1 - f_i)(\lambda + \epsilon) + f_i u_{n(k)-i}. \tag{2.4}$$

Letting first $k \to \infty$ and next $\epsilon \downarrow 0$ yields $\liminf u_{n(k)-i} \geq \lambda$ which is only possible if $u_{n(k)-i} \to \lambda$. Repeating the argument we see that this also holds for any i of the form $i = x_1 a_1 + \cdots + x_t a_t$ where $x_k \in \mathbb{N}$, $f_{a_k} > 0$. But since $\{f_n\}$ is aperiodic, it follows by A7.1(a) (see also Proposition 1.4) that any sufficiently large i, say $i \geq a$, can be represented in this form. Thus letting

$n = n(k) - a$ in (2.2) we obtain for any N

$$1 \ \geq \ \sum_{j=0}^{N} r_j u_{n(k)-a-j} \ \to \lambda \sum_{j=0}^{N} r_j. \tag{2.5}$$

Since $r_0 + r_1 + \cdots = \mu$, this proves $1 \geq \lambda\mu$.

It remains to show that $\nu = \liminf u_n \geq \mu^{-1}$. This is clear if $\mu = \infty$ and can be proved similarly as above if $\mu < \infty$. In fact, if $\{m(k)\}$ is chosen such that $u_{m(k)} \to \nu$, we obtain, instead of (2.3),

$$\begin{aligned} \nu + \epsilon \ &\geq \ u_{m(k)} \ \geq \ \sum_{j=1}^{N} f_j u_{m(k)-j} \ \geq \ (\nu - \epsilon) \sum_{j \leq N, j \neq i} f_j + f_i u_{m(k)-i} \\ &= \ (1 - f_i)(\nu - \epsilon) - r_N(\nu - \epsilon) + f_i u_{m(k)-i}. \end{aligned}$$

As above, this implies $\limsup u_{m(k)-i} \leq \nu$ and $u_{m(k)-i} \to \nu$. Hence for fixed N

$$1 \ \leq \ \sum_{j=0}^{N} r_j u_{m(k)-a-j} + \sum_{j=N+1}^{\infty} r_j \ \to \ \nu \sum_{j=0}^{N} r_j + \sum_{j=N+1}^{\infty} r_j,$$

which tends to $\nu\mu + 0$ as $N \to \infty$. □

Corollary 2.3 *Let $\{u_n\}$, $\{f_n\}$ have period $d > 1$. Then:* (i) *$\{u_{nd}\}_{n=1}^{\infty}$ is an aperiodic renewal sequence governed by $\{f_{nd}\}_{n=1}^{\infty}$;* (ii) *$u_m = 0$ whenever m is not of the form $m = nd$;* (iii) *$u_{nd} \to d/\mathbb{E}Y = d/\mu$ as $n \to \infty$.*

Proof. Here (i) and (ii) are obvious, and from Theorem 2.2 we get

$$u_{nd} \ \to \ \bigl(f_d + 2f_{2d} + 3f_{3d} + \cdots\bigr)^{-1} \ = \ d/\mathbb{E}Y.$$ □

Sometimes one also encounters defective governing distributions $\{f_n\}$, i.e. $f_\infty = 1 - f_1 - f_2 - \cdots > 0$. The corresponding renewal sequence is still uniquely determined by $u_0 = 1$ and (2.1), and can be interpreted in terms of a *terminating* or *transient* renewal process. This is defined simply by attaching the Y_k mass f_∞ at ∞. If $f_\infty > 0$, then $\sigma = \inf\{n \geq 1 : Y_n = \infty\}$ is finite a.s., and $S_n < \infty$ for $n = 0, \ldots, \sigma - 1$, $= \infty$ for $n \geq \sigma$. In particular, the number σ of renewals is finite a.s., and hence the probability u_n of a renewal at n tends to zero as $n \to \infty$. More precisely:

Proposition 2.4 *If $f_\infty > 0$, then the expected number of renewals is given by $\mathbb{E}\sigma = \sum_0^\infty u_n = 1/f_\infty$.*

Proof. Since u_n is the probability of a renewal at n, the expected number of renewals is indeed $\sum_0^\infty u_n$. But it is also

$$\begin{aligned} \mathbb{E}\sigma \ &= \ \sum_{n=1}^{\infty} \mathbb{P}(\sigma \geq n) \ = \ \sum_{n=1}^{\infty} \mathbb{P}(Y_k < \infty, k = 1, \ldots, n-1) \\ &= \ \sum_{n=1}^{\infty} (1 - f_\infty)^{n-1} \ = \ 1/f_\infty. \end{aligned}$$

□

Problems

2.1 Define the generating function of $\{f_n\}$ by $\widehat{f}[s] = \sum_0^\infty s^n f_n$ $(f_0 = 0)$. Show that $\widehat{u}[s] = \sum_0^\infty s^n u_n = (1 - \widehat{f}[s])^{-1}$.
2.2 Consider the geometric case $f_n = (1-\theta)\theta^{n-1}$. Show that u_n is constant for $n > 0$, $u_n = 1 - \theta$.
2.3 Show that $\{u_n v_n\}$ is a renewal sequence if $\{u_n\}$, $\{v_n\}$ are so.
2.4 Let $\{u_n\}$ be a renewal sequence with $\sum_1^\infty f_n \neq 1$. Assume that $\sum_1^\infty \rho^n f_n = 1$ for some ρ. Show that $\{\rho^n u_n\}$ is a renewal sequence and that $u_n \sim c\rho^{-n}$ for some $c \geq 0$ (provided $\{f_n\}$ is aperiodic).

Notes The proof of Theorem 2.2 is a classical argument due to Erdös *et al.* (1949) (many texts today use coupling instead and we return to this in VII.2). Additional material on renewal sequences and related topics can be found in Kingman (1972).

3 Stationarity

Let $\boldsymbol{\nu} = (\nu_i)_{i\in E}$ be any nonnegative measure on E (it is not assumed that $\boldsymbol{\nu}$ is a distribution, $|\boldsymbol{\nu}| = \sum \nu_i = 1$, neither that $\boldsymbol{\nu}$ is finite, $|\boldsymbol{\nu}| < \infty$, but just that all $\nu_i < \infty$). We can then define a new measure $\boldsymbol{\nu P}$ by usual matrix multiplication (viewing $\boldsymbol{\nu}$ as a row vector), so that $\boldsymbol{\nu P}$ attaches mass $\sum_{i\in E} \nu_i p_{ij}$ to j. We call $\boldsymbol{\nu} \neq \boldsymbol{0}$ *stationary* if all $\nu_i < \infty$ and $\boldsymbol{\nu P} = \boldsymbol{\nu}$, i.e. if in algebraic terms $\boldsymbol{\nu}$ is a left eigenvector of the transition matrix $\boldsymbol{P}$ corresponding to the eigenvalue 1.

Of particular importance is the case where $\boldsymbol{\nu}$ is a distribution. Irrespective of whether $\boldsymbol{\nu}$ is stationary or not, we then have

$$\mathbb{P}_{\boldsymbol{\nu}}(X_1 = j) \;=\; \sum_{i\in E} \mathbb{P}_{\boldsymbol{\nu}}(X_0 = i)p_{ij} \;=\; \sum_{i\in E} \nu_i p_{ij} \;=\; (\boldsymbol{\nu P})_j.$$

Thus $\boldsymbol{\nu P}$ can be interpreted as the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of X_1, and in a similar manner the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of X_m is $\boldsymbol{\nu P}^m$. In particular, if $\boldsymbol{\nu}$ is stationary, then $\boldsymbol{\nu P}^m = \boldsymbol{\nu}$ for all m so that the distribution of X_m is independent of m. More generally:

Theorem 3.1 *Suppose that $\boldsymbol{\nu}$ is a stationary distribution. Then:*
(i) *The chain is strictly stationary w.r.t. $\mathbb{P}_{\boldsymbol{\nu}}$, i.e. the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of $(X_n, X_{n+1}, \ldots)$ does not depend on n;*
(ii) *there exists a strictly stationary version $\{X_n\}_{n\in\mathbb{Z}}$ of the chain with doubly infinite time, such that $\mathbb{P}_{\boldsymbol{\nu}}(X_n = i) = \nu_i$ for all $n \in \mathbb{Z}$.*

Proof. (i) Clearly $(X_n, X_{n+1}, \ldots)$ is a Markov chain with transition matrix $\boldsymbol{P}$ w.r.t. $\mathbb{P}_{\boldsymbol{\nu}}$. Then the distribution of the whole sequence is uniquely given by the initial distribution which is $\boldsymbol{\nu P}^n = \boldsymbol{\nu}$, hence independent of n.

(ii) This is a standard construction based upon Kolmogorov's consistency theorem and valid for general stationary sequences: let $\mathbb{P}^{n(1),\ldots,n(k)}$ be the

$\mathbb{P}_{\boldsymbol{\nu}}$–distribution of $\bigl(X_0, X_{n(2)-n(1)}, \ldots, X_{n(k)-n(1)}\bigr)$, $n(1) < n(2) < \cdots < n(k)$, and note that (by stationarity) $\bigl\{\mathbb{P}^{n(1),\ldots,n(k)}\bigr\}$ is a consistent family (see Breiman, 1968, p. 105, for more detail). □

Question of existence and uniqueness of stationary distributions is one of the main topics of Markov chain theory. We start by an explicit construction (a generalization of which will also turn out to be basic for non–Markovian processes; cf. VI.1 and VII.6):

Theorem 3.2 *Let i be a fixed recurrent state. Then a stationary measure $\boldsymbol{\nu}$ can be defined by letting ν_j be the expected number of visits to j in between two consecutive visits to i,*

$$\nu_j \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(i)-1} I(X_n = j) \;=\; \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = j, \tau(i) > n). \qquad (3.1)$$

The proof is based upon the following lemma:

Lemma 3.3 *Let $\boldsymbol{\lambda}$ be an arbitrary initial distribution and σ a stopping time, and define new measures $\boldsymbol{\lambda}(\sigma)$, $\boldsymbol{\mu}(0)$, $\boldsymbol{\mu}(1)$ by $\lambda_j(\sigma) = \mathbb{P}_{\boldsymbol{\lambda}}(X_\sigma = j)$,*

$$\mu_j(0) \;=\; \mathbb{E}_{\boldsymbol{\lambda}} \sum_{n=0}^{\sigma-1} I(X_n = j), \quad \mu_j(1) \;=\; \mathbb{E}_{\boldsymbol{\lambda}} \sum_{n=1}^{\sigma} I(X_n = j).$$

Then $\boldsymbol{\lambda} + \boldsymbol{\mu}(1) = \boldsymbol{\mu}(0) + \boldsymbol{\lambda}(\sigma)$, $\boldsymbol{\mu}(1) = \boldsymbol{\mu}(0)\boldsymbol{P}$.

Proof. The first statement follows by computing $\mathbb{E}_{\boldsymbol{\lambda}} \sum_0^\sigma I(X_n = j)$ by splitting first into the contribution from $n = 0$ and the sum from 1 to σ, and next into the sum from 0 to $\sigma - 1$ and the contribution from $n = \sigma$. The second follows from

$$\begin{aligned}
\mu_j(1) \;&=\; \sum_{n=1}^{\infty} \mathbb{P}_{\boldsymbol{\lambda}}(X_n = j, \sigma \ge n) \;=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\bigl[\mathbb{P}\bigl(X_n = j, \sigma \ge n \,\big|\, \mathscr{F}_{n-1}\bigr)\bigr] \\
&=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\bigl[\mathbb{P}\bigl(X_n = j \,\big|\, \mathscr{F}_{n-1}\bigr);\, \sigma \ge n\bigr] \;=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\bigl[p_{X_{n-1}j};\, \sigma \ge n\bigr] \\
&=\; \sum_{k\in E} p_{kj} \sum_{n=0}^{\infty} \mathbb{P}_{\boldsymbol{\lambda}}(\sigma > n, X_n = k) \;=\; \sum_{k\in E} p_{kj}\mu_k(0) \;=\; (\boldsymbol{\mu}(0)\boldsymbol{P})_j.
\end{aligned}$$

Here in the third step we used the $\mathscr{F}_{n-1}$–measurability of $I(\sigma \ge n)$. □

Proof of Theorem 3.2. If in Lemma 3.3 we take $\boldsymbol{\lambda}$ as the one–point distribution at i and $\sigma = \tau(i)$, we have $\boldsymbol{\mu}(0) = \boldsymbol{\nu}$ and $\boldsymbol{\lambda}(\sigma) = \boldsymbol{\lambda}$. The conclusion of the lemma can be written $\boldsymbol{\lambda} + \boldsymbol{\nu}\boldsymbol{P} = \boldsymbol{\nu} + \boldsymbol{\lambda}$. Hence $\boldsymbol{\nu}\boldsymbol{P} = \boldsymbol{\nu}$, and we need only to check that $\nu_j < \infty$ for any j. Clearly, $\nu_i = 1$ and $\nu_j = 0$ if j is not in the same recurrent class as i. Otherwise observe first that $p_{ji}^m > 0$ for some m so that $\nu_j < \infty$ follows from

$$\nu_i \;=\; \sum_{k\in E} \nu_k p_{ki}^m \;\geq \nu_j p_{ji}^m. \tag{3.2}$$

□

Theorem 3.4 *If the chain is irreducible and recurrent, then a stationary measure $\boldsymbol{\nu}$ exists, satisfies $\nu_j > 0$ for all j and is unique up to a multiplicative constant.*

Here existence is immediate from Theorem 3.2 (we denote in the following the measure in (3.1) by $\boldsymbol{\nu}^{(i)}$). Also, $\nu_i > 0$ for any $i \in E$ and any stationary measure $\boldsymbol{\nu}$ is clear from (3.2) since we may choose j with $\nu_j > 0$. The key step for uniqueness is the following:

Lemma 3.5 *Let i be some fixed state and let $\boldsymbol{\nu}$ be superstationary* (i.e. $\boldsymbol{\nu P} \leq \boldsymbol{\nu}$) *with $\nu_i \geq 1$. Then $\nu_j \geq \nu_j^{(i)}$ for all $j \in E$.*

Proof. With $\widetilde{\boldsymbol{P}}$ the matrix obtained from $\boldsymbol{P}$ by replacing the ith column by zeros, it is easily seen by induction that $\widetilde{p}_{kj^n}$ is the *taboo probability* $\mathbb{P}_k(X_n = j, \tau(i) > n)$. In particular, if we let $k = i$ and sum over n, we get $\boldsymbol{\nu}^{(i)} = \boldsymbol{\delta}^{(i)} \sum_0^\infty \widetilde{\boldsymbol{P}}^n$ where $\boldsymbol{\delta}^{(i)}$ is the distribution degenerate at i. We next claim that $\nu_j \geq \delta_j^{(i)} + (\boldsymbol{\nu}\widetilde{\boldsymbol{P}})_j$. Indeed, for $j = i$ this follows from $\nu_i \geq 1 = \delta_i^{(i)}$, and for $j \neq i$ we have $(\boldsymbol{\nu}\widetilde{\boldsymbol{P}})_j = (\boldsymbol{\nu P})_j \leq \nu_j$. Hence

$$\begin{aligned} \boldsymbol{\nu} \;&\geq\; \boldsymbol{\delta}^{(i)} + \boldsymbol{\nu}\widetilde{\boldsymbol{P}} \;\geq\; \boldsymbol{\delta}^{(i)}(\boldsymbol{I} + \widetilde{\boldsymbol{P}}) + \boldsymbol{\nu}\widetilde{\boldsymbol{P}}^2 \;\geq\; \cdots \\ &\geq\; \boldsymbol{\delta}^{(i)} \sum_{n=0}^{N} \widetilde{\boldsymbol{P}}^n + \boldsymbol{\nu}\widetilde{\boldsymbol{P}}^{N+1} \;\geq\; \boldsymbol{\delta}^{(i)} \sum_{n=0}^{N} \widetilde{\boldsymbol{P}}^n, \end{aligned}$$

and letting $N \to \infty$, $\boldsymbol{\nu} \geq \boldsymbol{\nu}^{(i)}$ follows. □

Proof of Theorem 3.4. If $\boldsymbol{\nu}$ is stationary, then $\nu_i > 0$ as observed above. Thus we may assume $\nu_i = 1$ and the proof will be complete if we can show $\boldsymbol{\nu} = \boldsymbol{\nu}^{(i)}$. But according to the lemma, we have $\boldsymbol{\nu} \geq \boldsymbol{\nu}^{(i)}$. Hence $\boldsymbol{\mu} = \boldsymbol{\nu} - \boldsymbol{\nu}^{(i)}$ is nonnegative and $\boldsymbol{\mu P} = \boldsymbol{\mu}$. As noted above $\mu_i = 0$ then implies $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\nu} = \boldsymbol{\nu}^{(i)}$. □

Clearly, the total mass of the stationary measure $\boldsymbol{\nu}^{(i)}$ given by (3.1) is

$$|\boldsymbol{\nu}^{(i)}| \;=\; \sum_{j\in E} \nu_j^{(i)} \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(i)-1} 1 \;=\; \mathbb{E}_i \tau(i). \tag{3.3}$$

Now if the chain is irreducible and recurrent, it follows by uniqueness that the $|\boldsymbol{\nu}^{(i)}| = \mathbb{E}_i\tau(i)$ are either all finite or all infinite, i.e. that the states are all positive recurrent or all null recurrent, proving the remaining part of Proposition 1.3. In the first case, $\boldsymbol{\nu}$ hence can be normalized to a stationary distribution $\boldsymbol{\pi} = \boldsymbol{\nu}/|\boldsymbol{\nu}|$ which is unique. In particular, for each j we have $\pi_j = \nu_j^{(j)}/|\boldsymbol{\nu}^{(j)}| = 1/\mathbb{E}_j\tau(j)$ which yields an expression for $\boldsymbol{\pi}$ independent of the reference state i. In summary:

Corollary 3.6 *If the chain is irreducible and positive recurrent, there exists a unique stationary distribution $\boldsymbol{\pi}$ given by*

$$\pi_j \;=\; \frac{1}{\mathbb{E}_i\tau(i)}\mathbb{E}_i \sum_{n=0}^{\tau(i)-1} I(X_n = j) \;=\; \frac{1}{\mathbb{E}_j\tau(j)} \tag{3.4}$$

Corollary 3.7 *Any irreducible Markov chain with a finite state space is positive recurrent.*

Proof. With $S_i = \sum_0^\infty I(X_n = i)$, we have $\sum_{i\in E} S_i = \infty$ so that by finiteness $S_i = \infty$ for at least one i. But then i is recurrent, and therefore by irreducibility the chain is recurrent. Since obviously the stationary measure cannot have infinite mass if E is finite, we have positive recurrence. □

Example 3.8 Consider the backward and forward recurrence time chains $\{A_n\}$, $\{B_n\}$ of a renewal process governed by $\{f_n\}$. It is clear from the discussion in Section 2 that both chains are irreducible on the appropriate state spaces. It is also clear that 0 is recurrent for $\{A_n\}$ and 1 for $\{B_n\}$ with $\{f_n\}$ as recurrence time distribution in both cases. In particular, positive recurrence is equivalent to $\mu = \sum nf_n < \infty$. For $\{A_n\}$, the stationary measure (3.1) with $i = 0$ becomes $\nu_n = r_n = f_{n+1} + f_{n+2} + \cdots$, $n = 0, 1, \ldots$. Indeed, n is visited once in between two consecutive visits to 0 if the recurrence time is $\geq n{+}1$. This occurs w.p. r_n and otherwise n is not visited. In particular, if $\mu < \infty$, then the stationary distribution is $\pi_n = r_n/\mu$. In an entirely similar manner it is seen that the stationary measure for $\{B_n\}$ is $\nu_n = r_{n-1}$, $n = 1, 2, \ldots$, and if $\mu < \infty$ then $\pi_n = r_{n-1}/\mu$ defines the stationary distribution. □

The above assumption of irreducibility and recurrence (i.e. one recurrent class) can easily be weakened by invoking the decomposition (1.7) of the state space. For example, if $\boldsymbol{\nu}^{(r)}$ is a stationary measure on the rth recurrent class R_r, it is easy to see that $\boldsymbol{\nu} = \sum_r \boldsymbol{\nu}^{(r)}$ is stationary for the whole chain. Conversely, the restriction of a stationary $\boldsymbol{\nu}$ to R_r is stationary (for the chain restricted to R_r). Also, some *transient* chains have a stationary measure. The theory is more difficult than for the recurrent case and will not be discussed here. We remark only that a stationary *distribution* always attaches mass zero to the transient states because $\mathbb{P}(X_n = i) \to 0$ when i is transient. It is then easy to see that the most general form of a stationary distribution is a convex combination of the unique stationary distributions on the positively recurrent classes.

An alternative proof of the uniqueness of the stationary measure will be given in VII.3. It relies on restricting the Markov chain to a subset F of the state space, a procedure that also has other applications and which we now take the opportunity to discuss briefly. Let $\tau(F;k)$ be the time of the kth visit of $\{X_n\}$ to F, and define $\tau(F) = \tau(F;1)$, $X_k^F = X_{\tau(F;k)}$. In the recurrent case, $\tau(F;k) < \infty$ for all k, and by the strong Markov property $\{X_k^F\}$ is a Markov chain. The transition matrix has elements $p_{k\ell}^F = \mathbb{P}_k(X_{\tau(F)} = \ell)$, $k, \ell \in F$, but these cannot in general be found explicitly in terms of the p_{ij} (but see Problem 3.8). Nevertheless, we have the following result:

Proposition 3.9 *If $\{X_n\}$ is irreducible and recurrent with stationary measure $\boldsymbol{\nu}$, then $\{X_k^F\}$ is also irreducible and recurrent, and the stationary measure $\boldsymbol{\nu}^F = (\nu_\ell^F)_{\ell \in F}$ can be obtained by restricting $\boldsymbol{\nu}$ to F, i.e.* (up to a multiplicative constant) *$\nu_\ell^F = \nu_\ell$, $\ell \in F$. In particular, if $\{X_k^F\}$ is positive recurrent, then the stationary distribution is given by $\pi_\ell^F = \nu_\ell / \sum_{k \in F} \nu_k$.*

Proof. The first assertion is obvious. If we choose the initial state i in (3.1) in F, then both $\{X_n\}$ and $\{X_k^F\}$ visit $\ell \in F$ the same number of times in between visits to i. Hence, also constructing $\boldsymbol{\nu}^F$ according to (3.1) with the same i yields $\nu_\ell^F = \nu_\ell$. □

The formula which conversely expresses $\boldsymbol{\nu}$ in terms of $\boldsymbol{\nu}^F$ (and $\boldsymbol{P}$) is given in VII.5.

Occasionally, the following criterion is useful:

Lemma 3.10 *Let $\{X_n\}$ be irreducible and F a finite subset of the state space. Then the chain is positive recurrent if $\mathbb{E}_i\tau(F) < \infty$ for all $i \in F$.*

Proof. Define $\sigma(i) = \inf\{k \geq 1 : X_k^F = i\}$, $\tau(F;0) = 0$, $Y_k = \tau(F;k) - \tau(F;k-1)$. Then with $m = \max_{j \in F} \mathbb{E}_j \tau(F)$ we have for $i \in F$ that

$$\begin{aligned}
\mathbb{E}_i\tau(i) &= \mathbb{E}_i \sum_{k=1}^{\sigma(i)} Y_k = \sum_{k=1}^{\infty} \mathbb{E}_i\big[\mathbb{E}\big[Y_k \,\big|\, \mathscr{F}_{\tau(F;k-1)}\big];\, k \leq \sigma(i)\big] \\
&\leq m \sum_{k=1}^{\infty} \mathbb{P}_i\big(k \leq \sigma(i)\big) = m\mathbb{E}_i\sigma(i).
\end{aligned}$$

Since E is finite, $\{X_n^F\}$ is positive recurrent. Thus $\mathbb{E}_i\sigma(i) < \infty$, implying $\mathbb{E}_i\tau(i) < \infty$ and positive recurrence of $\{X_n\}$. □

Problems

3.1 Compute a stationary measure if $\boldsymbol{P}$ is *doubly stochastic*, i.e. both the rows and columns sum to 1.

3.2 Show that a Bernoulli random walk ($E = \mathbb{Z}$, $p_{n(n+1)} = \theta$, $p_{n(n-1)} = 1 - \theta$) is doubly stochastic and, if in addition $\theta \neq 1/2$, transient. Show that both $\nu_n = 1$ and $\mu_n = \theta^n/(1-\theta)^n$ are stationary.

3.3 (Continuation of Problem 2.1). Show that the generating function $\widehat{\nu}[s]$ of the stationary measure of the backward recurrence–time chain of a renewal process is given by $\widehat{\nu}[s] = (\widehat{f}[s] - 1)/(s-1)$.

3.4 A set A of states i called an *atom* if $\boldsymbol{p}_{i\cdot}$ is the same for all $i \in A$. Show that $\tau(A)$ is finite $\mathbb{P}_i$–a.s. either for all $i \in A$ or for no $i \in A$, and that in the first case a stationary measure can be defined by

$$\nu_j \;=\; \mathbb{E}_i \sum_{n=1}^{\tau(A)} I(X_n = j) \;\;\text{with } i \in A \text{ arbitrary.}$$

3.5 Consider the recurrence times A_n, B_n of a renewal process. Show that $\{(A_n, B_n)\}$ is Markov with the set of states of the form $(i, 1)$ being an atom, and that the stationary measure is given by $\nu_{ij} = f_{i+j}$.

3.6 Show that $\{(X_n, X_{n+1})\}$ is a Markov chain, and compute the stationary measure in terms of that of $\{X_n\}$.

3.7 Let $\{X_n\}$ have stationary distribution $\boldsymbol{\pi}$ and let $\tau = \inf\{n \geq 1 : X_n = X_0\}$ be the time of return to the initial state. Evaluate $\mathbb{E}_{\boldsymbol{\pi}}\tau$.

3.8 In block notation corresponding to $E = F + F^c$, write the transition matrix as

$$\boldsymbol{P} \;=\; \begin{pmatrix} \boldsymbol{P}_{FF} & \boldsymbol{P}_{FF^c} \\ \boldsymbol{P}_{F^cF} & \boldsymbol{P}_{F^cF^c} \end{pmatrix}.$$

Show that $\left\{X_n^F\right\}$ has transition matrix

$$\boldsymbol{P}^F \;=\; \boldsymbol{P}_{FF} + \boldsymbol{P}_{FF^c}(\boldsymbol{I} - \boldsymbol{P}_{F^cF^c})^{-1}\boldsymbol{P}_{F^cF}.$$

Notes A concept somewhat related to a stationary distribution is that of a *quasi–stationary distribution.* For the precise definition, assume that a special state, say $0 \in E$, is absorbing, and write $E_0 = E\backslash\{0\}$. Then $\boldsymbol{\lambda} = (\lambda_i)_{i\in E_0}$ is called quasi–stationary if $\mathbb{P}_{\boldsymbol{\lambda}}(X_1 = j \,|\, \tau(0) > 1) = \lambda_j$. Closely related are *Yaglom limits*, defined as limits λ_j of $\mathbb{P}_i(X_n = j \,|\, \tau(0) > n)$. A main result in the area states that a (proper) Yaglom limit is necessarily quasi–stationary. However, it is more difficult to assess when a quasi–stationary distribution or a Yaglom limit is unique (the finite case is, however, easy).

Under weak irreducibility conditions, it is trivial to check that when a quasi–stationary distribution $\boldsymbol{\lambda}$ exists, then $\mathbb{P}_{\boldsymbol{\lambda}}(\tau(0) > n) = \theta^n$ where $\theta = \mathbb{P}_{\boldsymbol{\lambda}}(\tau(0) > 1)$ $= \mathbb{P}_{\boldsymbol{\lambda}}(X_1 \neq 0)$. This implies in particular $\mathbb{E}_i R^{\tau(0)} < \infty$ for $R < 1/\theta$. A recent result of Ferrari *et al.* (1995) goes the other way and states that under mild additional conditions, $\mathbb{E}_i R^{\tau(0)} < \infty$ for some $R > 1$ is necessary and sufficient for the existence of a quasi–stationary distribution. Further recent references in the area include Seneta (1994) and Glynn and Thorisson (2001).

4 Limit Theory

The aim is to obtain the limiting behaviour of the p_{ij}^n. We start by noting that this is nontrivial only in the positive recurrent case:

Proposition 4.1 *If state j is either transient or null recurrent, then $p_{ij}^n \to 0$ for any $i \in E$.*

Proof. In the transient case, $I(X_n = j) = 0$ eventually so that the $\mathbb{P}_i$–expectation p_{ij}^n must tend to zero. In the null recurrent case, write

$$p_{ij}^n \;=\; \sum_{k=1}^{n} \mathbb{P}_i(\tau(j) = k) u_{n-k} \;\text{ where }\; u_n = p_{jj}^n. \tag{4.1}$$

Now $\{u_n\}$ is a renewal sequence governed by a distribution by infinite mean and therefore by Corollary 2.3 $u_n \to 0$. Letting $n \to \infty$ in (4.1) and appealing to dominated convergence yields $p_{ij}^n \to 0$. □

Theorem 4.2 (ERGODIC THEOREM FOR MARKOV CHAINS) *Suppose that the chain is irreducible, positive recurrent and aperiodic with stationary distribution* $\boldsymbol{\pi}$. *Then* $p_{ij}^n \to \pi_j$ *for all* j. *That is,* $\boldsymbol{P}^n \to \mathbf{1}\boldsymbol{\pi}$.

Proof. We use again (4.1). By Theorem 2.2, $u_n \to \mu^{-1}$ where μ is the mean recurrence time $\mathbb{E}_j\tau(j) = \pi_j^{-1}$. Appeal to dominated convergence once more to get $p_{ij}^n \to \pi_j$. □

The conclusion is that the limiting distribution of X_n is $\boldsymbol{\pi}$, irrespective of the initial state. Replacing $\mathbb{P}_i$ by $\mathbb{P}_{\boldsymbol{\nu}}$ shows that the same conclusion more generally holds for *any* initial distribution $\boldsymbol{\nu}$.

The case $d > 1$ can be quite easily reduced to the case $d = 1$. To this end, we need the concept of *cyclic classes*, i.e. a partitioning of E into disjoint sets $E_0, \ldots, E_{d-1}$ with the property that the only possible transitions are of the form $E_r \to E_{r+1}$ (here we identify E_d with E_0, E_{d+1} with E_1 and so on).

Proposition 4.3 *Consider an irreducible chain with period* $d > 1$, *let* i *be some arbitrary but fixed state and define*

$$E_r \;=\; \left\{ j \in E : P_{ij}^{nd+r} > 0 \text{ for some } n \geq 0 \right\}, \quad r = 0, \ldots, d-1.$$

Then $E_0, \ldots, E_{d-1}$ *partition* E *into nonempty disjoint sets, and if* $j \in E_r$, *then* $\mathbb{P}_j(X_1 \in E_{r+1}) = 1$ *and more generally* $\mathbb{P}_j(X_m \in E_{r+m}) = 1$. *Furthermore, these properties determine the* E_r *uniquely up to a cyclic rotation.*

Proof. It is obvious that $E_r \neq \emptyset$ (take $n = 0$). By irreducibility, each j is in some E_r so that $\cup_0^{d-1} E_r = E$. Suppose that p_{ij}^{nd+r} and p_{ij}^{md+s} are both > 0, and choose t with $p_{ji}^t > 0$. Then $nd+r+t$ and $md+s+t$ must both be multiples of d, so that $r - s = 0 \pmod d$, showing that the E_r are disjoint. Clearly, $j \in E_r$ and $p_{jk}^m > 0$ implies $k \in E_{r+m}$. Summing over all such k yields $\mathbb{P}_j(X_m \in E_{r+m}) = 1$. Uniqueness is easy and is omitted. □

It follows that if $d > 1$, then the chain $X_0, X_d, X_{2d}, \ldots$ has $E_0, \ldots, E_{d-1}$ as disjoint closed sets. In the irreducible positive recurrent case it is furthermore clear that $\{X_{nd}\}$ is aperiodic positive recurrent on each E_r, i.e. admits a unique stationary distribution $\boldsymbol{\pi}^{(r)}$ concentrated on E_r. Now if $\boldsymbol{\pi}$ is stationary for $\{X_n\}$, its restriction to E_r is also stationary for $\{X_{nd}\}$, and thus by uniqueness $\boldsymbol{\pi}$ is a convex combination $\sum_0^{d-1} \alpha_r \boldsymbol{\pi}^{(r)}$ of the $\boldsymbol{\pi}^{(r)}$. Since

$$\alpha_{r+1} \;=\; \mathbb{P}_{\boldsymbol{\pi}}(X_1 \in E_r) \;=\; \mathbb{P}_{\boldsymbol{\pi}}(X_0 \in E_r) \;=\; \alpha_r,$$

we must even have $\alpha_r = d^{-1}$. Also, the limiting behaviour of p_{jk}^n can easily be seen from $p_{j\ell}^{nd} \to \pi_\ell^{(r)}$ if $j, \ell \in E_r$. Indeed, if $j \in E_r$ then $p_{jk}^{nd+s} = 0$ for all n if $k \notin E_{r+s}$, whereas if $k \in E_{r+s}$ then by dominated convergence

$$p_{jk}^{nd+s} \;=\; \sum_{\ell \in E_{r+s}} p_{j\ell}^s p_{\ell k}^{nd} \;\to\; \sum_{\ell \in E_{r+s}} p_{j\ell}^s \pi_k^{(r+s)} \;=\; \pi_k^{(r+s)} \;=\; d\pi_k. \qquad (4.2)$$

In view of this discussion one can assume aperiodicity in most cases. An irreducible aperiodic positive recurrent chain is simply called *ergodic*.

A further noteworthy property of the stationary distribution is as the limit of time averages (aperiodicity is not required),

$$\frac{1}{n}\sum_{k=0}^{n} f(X_k) \;\to\; \pi(f) \;=\; \boldsymbol{\pi f} \;=\; \mathbb{E}_{\boldsymbol{\pi}} f(X_k) \;=\; \sum_{i \in E} f(i)\pi_i, \qquad (4.3)$$

which holds if f is say bounded or nonnegative. The (easy) proof is carried out in a more general setting in VI.3; a corresponding CLT is in Section 7.

It is reasonable to ask what is the rate of convergence of p_{ij}^n to π_i. In particular, there has been considerable interest in *geometrical ergodicity*, defined by the requirement $p_{ij}^n - \pi_j = \mathrm{O}(\delta^n)$ for some $\delta < 1$ independent of i, j. One has:

Proposition 4.4 (a) *An ergodic Markov chain is geometrically ergodic provided* $\mathbb{E}_i z^{\tau(i)} < \infty$ *for some* $i \in E$ *and some* $z > 1$*;* (b) *any irreducible aperiodic finite Markov chain is geometrically ergodic.*

Proof. Part (a) is a contained in the more general VII.2.11 proved later. For (b), we can choose m_{ki} such that $p_{ki}^m > \pi_i/2$ for all $m \ge m_{ki}$. By finiteness, this implies the existence of $\epsilon > 0$ and $M < \infty$ such that $p_{ki}^m > \epsilon$ for all $m \ge M$ and all k. Hence $\mathbb{P}_i(\tau(i) > (n+1)M \,|\, \tau(i) \ge nM) \le 1 - \epsilon$, and hence by the geometrical trials lemma A6.1 $\mathbb{E}_i z^{\tau(i)} < \infty$ when $z > 1$ is chosen with $z^M(1-\epsilon) < 1$. Now just appeal to (a). □

Also in the null recurrent case it is sometimes possible in various ways to obtain limit statements in terms of the stationary measure which are more refined than just $p_{ij}^n \to 0$. For example:

Proposition 4.5 *If the chain is irreducible recurrent with stationary measure* $\boldsymbol{\nu}$*, then for all* $i, j, k, \ell \in E$

$$\frac{\sum_{n=0}^m p_{ij}^n}{\sum_{n=0}^m p_{\ell k}^n} \;\to\; \frac{\nu_j}{\nu_k}, \quad m \to \infty. \qquad (4.4)$$

For the proof, we need two lemmas (the proof of the first is a straightforward verification and omitted; generalizations are in Problem 5.1 and Section 6).

Lemma 4.6 *The matrix $\widetilde{\boldsymbol{P}}$ with elements $\widetilde{p}_{ij} = \nu_j p_{ji}/\nu_i$ is a transition matrix. Furthermore, the ijth element $\widetilde{p}_{ij}^m$ of $\widetilde{\boldsymbol{P}}^m$ is given by $\widetilde{p}_{ij}^m = \nu_j p_{ji}^m/\nu_i$.*

Lemma 4.7 *Define $N_i^m = \sum_{n=0}^m I(X_n = i)$ as the number of visits to i before time m. Then in the irreducible recurrent case, $\lim_{m\to\infty} \mathbb{E}_j N_i^m / \mathbb{E}_k N_i^m = 1$ for any $j, k \in E$.*

Proof. It may be assumed that $k = i$. By recurrence, $N_i^m \uparrow \infty$ and hence $\mathbb{E}_i N_i^m \uparrow \infty$. Since $N_i^{m-n} = N_i^m + \mathrm{O}(1)$, dominated convergence yields

$$\frac{\mathbb{E}_j N_i^m}{\mathbb{E}_i N_i^m} = \sum_{n=0}^{m} \mathbb{P}_j(\tau(i) = n) \frac{\mathbb{E}_i N_i^{m-n}}{\mathbb{E}_i N_i^m} \to \sum_{n=0}^{\infty} \mathbb{P}_j(\tau(i) = n) = 1. \qquad \square$$

Proof of Proposition 4.5. Consider a Markov chain $\{\widetilde{X}_n\}$ with transition matrix $\widetilde{\boldsymbol{P}}$ given by Lemma 4.6. The expression for $\widetilde{p}_{ij}^n$ shows that $\{X_n\}$ and $\{\widetilde{X}_n\}$ are irreducible at the same time and (sum over n and use Proposition 1.2) recurrent at the same time. Hence $\{\widetilde{X}_n\}$ satisfies the assumptions of Lemma 4.7, and we obtain

$$\begin{aligned} 1 &= \lim_{m\to\infty} \frac{\widetilde{\mathbb{E}}_j N_i^m}{\widetilde{\mathbb{E}}_i N_i^m} \cdot \frac{\mathbb{E}_i N_k^m}{\mathbb{E}_\ell N_k^m} = \lim_{m\to\infty} \frac{\sum_{n=0}^m \widetilde{p}_{ji}^n}{\sum_{n=0}^m \widetilde{p}_{ki}^n} \cdot \frac{\sum_{n=0}^m p_{ik}^n}{\sum_{n=0}^m p_{\ell k}^n} \\ &= \frac{\nu_k}{\nu_j} \lim_{m\to\infty} \frac{\sum_{n=0}^m p_{ij}^n}{\sum_{n=0}^m p_{ik}^n} \cdot \frac{\sum_{n=0}^m p_{ik}^n}{\sum_{n=0}^m p_{\ell k}^n} = \frac{\nu_k}{\nu_j} \lim_{m\to\infty} \frac{\sum_{n=0}^m p_{ij}^n}{\sum_{n=0}^m p_{\ell k}^n}. \end{aligned}$$

$\square$

Notes The terminology "ergodic" as used above is standard, but one should beware not to confuse it with the meaning it has in general stationary process theory (e.g. Breiman, 1968, Ch. 6), namely that the invariant σ–field is trivial. In the Markov chain setting, this does not require aperiodicity, whereas the tail σ–field of a positive recurrent Markov chain being trivial is equivalent to aperiodicity; see e.g. Freedman (1971).

For further results on geometric convergence rates, see VII.2.10. Studying convergence rates via asymptotics of $p_{ij}^n - \pi_j$ as $n \to \infty$ is not the only possible point of view. For example, in a number of models one has observed that $\|\boldsymbol{\nu}\boldsymbol{P}^n - \boldsymbol{\pi}\|$ (t.v. distance) changes from $\|\boldsymbol{\nu} - \boldsymbol{\pi}\|$ to 0 rather abrubtly at a certain time point N, and this N may be a more appropriate measure of the convergence rate than sharp estimates of the deviation of p_{ij}^n from π_j when n is so large that the difference is negligible anyway. Surveys of such broader aspects are in Rosenthal (1995) and Saloff–Coste (1996).

One might expect from Proposition 4.5 and the ergodic theorem for Markov chains that if the chain is also aperiodic, then the *strong ratio property* $p_{ij}^n/p_{\ell k}^n \to \nu_j/\nu_k$ holds. This is, however, not true for all null recurrent chains and presents in fact difficult and not completely solved problems; see Orey (1971). There has also been much discussion of the strong ratio property in relation to quasi–stationarity; see Kesten (1995).

5 Harmonic Functions, Martingales and Test Functions

There is a concept dual to that of a stationary measure, namely that of a *harmonic function* h defined as a right eigenvector $\boldsymbol{h}$ of $\boldsymbol{P}$ corresponding to the eigenvalue 1.[2] The requirement $\boldsymbol{Ph} = \boldsymbol{h}$ means

$$h(i) \;=\; \sum_{j\in E} p_{ij}h(j) \;=\; \mathbb{E}_i h(X_1) \;=\; \mathbb{E}\big[h(X_{n+1})\,\big|\,X_n = i\big],$$

i.e. that $\{h(X_n)\}$ is a martingale. Similarly, one defines h to be *subharmonic* if $\boldsymbol{Ph} \ge \boldsymbol{h}$, i.e. $\{h(X_n)\}$ is a submartingale, and *superharmonic* or *excessive* if $\boldsymbol{Ph} \le \boldsymbol{h}$, i.e. $\{h(X_n)\}$ is a supermartingale.

Proposition 5.1 *If the chain is irreducible and recurrent, then any non-negative superharmonic function h is necessarily constant. Similarly, any bounded subharmonic function h is constant.*

Proof. We must show that $h(i) = h(j)$ for $i \ne j$. Now from the convergence of any non–negative supermartingale we have that $Z = \lim h(X_n)$ exists $\mathbb{P}_i$–a.s. Since $\mathbb{P}_i(X_n = i \text{ i.o.}) = 1$, it follows that $Z = h(i)$ $\mathbb{P}_i$–a.s. Similarly, $\mathbb{P}_i(X_n = j \text{ i.o.}) = 1$ implies that $Z = h(j)$ $\mathbb{P}_i$–a.s. and hence $h(i) = h(j)$. The subharmonic case is similar, using the a.s. convergence of any bounded submartingale. □

When concerned with the recurrent case as in most of this book, the implication is that (super– or sub–) harmonic functions do not play a major role. In the rest of this section we will see, however, that a number of useful recurrence/transience criteria and other properties can be stated in terms of functions h (commonly referred to as *test functions* or *Lyapounov functions*), having properties which are rather similar and allowing for arguments along the lines of the proof of Proposition 5.1.

The problems we study are trivial if E is finite, and in the infinite case we write $h(j) \to \infty$ if the set $\{j :\, h(j) \le a\}$ is finite for any $a < \infty$.

Proposition 5.2 *Suppose the chain is irreducible and let i be some fixed state. Then the chain is transient if and only if there is a bounded nonzero function $h :\, E\backslash\{i\} \to \mathbb{R}$ satisfying*

$$h(j) \;=\; \sum_{k\ne i} p_{jk}h(k), \quad j \ne i. \tag{5.1}$$

Proof. Obviously $h(j) = \mathbb{P}_j(\tau(i) = \infty)$ is bounded and satisfies (5.1). If the chain is transient, then furthermore $h \ne 0$. Suppose, conversely, that there is an h as stated and define $\widetilde{h}(j) = h(j)$, $j \ne i$, $\widetilde{h}(i) = 0$, $\alpha = P\widetilde{h}(i)$. By changing the sign if necessary, we may assume $\alpha \ge 0$ so that $P\widetilde{h}(i) \ge \widetilde{h}(i)$.

[2] See Notes to Section 1 for notation, identication of h with $\boldsymbol{h}$, of $\boldsymbol{Ph}$ with Ph, etc.

Since $P\widetilde{h}(j) = \widetilde{h}(j)$ for $j \neq i$, $\widetilde{h}$ is thus subharmonic. Hence if the chain is recurrent, we have by Proposition 5.1 that $h(j) = \widetilde{h}(j) = \widetilde{h}(i) = 0$ for all $j \neq i$, contradicting $h \neq 0$. Hence the chain is transient. □

Proposition 5.3 *Suppose the chain is irreducible and let E_0 be a finite subset of the state space E. Then:*
(i) *the chain is recurrent if there exists a function $h : E \to \mathbb{R}$ such that $h(x) \to \infty$ and*

$$\sum_{k\in E} p_{jk}h(k) \;\le\; h(j), \quad j \notin E_0. \tag{5.2}$$

(ii) *the chain is positive recurrent if for some $h : E \to \mathbb{R}$ and some $\epsilon > 0$ we have* $\inf_x h(x) > -\infty$ *and*

$$\sum_{k\in E} p_{jk}h(k) \;<\; \infty, \quad j \in E_0, \tag{5.3}$$

$$\sum_{k\in E} p_{jk}h(k) \;\le\; h(j) - \epsilon, \quad j \notin E_0. \tag{5.4}$$

An often encountered compact way to write (5.3)–(5.4) is

$$Ph(j) \;\le\; h(j) - \epsilon + bI(j \in E_0).$$

The intuitive content of (5.2) is that the "center" of the state space in the h–scale corresponds to small values, and that the drift points to the center; similarly, (5.4) can be interpreted as a uniformly positive drift towards the center.

Proof. By adding a constant if necessary, we may assume $h \ge 0$. Write $T = \tau(E_0)$ and define $Y_n = h(X_n)I(T > n)$.
(i) Note first that (5.2) may be rewritten as $\mathbb{E}[h(X_{n+1}) \,|\, X_n = j] \le h(j)$ for $j \notin E_0$. Let $X_0 = i \notin E_0$. Then on $\{T > n\}$, $X_n \notin E_0$ (this fails for $n = 0$ if $X_0 \in E_0$) and hence

$$\begin{aligned} \mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \;&\le\; \mathbb{E}_i\bigl[h(X_{n+1});\, T > n \,\big|\, \mathscr{F}_n\bigr] \\ &=\; I(T > n)\mathbb{E}_i[h(X_{n+1}) \,|\, \mathscr{F}_n] \;\le\; I(T > n)h(X_n) \;=\; Y_n. \end{aligned} \tag{5.5}$$

If $T \le n$, then $Y_n = Y_{n+1} = 0$, and thus $\mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \le Y_n$, i.e. $\{Y_n\}$ is a nonnegative supermartingale and hence converges a.s., $Y_n \overset{\text{a.s.}}{\to} Y_\infty$. Suppose the chain is transient. Then $h(X_n) \le a$ only finitely often, i.e. $h(X_n) \to \infty$, and since $Y_\infty < \infty$, we must have $\mathbb{P}_i(T = \infty) = 0$. But $\mathbb{P}_i(T < \infty) = 1$ for all $i \notin E_0$ implies that some $j \in E_0$ is recurrent, a contradiction.
(ii) Again let $X_0 = i \notin E_0$. Then as in (5.5), we get on $\{T > n\}$ that

$$\mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \;\le\; I(T > n)\mathbb{E}_i[h(X_{n+1}) \,|\, \mathscr{F}_n] \;\le\; Y_n - \epsilon I(T > n).$$

Again, the same is obvious on $\{T \le n\}$ and hence

$$0 \le \; \mathbb{E}_iY_{n+1} \;\le\; \mathbb{E}_iY_n - \epsilon\mathbb{P}_i(T > n) \;\le\; \cdots \;\le \mathbb{E}_iY_0 - \epsilon\sum_{k=0}^{n}\mathbb{P}_i(T > k).$$

Letting $n \to \infty$ and using $Y_0 = h(i)$ yields $\mathbb{E}_i T \le \epsilon^{-1} h(i)$. Thus for $j \in E_0$,

$$\mathbb{E}_j T \;=\; \sum_{i \in E_0} p_{ji} + \sum_{i \notin E_0} p_{ji} \mathbb{E}_i (T+1) \;\le\; 1 + \epsilon^{-1} \sum_{i \notin E_0} p_{ji} h(i)$$

which is finite by (5.3). That the chain is positive recurrent now follows by Lemma 3.10. □

Proposition 5.4 *Suppose the chain is irreducible and let E_0 be a finite subset of E and h a function such that*

$$\sum_{k \in E} p_{jk} h(k) \;\ge\; h(j), \quad j \notin E_0, \tag{5.6}$$

and that $h(i) > h(j)$ for some $i \notin E_0$ and all $j \in E_0$. Then: (i) *if h is bounded, then the chain is transient;* (ii) *if h is bounded below and*

$$\sum_{k \in E} p_{jk} |h(k) - h(j)| \;\le A, \quad j \in E, \tag{5.7}$$

for some $A < \infty$, then the chain is null recurrent or transient .

Proof. Define T as above but let now $Y_n = h(X_{n \wedge T})$. It is then readily verified that $\{Y_n\}$ is a submartingale when $X_0 = i \notin E_0$. In (i), boundedness then implies $Y_n \overset{\text{a.s.}}{\to} Y_\infty$ where $\mathbb{E}_i Y_\infty \ge \mathbb{E}_i Y_0 = h(i)$. But $Y_\infty < h(i)$ on $\{T < \infty\}$ so that $\mathbb{P}_i(T < \infty) < 1$, showing transience.

For (ii), we can choose $j \in E_0$ such that $\alpha = \mathbb{P}_j(\tau(i) < T) > 0$. Then $\mathbb{E}_j \tau(j) \ge \mathbb{E}_j T \ge \alpha \mathbb{E}_i T$ so that is suffices to show $\mathbb{E}_i T = \infty$. Suppose $\mathbb{E}_i T < \infty$. Then in particular, $T < \infty$ $\mathbb{P}_i$–a.s. and by (5.7),

$$\mathbb{E}_i \sum_{n=1}^{T} |Y_n - Y_{n-1}| \;=\; \mathbb{E}_i \sum_{n=1}^{\infty} I(T \ge n) \mathbb{E}\big[|Y_n - Y_{n-1}| \,\big|\, \mathscr{F}_{n-1}\big] \;\le\; A \mathbb{E}_i T \;<\; \infty.$$

Thus we can interchange summation and expectation to get

$$\begin{aligned} \mathbb{E}_i Y_T \;&=\; \mathbb{E}_i Y_0 + \mathbb{E}_i \sum_{n=1}^{T} (Y_n - Y_{n-1}) \;=\; h(i) + \sum_{n=1}^{\infty} \mathbb{E}_i [Y_n - Y_{n-1};\, T \ge n] \\ &=\; h(i) + \sum_{n=1}^{\infty} \mathbb{E}_i \big(I(T \ge n) \mathbb{E}\big[Y_n - Y_{n-1} \,\big|\, \mathscr{F}_{n-1}\big]\big) \;\ge\; h(i), \end{aligned}$$

using the submartingale property in the last step. This is a contradiction since $Y_T < h(i)$. □

Proposition 5.5 *Suppose the chain is irreducible and recurrent, and let E_0 be a finite subset of the state space E. Then the chain is geometrically ergodic if for some $h \ge 0$ with $h(i) > A > 0$, $i \in E_0$, and some $r > 1$*

$$\sum_{k \in E} p_{jk} h(k) \;<\; \infty, \quad j \in E_0, \tag{5.8}$$

$$\sum_{k\in E} p_{jk}h(k) \quad \le \quad h(j)/r, \quad j \notin E_0. \tag{5.9}$$

Proof. Let $X_0 = i \notin E_0$, $Y_n = r^n h(X_{n\wedge T})$. Then it follows easily from (5.9) that $\{Y_n\}$ is a nonnegative supermartingale. By recurrence, the limit is $Y_\infty = r^T h(X_T) \ge Ar^T$. On the other hand, $\mathbb{E}_i Y_\infty \le \mathbb{E}_i Y_0 = h(i)$. For $j \in E_0$, (5.8) then yields

$$\mathbb{E}_j r^T \;\le\; r + r\sum_{i\notin E_0} p_{ji}\mathbb{E}_i r^T \;\le\; 1 + A^{-1}\sum_{i\notin E_0} p_{ji}h(i) \;<\; \infty.$$

It remains to show that $\mathbb{E}_j r^T \le$ for all $j \in E_0$ implies geometric ergodicity. By Proposition 4.4, this will follow if we can show $\mathbb{E}_i r^T < \infty$ for all $i \in E_0$. This in turn follows by a variant of the proof of Lemma 3.10, left as Problem 5.3. □

Proposition 5.6 *Suppose the chain is irreducible and positive recurrent with stationary distribution π, and let f, g, h be nonnegative functions on E such that*

$$\sum_{j\in E} p_{ij}h(j) \;\le\; h(i) - f(i) + g(i), \quad i \in E. \tag{5.10}$$

If $\pi(g) < \infty$, $\pi(h) < \infty$, then also $\pi(f) < \infty$.

Proof. We can rewrite (5.10) as $f \le h - Ph + g$. Thus $P^k f \le P^k h - P^{k+1}h + P^k g$ and for any i,

$$\sum_{k=1}^{n} P^k f(i) \le \; Ph(i) - P^{n+1}h(i) + \sum_{k=1}^{n} P^k g(i) \le \; Ph(i) + \sum_{k=1}^{n} P^k g(i).$$

Applying π to the left and noting that $\pi(Ph)/n = \pi(h)/n \to 0$ yields $\pi(f) \le \pi(g) < \infty$. □

Example 5.7 Consider a queue where service takes place at a discrete sequence of instants $n = 0, 1, 2, \ldots$, let X_n be the queue length at time n, B_n the number of customers arriving between n and $n+1$ and A_n the maximal number of customers that can be served at the $(n+1)$th service epoch. Thus with $Y_n = B_n - A_n$

$$X_{n+1} \;=\; (X_n + Y_n)^+, \tag{5.11}$$

a recurrence relation (the Lindley recursion) also typical for many other queueing situations and discussed in length in III.6. For example, this could describe the queue at the stop of a bus with regular schedule, with A_n the number of free seats in the nth bus.

Assume further that the random vectors (A_n, B_n) are i.i.d.; then $\{X_n\}$ is a Markov chain on $\mathbb{N}$. Let $\mu = \mathbb{E}Y_n$. With $h(i) = i$, (5.11) then yields $\mathbb{E}_i W_1 = \mathbb{E}(i + Y_1)^+ \ge i + \mu$. Thus, if $\mu \ge 0$, Proposition 5.4(ii) shows immediately that $\{X_n\}$ cannot be positive recurrent. Suppose on the other hand that $\mu < 0$ and let $\mu_i = \mathbb{E}[Y_n;\, Y_n > -i]$. Then $\mu_i \to \mu$, $i \to \infty$, and hence for i so large, say $i > i_0$, that $\mu_i \le \mu/2$,

$$\mathbb{E}(i+Y_1)^+ \;=\; \mathbb{E}[i+Y_1;\, Y_1 > -i] \;\le\; i+\mu_i \;\le\; i+\mu/2.$$

Thus Proposition 5.3(ii) with $E_0 = \{0,\ldots,i_0\}$, $h(i) = i$, $\epsilon = -\mu/2$ yields positive recurrence.

For geometrical ergodicity, assume $\mu < 0$ and that $\mathbb{E}z^{B_1} < \infty$ for some $z > 1$. By replacing z by a smaller z if necessary, we may assume $r_1 = \mathbb{E}z^{Y_1} < 1$. We have $\mathbb{E}z^{i+Y_1} = z^i r_1$, and as above, one then gets $\mathbb{E}_i z^{W_1} < z^i r$ for $i \ge i_0$ and some $r \in (r_1, 1)$. Thus Proposition 5.5 with $h(i) = z^i$ yields geometric ergodicity.

Finally, assume $\mu < 0$, $\mu_2 = \mathbb{E}Y_n^2 < \infty$. With $h(i) = i^2$, we then have $\mathbb{E}h(i+Y_1) = h(i)+\mu_2+2i\mu$. As above, this implies $Ph(i) \le h(i)-f(i)+g(i)$ for $i \ge i_0$ where $g(i) = \mu_2/2$, $f(i) = -i\mu$. Since $\pi(g) < \infty$, Proposition 5.6 yields $\pi(f) < \infty$. I.e., the stationary distribution has finite mean when $\mu_2 < \infty$ [see further X.2]. □

Problems

5.1 (DOOB'S h–TRANSFORM) Suppose the chain is irreducible and $h \ge 0$ harmonic with $h \ne 0$. Show that $h(i) > 0$ for all i and that the matrix $\widetilde{\boldsymbol{P}}$ with elements $\widetilde{p}_{ij} = h(j)p_{ij}/h(i)$ is a transition matrix.

5.2 Consider a population process satisfying the assumptions of Problem 1.4 and with all states $i, j \ge 1$ communicating. Show that the extinction probability $q_i = \mathbb{P}_i(\tau(0) < \infty)$ is either 1 for all $i \ge 1$ or 0 for all $i \ge 1$. Let E_0 be finite and suppose (5.2) holds. Show that $q_i = 1$ if $h(j) \to \infty$ and that $q_i < 1$ if h is bounded with $h(i) < h(j)$ for some $i \notin E_0$ and all $j \in E_0$. [*Hint*: Consider $\{\widetilde{X}_n\}$ evolving as $\{X_n\}$ except that $\widetilde{p}_{01} = 1$ rather than $\widetilde{p}_{01} = 0$, and use h as test function for $\{\widetilde{X}_n\}$.] Show in particular that if $\mathbb{E}[X_{n+1} \,|\, X_n] \le X_n$ (i.e. the expected number of children per individual does not exceed 1), then extinction occurs a.s.

5.3 Carry out the last part of the proof of Proposition 5.5.

Notes Results of type Proposition 5.3, 5.4(i) have a long history and are often referred to as *Foster's criteria*. A main reference for test function techniques is Meyn and Tweedie (1993), who also treat the case of an uncountable state space (essentially, all results carry over at the cost of more tedious proofs and formulations). It is known that many of the sufficient conditions given above are also necessary in the sense that a test function with the stated properties must exist. However, finding the appropriate one is far from easy in more complicated models; Brémaud (1999) surveys a number of examples dealing with nonstandard queueing models. See also Fayolle *et al.* (1995).

6 Nonnegative Matrices

Finite square matrices with nonnegative elements occur in a variety of contexts in applied probability. The so–called *Perron–Frobenius theory* of such matrices describes in quite some detail their spectral properties (and therefore also the asymptotic properties of their powers), and is therefore a powerful and indispensable tool for many applications. We shall here develop this theory by exploiting the intimate connection to Markov chains with a finite number of states.

We start by recalling some facts from linear algebra. Let $\boldsymbol{A}$ be any $p \times p$ matrix and define for $\lambda \in \mathbb{C}$ $E_\lambda = \{\boldsymbol{x} \in \mathbb{C}^p : \boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}\}$. Thus $\mathrm{sp}(\boldsymbol{A}) = \{\lambda : E_\lambda \neq \emptyset\}$ is the set of eigenvalues of $\boldsymbol{A}$ or the *spectrum* of $\boldsymbol{A}$, and $\mathrm{spr}(\boldsymbol{A}) = \sup\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{A})\bigr\}$ is the *spectral radius* of $\boldsymbol{A}$. If $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then λ is a root in the characteristic polynomial $\det(\boldsymbol{A} - \lambda\boldsymbol{I})$, and if the multiplicity is 1, we call λ *simple*. Then also the geometric multiplicity $\dim\bigl(E_\lambda \cup \{0\}\bigr)$ is 1, i.e. the eigenvector is unique up to a constant. If $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then λ is also eigenvalue for the transposed matrix $\boldsymbol{A}^\mathsf{T}$. The existence of an eigenvector for $\boldsymbol{A}^\mathsf{T}$ then means that $\boldsymbol{\nu}\boldsymbol{A} = \lambda\boldsymbol{\nu}$ for some row vector $\boldsymbol{\nu} \neq \boldsymbol{0}$, called a *left eigenvector* for $\boldsymbol{A}$ ($\boldsymbol{x} \in E_\lambda$ is a *right eigenvector*). The following lemma is standard (all statements are easy to verify if one writes $\boldsymbol{A}$ on the Jordan canonical form):

Lemma 6.1 (i) $\mathrm{sp}(\boldsymbol{A}^m) = \bigl\{\lambda^m : \lambda \in \mathrm{sp}(\boldsymbol{A})\bigr\}$; (ii) *the* $\boldsymbol{A}^m$*–multiplicity of* $\lambda \in \mathrm{sp}(\boldsymbol{A})$ *is the sum of the* $\boldsymbol{A}$*–multiplicities of the* $\lambda_i \in \mathrm{sp}(\boldsymbol{A})$ *with* $\lambda_i^m = \lambda$; (iii) *if* $\lambda \in \mathrm{sp}(\boldsymbol{A})$ *is not simple, then either* $\dim\bigl(E_\lambda \cup \{0\}\bigr) > 1$ *or for any* $\boldsymbol{h} \in E_\lambda$ *we can find* $\boldsymbol{k}$ *with* $\boldsymbol{A}\boldsymbol{k} = \boldsymbol{h} + \lambda\boldsymbol{k}$; (iv) $\boldsymbol{A}^n = \mathrm{O}\bigl(n^k[\mathrm{spr}(\boldsymbol{A})]^n\bigr)$ *for some* $k = 0, 1, 2, \ldots$.

We start by examining the spectral properties of ergodic transition matrices:

Proposition 6.2 *Let* $\boldsymbol{P} = (p_{ij})_{i,j=1,\ldots,p}$ *be an ergodic* $p \times p$ *transition matrix with stationary distribution* $\boldsymbol{\pi}$. *Then* $\mathrm{spr}(\boldsymbol{P}) = 1$ *and* 1 *is a simple eigenvalue of* $\boldsymbol{P}$ *with* $\boldsymbol{\pi}$ *and* $\boldsymbol{1} = (1 \cdots 1)^\mathsf{T}$ *as corresponding left and right eigenvectors. Furthermore for* $\lambda \in \mathrm{sp}(\boldsymbol{P})$, $\lambda \neq 1$, *we have* $|\lambda| < 1$ *and with* $\lambda_1 = \max\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{P}), \lambda \neq 1\bigr\}$ *it holds for some* k *that the powers* $\boldsymbol{P}^n = (p_{ij}^n)$ *satisfy*

$$p_{ij}^n \;=\; \pi_j + \mathrm{O}(n^k\lambda_1^n), \quad n \to \infty. \tag{6.1}$$

Proof. It is clear that $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$, $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{1}$ and hence $1 \in \mathrm{sp}(\boldsymbol{P})$. Also $\boldsymbol{h} \in E_1$ means that $\boldsymbol{h}$ is harmonic and thus $\boldsymbol{h} = c\boldsymbol{1}$ (cf. Proposition 5.1; the extension to the complex case is easy). Thus if 1 is not simple, Lemma 6.1(iii) shows that we can find $\boldsymbol{k}$ with $\boldsymbol{P}\boldsymbol{k} = \boldsymbol{1} + \boldsymbol{k}$. But then $\boldsymbol{P}^n\boldsymbol{k} = n\boldsymbol{1} + \boldsymbol{k}$ which in Markov chain terms means that $\mathbb{E}_i k_{X_n} = n + k_i$, contradicting that $\boldsymbol{k}$ is bounded in the finite case. Similarly, the ergodic theorem means that $\boldsymbol{P}^n \to \mathbf{1}\boldsymbol{\pi}$ and hence if $\lambda \in \mathrm{sp}(\boldsymbol{P})$, $\boldsymbol{k} \in E_\lambda$, we have $\lambda^n \boldsymbol{k} = \boldsymbol{P}^n \boldsymbol{k} \to \mathbf{1}\boldsymbol{\pi}\boldsymbol{k}$. But $\lambda^n \boldsymbol{k}$ can only converge if $|\lambda| < 1$ or $\lambda = 1$.

It only remains to prove (6.1). Write $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$ with $\boldsymbol{P}_1 = \mathbf{1}\boldsymbol{\pi}$, $\boldsymbol{P}_2 = \boldsymbol{P} - \mathbf{1}\boldsymbol{\pi}$. It is then readily checked that $\boldsymbol{P}_1\boldsymbol{P}_2 = \boldsymbol{P}_2\boldsymbol{P}_1 = \mathbf{0}$ and hence $\boldsymbol{P}^n = \boldsymbol{P}_1^n + \boldsymbol{P}_2^n$. It is also easily seen that $\boldsymbol{P}_1^n = \boldsymbol{P}_1 = \mathbf{1}\boldsymbol{\pi}$. Hence by Lemma 6.1(iv) it suffices to show that if $\lambda \in \mathrm{sp}(\boldsymbol{P}_2)\setminus\{0\}$, then $|\lambda| \le \lambda_1$. But from $\boldsymbol{P}_2\boldsymbol{k} = \lambda\boldsymbol{k}$ we get $\boldsymbol{P}\boldsymbol{k} = (\boldsymbol{P}_1 + \boldsymbol{P}_2)\lambda^{-1}\boldsymbol{P}_2\boldsymbol{k} = \lambda\boldsymbol{k}$, i.e. $\lambda \in \mathrm{sp}(\boldsymbol{P})$. If $\lambda = 1$, we would have $\boldsymbol{k} = c\mathbf{1}$ and hence $\boldsymbol{P}_2\boldsymbol{k} = 0$ which is impossible. Hence $|\lambda| \le \lambda_1$. □

If $\lambda_1 < \delta < 1$, (6.1) may be rewritten as $p_{ij}^n = \pi_j + \mathrm{O}(\delta^n)$, and we have obtained a second proof of Proposition 4.4(b), stating that any irreducible finite Markov chain is geometrically ergodic.

A matrix $\boldsymbol{Q}$ is *nonnegative* if $q_{ij} \ge 0$ for all i, j, and *substochastic* if also $\boldsymbol{Q}\mathbf{1} \le \mathbf{1}$, i.e. the rows sums are at most 1. The following result is often used and holds under weaker conditions than irreducibility:

Proposition 6.3 *Let $\boldsymbol{Q}$ be substochastic, such that to each i there is a k and $j_1, \ldots, j_m$ with $\sum_\ell q_{k\ell} < 1$ and $q_{ij_1}q_{j_1j_2}\ldots q_{j_mk} > 0$. Then* $\mathrm{spr}(\boldsymbol{Q}) < 1$.

Proof. Let λ be an eigenvalue of absolute value $\mathrm{spr}(\boldsymbol{Q})$ and let $\boldsymbol{h} \in E_\lambda$. Consider a Markov chain $\{X_n\}$ on $\{0, 1, \ldots, p\}$ such that 0 is absorbing, and the probability of a transition $i \to j$ is q_{ij} for $i, j \ge 1$ and $1 - \sum_\ell q_{i\ell}$ for $j = 0$. The assumptions on $\boldsymbol{Q}$ and a geometrical trials argument (cf. A6.1) then easily yield that $X_n = 0$ eventually and that taking $h_0 = 0$ makes $\lambda^{-n}h_{X_n}$ a martingale. If $|\lambda| \ge 1$, boundedness would imply L_1–convergence (necessarily to h_0) so that taking $X_0 = i$ yields $h_i = h_0 = 0$ which contradicts $\boldsymbol{h} \ne \mathbf{0}$. Hence $|\lambda| < 1$ and $\mathrm{spr}(\boldsymbol{Q}) < 1$. □

We shall now derive a close analogue of Proposition 6.2 for nonnegative matrices $\boldsymbol{A}$. We shall adopt the definitions of irreducibility and the period d from transition matrices to nonnegative matrices by noting that they depend only on the pattern of entries i, j with $a_{ij} > 0$. Thus $\boldsymbol{A}$ is *irreducible* if for any i, j we can find m such that $a_{ij}^m > 0$, and we have:

Lemma 6.4 *If $\boldsymbol{A}$ is an irreducible nonnegative matrix, then the greatest common divisor d of the m with $a_{ii}^m > 0$ does not depend on i. If $d = 1$, then it holds for all sufficiently large m that $a_{ij}^m > 0$ for all i, j.*

Proof. Choose a transition matrix $\boldsymbol{P}$ with $p_{ij} > 0$ for exactly the same i, j as for which $a_{ij} > 0$. Then $a_{ij}^m > 0$ precisely when $p_{ij}^m > 0$ and results from Section 1 complete the proof. □

The d in Lemma 6.4 is called the *period* of $\boldsymbol{A}$, and $\boldsymbol{A}$ is *aperiodic* if $d = 1$.

Theorem 6.5 (PERRON–FROBENIUS) *Let $\boldsymbol{A}$ be an irreducible non–negative $p \times p$ matrix. Then:*
(a) *the spectral radius λ_0 of $\boldsymbol{A}$ is strictly positive and a simple eigenvalue of $\boldsymbol{A}$ with the corresponding left and right eigenvectors $\boldsymbol{\nu}$, $\boldsymbol{h}$ satisfying $\nu_i > 0$,*

$h_i > 0$ for all i;
(b) *if $\boldsymbol{A}$ is also aperiodic, then $\lambda_1 = \max\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{A})\backslash\{\lambda_0\}\bigr\} < \lambda_0$. Furthermore, if we normalize $\boldsymbol{\nu}, \boldsymbol{h}$ by $\boldsymbol{\nu h} = \sum_1^p \nu_i h_i = 1$, then for some k*

$$\boldsymbol{A}^n \;=\; \lambda^n \boldsymbol{h\nu} + \mathrm{O}(n^k \lambda_1^n), \quad n \to \infty; \tag{6.2}$$

(c) *if $\boldsymbol{A}$ has period $d > 1$, then $|\lambda| \le \lambda_0$ for any $\lambda \in \mathrm{sp}(\boldsymbol{A})$. Furthermore, $\lambda \in \mathrm{sp}(\boldsymbol{A})$, $|\lambda| = \lambda_0$ holds exactly when λ is of the form $\lambda_0\theta^k$, $k = 0, 1, \ldots, d-1$, with $\theta^k = \mathrm{e}^{2\pi k/d}$ the roots of unity.*

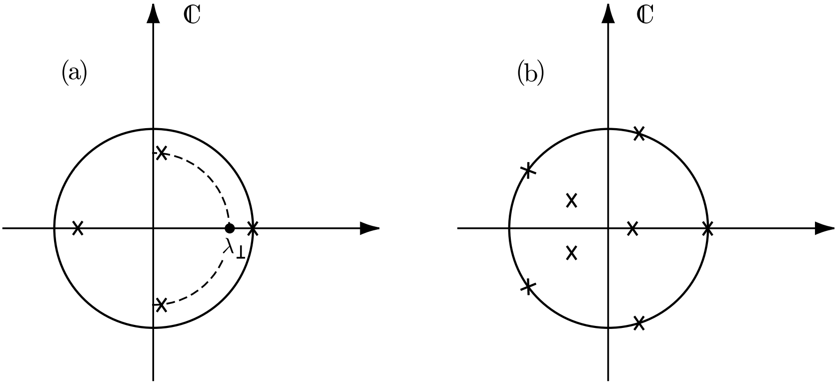

Figure 6.1

Figure 6.1 depicts sp(A) for the aperiodic case in (a) and for the periodic case $d = 5$ in (b). The eigenvalues fall in pairs of complex conjugates since $\boldsymbol{A}$ is real. We shall refer to λ_0 as the *Perron–Frobenius root* of $\boldsymbol{A}$.

The proof of the Perron–Frobenius theorem will be reduced to the Markov case in Proposition 6.2. We need some lemmas.

Lemma 6.6 *If $\boldsymbol{A}$ has all $a_{ij} > 0$, then there exists $\lambda \in \mathrm{sp}(\boldsymbol{A})$, $x \in E_\lambda$ with $\lambda > 0$, $x_i > 0$, $i = 1, \ldots, p$.*

Proof. The basic observation is that all $a_{ij} > 0$ implies

$$x_i \ge 0,\ \sum_{i=1}^p x_i > 0 \quad \Rightarrow \quad \text{all components of } \boldsymbol{Ax} \text{ are } > 0. \tag{6.3}$$

Define

$$\begin{aligned} K &= \Bigl\{\boldsymbol{x} \in \mathbb{R}^p : 0 \le x_i \le 1, \sum_{i=1}^p x_i = 1\Bigr\}, \\ S &= \bigl\{\mu \ge 0 : \boldsymbol{Ax} \ge \mu\boldsymbol{x} \text{ for some } \boldsymbol{x} \in K\bigr\}, \end{aligned}$$

$\lambda = \sup\{\mu : \mu \in S\}$. Since $\boldsymbol{A}K$ is compact, $\lambda < \infty$. For a given $\boldsymbol{x} \in K$, (6.3) implies $\boldsymbol{Ax} \ge \epsilon\boldsymbol{x}$ for small enough ϵ, and hence $\lambda > 0$. Now choose $\lambda_n \in S$, $\boldsymbol{x}_n \in K$ with $\lambda_n \uparrow \lambda$, $\boldsymbol{Ax}_n \ge \lambda_n\boldsymbol{x}_n$. Passing to a subsequence if necessary, we may assume that $\boldsymbol{x} = \lim \boldsymbol{x}_n$ exists. Then $\boldsymbol{Ax} \ge \lambda\boldsymbol{x}$ and we shall complete the proof by showing that indeed $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ ($x_i > 0$ is then ensured by (6.3)). Otherwise let $\boldsymbol{y} = c\boldsymbol{A}\boldsymbol{x}$ with $c > 0$ chosen so that $\boldsymbol{y} \in K$. Then $\boldsymbol{A}\boldsymbol{y} - \lambda\boldsymbol{y} = c\boldsymbol{A}(\boldsymbol{A}\boldsymbol{x} - \lambda\boldsymbol{x})$ has all components > 0 by (6.3). Hence $\boldsymbol{A}\boldsymbol{y} \geq (\lambda + \epsilon)\boldsymbol{y}$ for some $\epsilon > 0$, a contradiction. □

Lemma 6.7 *Suppose that $\boldsymbol{A}\boldsymbol{k} = \lambda\boldsymbol{k}$ with $\lambda > 0$ and all $k_i > 0$. Then the matrix $\boldsymbol{P}$ with elements $a_{ij}k_j/\lambda k_i$ is a transition matrix, $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{1}$, and the formulas*

$$\lambda^{\boldsymbol{A}} = \lambda\lambda^{\boldsymbol{P}}, \quad h_i^{\boldsymbol{A}} = k_i h_i^{\boldsymbol{P}}, \quad \pi_i^{\boldsymbol{A}} = \pi_i^{\boldsymbol{P}}/k_i$$

establish a one–to–one correspondence between $\lambda^{\boldsymbol{A}} \in$ sp$(\boldsymbol{A})$ and $\lambda^{\boldsymbol{P}} \in$ sp$(\boldsymbol{P})$ and the corresponding right and left eigenvectors ($\boldsymbol{\pi}^{\boldsymbol{A}}\boldsymbol{A} = \lambda^{\boldsymbol{A}}\boldsymbol{\pi}^{\boldsymbol{A}}$ etc.). Furthermore, $\lambda^{\boldsymbol{A}}$ is simple for $\boldsymbol{A}$ if and only if $\lambda^{\boldsymbol{P}}$ is simple for $\boldsymbol{P}$.

Proof. Everything is a straightforward verification except for the last statement which follows from

$$\det(\boldsymbol{P} - \mu\boldsymbol{I}) \;=\; \det(\lambda^{-1}\boldsymbol{A} - \mu\boldsymbol{I}) \;=\; \lambda^{-p}\det(\boldsymbol{A} - \mu\lambda\boldsymbol{I}).$$

Indeed, multiplying the ith row by k_i and the jth column by k_j^{-1} leaves the determinant unchanged and transform $\boldsymbol{P}$ into $\lambda^{-1}\boldsymbol{A}$, $\boldsymbol{I}$ into $\boldsymbol{I}$. □

Proof of Theorem 6.5 *in the aperiodic case.* Choose first m with all $a_{ij}^m > 0$, cf. Lemma 6.4, and next λ, k with $\boldsymbol{A}^m\boldsymbol{k} = \lambda\boldsymbol{k}$, $\lambda > 0$, all $k_i > 0$, cf. Lemma 6.6. Then by Lemma 6.7 1 is simple for $\boldsymbol{P}^m = (a_{ij}^m k_j/k_i)$ and hence λ simple for $\boldsymbol{A}^m$. If $\lambda_0 \in$ sp$(\boldsymbol{A})$ satisfies $\lambda_0^m = \lambda$, then by Lemma 6.1(ii) λ_0 is simple for $\boldsymbol{A}$. Choose $\boldsymbol{h} \in E_{\lambda_0}$. Then $\boldsymbol{A}^m\boldsymbol{h} = \lambda_0^m\boldsymbol{h} = \lambda\boldsymbol{h}$, and since λ is simple for $\boldsymbol{A}^m$, it follows that we may take $\boldsymbol{h} = \boldsymbol{k}$. Then by nonnegativity, $\boldsymbol{A}\boldsymbol{h} = \lambda_0\boldsymbol{h}$ implies $\lambda_0 > 0$ and $\boldsymbol{P} = (a_{ij}k_j/\lambda_0 k_i)$ is a transition matrix. Applying Proposition 6.2 and Lemma 6.7 everything then comes out in a straightforward manner. For (6.2), note that if $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$, $\boldsymbol{\pi}\boldsymbol{1} = 1$ and we let $\nu_i = \pi_i/h_i$, then $\boldsymbol{\nu}\boldsymbol{A} = \lambda_0\boldsymbol{\nu}$, $\boldsymbol{\nu}\boldsymbol{h} = 1$ and

$$a_{ij}^n \;=\; \lambda_0^n \frac{p_{ij}^n h_i}{h_j} \;=\; \lambda_0^n \frac{h_i}{h_j}\left\{\pi_j + \mathrm{O}\Big(n^k\Big(\frac{\lambda_1}{\lambda_0}\Big)^n\Big)\right\} \;=\; \lambda_0^n h_i\nu_j + \mathrm{O}(n^k\lambda_1^n).$$

□

Proof of Theorem 6.5 *in the periodic case.* We can reorder the coordinates by a cyclic class argument so that $\boldsymbol{A}$ has the form

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{A}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & & \boldsymbol{0} \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_{d-1} \\ \boldsymbol{A}_d & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} \end{pmatrix}.$$

Letting $\boldsymbol{B}_k = \boldsymbol{A}_k\boldsymbol{A}_{k+1}\cdots\boldsymbol{A}_d\boldsymbol{A}_1\cdots\boldsymbol{A}_{k-1}$, it follows that $\boldsymbol{A}^d$ is block–diagonal with diagonal elements $\boldsymbol{B}_k$ which are irreducible aperiodic. Let μ_k be the Perron–Frobenius root of $\boldsymbol{B}_k$ and $\boldsymbol{B}_k\boldsymbol{h}^{(k)} = \mu_k\boldsymbol{h}^{(k)}$ with $h_i^{(k)} > 0$.

Now

$$\boldsymbol{B}_k\boldsymbol{A}_k\boldsymbol{h}^{(k+1)} \;=\; \boldsymbol{A}_k\boldsymbol{B}_{k+1}\boldsymbol{h}^{(k+1)} \;=\; \mu_{k+1}\boldsymbol{A}_k\boldsymbol{h}^{(k+1)}$$

(identifying $d+1$ with 1). Since $\boldsymbol{A}_k\boldsymbol{h}^{(k+1)} \neq \boldsymbol{0}$, it follows that $\mu_{k+1} \in \mathrm{sp}(\boldsymbol{B}_k)$ and hence $\mu_{k+1} \le \mu_k$. Hence all μ_k are equal, say $\mu_k = \mu$, and we may take $\boldsymbol{h}^{(k)} = \boldsymbol{A}_k\boldsymbol{h}^{(k+1)} = \boldsymbol{A}_k \ldots \boldsymbol{A}_{d-1}\boldsymbol{h}^{(d)}$. Now

$$\det(\boldsymbol{A}^d - \eta\boldsymbol{I}) \;=\; \prod_{k=1}^{d}\det(\boldsymbol{B}_k - \eta\boldsymbol{I}).$$

This shows that if $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then $\eta = \lambda^d$ is in $\mathrm{sp}(\boldsymbol{B}_k)$ for some k. Hence $|\lambda| = |\eta|^{1/d} \le |\mu|^{1/d} = \lambda_0$ (say) and $|\lambda| = \lambda_0$ can only occur if $\lambda^d = \mu$, i.e. λ is of the form $\lambda_0\theta^k$ for some k. Also the $\boldsymbol{A}^d$–multiplicity of μ is exactly d. By Lemma 6.1(ii) the proof is now complete if we can show that each $\lambda_0\theta^k$ is an eigenvalue and that $\boldsymbol{z}^{(0)} \in E_{\lambda_0}$ may be taken with all $z_i^{(0)} > 0$. But an easy calculation shows that

$$\boldsymbol{z}^{(k)} \;=\; \Bigl((\lambda_0\theta^k)^0\boldsymbol{h}^{(1)^{\mathsf{T}}} \;\cdots\; (\lambda_0\theta^k)^{d-1}\boldsymbol{h}^{(d)^{\mathsf{T}}}\Bigr)^{\mathsf{T}}$$

satisfies $\boldsymbol{A}\boldsymbol{z}^{(k)} = \lambda_0\theta^k\boldsymbol{z}^{(k)}$. □

Problems

6.1 Is it true that if $\boldsymbol{P}$ is an *infinite* ergodic transition matrix, then all $p_{ij}^n > 0$ for some n?

6.2 Suppose that $\boldsymbol{A}$ is an irreducible aperiodic nonnegative matrix such that $\boldsymbol{A}^m$ is a transition matrix for some $m = 1, 2, \ldots$ Show that then $\boldsymbol{A}$ is itself a transition matrix. Show also that the result fails in the periodic case.

6.3 Let $\boldsymbol{A}$ be irreducible and nonnegative, and assume that $\boldsymbol{A}\boldsymbol{x} \le \lambda\boldsymbol{x}$ with $\boldsymbol{x} \ge \boldsymbol{0}$, $\boldsymbol{x} \neq \boldsymbol{0}$ and $\lambda > 0$. Show that $\mathrm{spr}(\boldsymbol{A}) \le \lambda$ provided either (i) $\boldsymbol{A}$ is irreducible, or (ii) all $x_i > 0$. Show also in case (i) that $\mathrm{spr}(\boldsymbol{A}) < \lambda$ if in addition $\boldsymbol{A}\boldsymbol{x} \neq \lambda\boldsymbol{x}$.

Notes Standard references for nonnegative matrices are Berman and Plemmons (1994) and Seneta (1994). Of extensions of the Perron–Frobenius theorem, we mention in particular operator versions such as the Krein–Rutman theorem, e.g. Schaefer (1970), and the more probabilistic inspired discussion of Nummelin (1984).

7 The Fundamental Matrix, Poisson's Equation and the CLT

We assume throughout this section that $\{X_n\}$ is irreducible positive recurrent with stationary distribution $\boldsymbol{\pi}$.

Let f be a real–valued function on E, sometimes written as a column vector $\boldsymbol{f}$ (see Notes to Section 1 for this and other notational issues like $\pi(f)$ versus $\boldsymbol{\pi f}$, Pf versus $\boldsymbol{Pf}$, etc.). The equation

$$\boldsymbol{g} \;=\; \boldsymbol{f} + \boldsymbol{Pg}, \tag{7.1}$$

with $\boldsymbol{g}$ the unknown, is referred to as *Poisson's equation.*

Proposition 7.1 *Assume that f is π–integrable. Then:* (i) *a necessary condition for the existence of a $\boldsymbol{\pi}$–integrable solution to Poisson's equation is $\pi(f) = 0$;* (ii) *a π–integrable solution is unique up to a multiple of* $\mathbf{1}$*;* (iii) *if $\pi(f) = 0$, then for any k $g(i) = \mathbb{E}_i \sum_0^{\tau(k)-1} f(X_n)$ is a finite solution satisfying $g(k) = 0$.*

Proof. Multiplying (7.1) by $\boldsymbol{\pi}$ immediately gives (i). If $\boldsymbol{g}_1, \boldsymbol{g}_2$ are solutions, then $\boldsymbol{d} = \boldsymbol{g}_1 - \boldsymbol{g}_2$ satifies $\boldsymbol{d} = \boldsymbol{Pd}$, i.e. $\boldsymbol{d}$ is harmonic and must therefore be constant by Proposition 5.1, showing (ii). In (iii), we have from Corollary 3.6 that

$$\pi(|f|)\mathbb{E}_k\tau(k) \;=\; \mathbb{E}_k \sum_{n=0}^{\tau(k)-1} |f|(X_n) \;\geq\; \mathbb{P}_k\bigl(\tau(j) < \tau(k)\bigr)\,\mathbb{E}_j \sum_{n=0}^{\tau(k)-1} |f|(X_n).$$

This shows first that $g(j)$ is finite and next, upon replacing $|f|$ by f in the left identity, that $g(k) = 0$. Conditioning upon X_1 and using the definition of g then gives

$$g(i) \;=\; f(i) + \sum_{j\neq k} p_{ij} g(j) \;=\; f(i) + \sum_{j\in E} p_{ij} g(j) \;=\; f(i) + Pg(i),$$

which is the same as (7.1). □

Theorem 7.2 *Let $\boldsymbol{f}$ be a $\boldsymbol{\pi}$–integrable function on E and define $\widetilde{f}(i) = f(i) - \pi(f)$. Assume that g is a solution of $g = \widetilde{f} + Pg$ and that $\pi(g^2) < \infty$. Then*

$$\frac{1}{\sqrt{n}}\bigl(f(X_0) + \cdots + f(X_{n-1}) - n\pi(f)\bigr) \;\overset{\mathscr{D}}{\to}\; N(0, \sigma^2(f)) \tag{7.2}$$

where $\sigma^2(f) = \pi(g^2) - \pi\bigl((Pg)^2\bigr)$.

Proof. We may assume w.l.o.g. that $\pi(f) = 0$ so that $f = \widetilde{f}$. Let $\Delta_k = g(X_k) - Pg(X_{k-1})$. Then $g = f + Pg$ implies

$$\sum_{k=0}^{n} f(X_k) \;=\; g(X_0) - Pg(X_n) + \sum_{k=1}^{n} \Delta_k. \tag{7.3}$$

Since $Pg(X_{k-1}) = \mathbb{E}(g(X_k) \,|\, \mathscr{F}_{k-1})$, the sequence $\{\Delta_k\}$ is a martingale difference sequence, and we have

$$\mathbb{V}ar(\Delta_k \,|\, \mathscr{F}_{k-1}) \;=\; \mathbb{V}ar(g(X_k) \,|\, \mathscr{F}_{k-1}) \;=\; \omega^2(X_{k-1})$$

where $\omega^2(i) = g^2(i) - (Pg)^2(i)$. Here $\pi(\omega^2) = \sigma^2(f)$ is finite by assumption so that $\sum_1^n \mathbb{V}ar(\Delta_k \,|\, \mathscr{F}_{k-1})/n \to \sigma^2(f)$ by the LLN (4.3). Therefore an appropriate martingale CLT (e.g. Hall and Heyde, 1980, p. 58, or Shiryaev, 1996, p. 541) shows that $\sum_1^n \Delta_k/n^{1/2}$ has a limiting $N(0, \sigma^2(f))$ distribution. In view of (7.3), this is equivalent to the assertion of the theorem.

□

Now assume that E is finite and define the *fundamental matrix* $\boldsymbol{Z}$ by

$$\boldsymbol{Z} \;=\; (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{1\pi})^{-1} \;=\; \sum_{n=0}^{\infty} (\boldsymbol{P} - \boldsymbol{1\pi})^n \;=\; \boldsymbol{I} + \sum_{n=1}^{\infty} (\boldsymbol{P}^n - \boldsymbol{1\pi}). \quad (7.4)$$

Note that by Proposition 6.2 we have $|\lambda| < 1$ for any eigenvalue of $\boldsymbol{P} - \boldsymbol{1\pi}$ when $\boldsymbol{P}$ is aperiodic, so that the first series converges and equals the inverse; the last expression for $\boldsymbol{Z}$ follows by verifying by induction that $(\boldsymbol{P} - \boldsymbol{1\pi})^n = \boldsymbol{P}^n - \boldsymbol{1\pi}$ (we omit the easy proof that (7.4) also holds in the periodic case). Some easily verified identities are

$$\boldsymbol{\pi Z} = \boldsymbol{\pi}, \quad \boldsymbol{Z1} = \boldsymbol{1}, \quad \boldsymbol{PZ} = \boldsymbol{ZP} = \boldsymbol{Z} - \boldsymbol{I} + \boldsymbol{1\pi}. \quad (7.5)$$

Proposition 7.3 *Assume that E is finite. Then if $\boldsymbol{\pi f} = \boldsymbol{0}$, the unique solution $\boldsymbol{g}$ of Poisson's equation satisying $\boldsymbol{\pi g} = 0$ is $\boldsymbol{g} = \boldsymbol{Zf}$.*

Proof. From (7.5), we first get $\boldsymbol{\pi g} = \boldsymbol{\pi f} = 0$ and next

$$\boldsymbol{Pg} \;=\; (\boldsymbol{Z} - \boldsymbol{I} + \boldsymbol{1\pi})\boldsymbol{f} \;=\; \boldsymbol{g} - \boldsymbol{f} + \boldsymbol{0}.$$

□

Proposition 7.4 $z_{ij} \;=\; \begin{cases} \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j) & i = j \\ \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j) - \pi_j \mathbb{E}_i \tau(j) & i \neq j \end{cases}.$

Note in particular that whereas the calculation of $\mathbb{E}_i \tau(i) = 1/\pi_i$ is easy, so is not the case for $\mathbb{E}_i \tau(j)$, and the answer $(z_{jj} - z_{ij})/\pi_j$ is provided by Proposition 7.4.

Proof. Define $\boldsymbol{f} = \boldsymbol{1}_j - \pi_j \boldsymbol{1}$. Then $\pi(f) = 0$, and so by Proposition 7.1(iii) the solution g of Poisson's equation with $g(j) = 0$ is given by

$$g(i) \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(j)-1} I(X_n = j) - \pi_j \mathbb{E}_i \tau(j) \;=\; \delta_{ij} - \pi_j \mathbb{E}_i \tau(j).$$

Thus the solution g^* satifying $\pi(g^*) = 0$ is

$$g^*(i) \;=\; g(i) - \pi(g) \;=\; \delta_{ij} - \pi_j \mathbb{E}_i \tau(j) - \pi_j + \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j).$$

On the other hand, by Proposition 7.3 we have

$$g^*(i) \;=\; \boldsymbol{1}_i' \boldsymbol{Zf} \;=\; z_{ij} - \pi_j.$$

Equating these two expressions yields the result (if $i = j$, note that $\pi_j \mathbb{E}_i \tau(j) = \delta_{ij} = 1$).

□

Corollary 7.5 *In the finite case, $\sigma^2(\boldsymbol{f}) = \boldsymbol{\pi}(2\boldsymbol{f} \bullet \boldsymbol{Zf} - \boldsymbol{f} \bullet \boldsymbol{f}) - \overline{f}^2$ where $\overline{f} = \boldsymbol{\pi f}$ and $\bullet$ denotes multiplication element by element, $(\boldsymbol{a} \bullet \boldsymbol{b})_i = a_i b_i$.*

Proof. We have $\pi(\boldsymbol{f}) = \overline{f}$, $\boldsymbol{g} = \boldsymbol{Z}(\boldsymbol{f} - \overline{f}\mathbf{1})$ in Theorem 7.2 and thus

$$\begin{aligned}
\sigma^2(\boldsymbol{f}) &= \boldsymbol{\pi}(\boldsymbol{g} \bullet \boldsymbol{g} - \boldsymbol{P}\boldsymbol{g} \bullet \boldsymbol{P}\boldsymbol{g}) \\
&= \boldsymbol{\pi}\big((\boldsymbol{Z}\boldsymbol{f} - \overline{f}\mathbf{1}) \bullet (\boldsymbol{Z}\boldsymbol{f} - \overline{f}\mathbf{1}) - (\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f}) \bullet (\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f})\big) \\
&= \boldsymbol{\pi}\big(\overline{f}^2\mathbf{1} - 2\overline{f}\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f} \bullet \boldsymbol{f} + 2\boldsymbol{f} \bullet \boldsymbol{Z}\boldsymbol{f}\big) \\
&= \overline{f}^2 - 2\overline{f}^2 + \boldsymbol{\pi}(-\boldsymbol{f} \bullet \boldsymbol{f} + 2\boldsymbol{f} \bullet \boldsymbol{Z}\boldsymbol{f}),
\end{aligned}$$

where we used (7.5) repeatedly in the second step. □

8 Foundations of the General Theory of Markov Processes

We shall consider two generalizations, first that of a general (not necessarily countable) state space E, and next that of a continuous time parameter $t \in [0, \infty)$.

In the general state space case, one needs to assume that E is equipped with a measurable structure, i.e. a σ–algebra $\mathscr{E}$ to which all subsets of E considered in the following are assumed to belong. Instead of the transition matrix we have a *transition* (or *Markov*) *kernel*, i.e. a function $P(x, A)$ of $x \in E$ and $A \in \mathscr{E}$ such that $P(x, \cdot)$ is a probability on $(E, \mathscr{E})$ for each x and $P(\cdot, A)$ is $\mathscr{E}$–measurable for each A.

Markov chains with transition kernel P and the corresponding Markov probabilities $\mathbb{P}_\mu$ (with μ a distribution on $(E, \mathscr{E})$) are defined by the requirements $\mathbb{P}_\mu(X_0 \in A) = \mu(A)$,

$$\mathbb{P}_\mu(X_{n+1} \in A \,|\, \mathscr{F}_n) \;=\; P(X_n, A) \tag{8.1}$$

where $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$. With the usual a.s. interpretation of conditional probabilities, it follows from (8.1) that

$$\mathbb{P}_\mu(X_{n+1} \in A \,|\, X_n = x) \;=\; P(x, A) \tag{8.2}$$

Also, say by induction, one easily gets

$$\begin{aligned}
&\mathbb{P}_\mu(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \\
&\quad = \int_{A_0} \mu(\mathrm{d}x_0) \int_{A_1} P(x_0, \mathrm{d}x_1) \cdots \int_{A_{n-1}} P(x_{n-2}, \mathrm{d}x_{n-1}) P(x_{n-1}, A_n).
\end{aligned} \tag{8.3}$$

This formula also immediately suggests how to define the Markov probabilities and the Markov chain: take $X_0, X_1, \ldots$ as the projections $E^{\mathbb{N}} \to E$ and let

$$\mathscr{E}_n = \sigma(X_0, \ldots, X_n), \quad \mathscr{E}_\infty = \sigma(X_0, X_1, \ldots) = \mathscr{E}^{\mathbb{N}}.$$

Then by standard arguments from measure theory it can be seen that the r.h.s. of (8.3) in a unique way corresponds to a probability $\mathbb{P}_\mu^n$ on

$(E^{\mathbb{N}},\mathscr{E}_n)$. The $\mathbb{P}_\mu^n$ have the consistency property $\mathbb{P}_\mu^n(A)=\mathbb{P}_\mu^m(A)$, $m\le n$, $A\in\mathscr{E}_m$, and hence define a finitely additive probability on the algebra $\cup_0^\infty\mathscr{E}_n$. The desired $\mathbb{P}_\mu$ is the (necessarily unique) extension to $\mathscr{E}_\infty = \sigma\left(\cup_0^\infty\mathscr{E}_n\right)$. The existence, i.e. the σ–additivity on $\cup_0^\infty\mathscr{E}_n$, may be seen either from Kolmogorov's consistency theorem which requires some topological assumptions like E being Polish and $\mathscr{E}$ the Borel σ–algebra, or by a measure–theoretic result of Ionescu Tulcea (see Neveu, 1965).

The continuous–time case is substantially more involved. What will be needed in later chapters is, however, only a few basic facts and we shall therefore just outline a theory which needs several amendments when pursuing Markov process theory in its full generality.

One does not get very far without topology, so we assume right from the start that E is Polish with $\mathscr{E}$ the Borel σ–algebra. That a process $\{X_t\}_{t\ge 0}$ with state space E is Markov means intuitively just the same as in discrete time: given the history $\mathscr{F}_t=\sigma(X_s;\,s\le t)$, the process evolves from then on as restarted at time 0 in state X_t and depending on $\mathscr{F}_t$ through X_t only. Formally, this may be expressed by the existence of a family of probability measures $\mathbb{P}_\mu$ with the property $\mathbb{P}_\mu(X_0\in A)=\mu(A)$,

$$\mathbb{E}_\mu\bigl[h(X_{s+t};\,t\ge 0)\,\big|\,\mathscr{F}_s\bigr] \;=\; \mathbb{E}_{X_s}h(X_t;\,t\ge 0) \tag{8.4}$$

where $\mathbb{P}_x,\mathbb{E}_x$ refer to $X_0=x$ and (8.4) should hold for a class of functions h of the process sufficiently rich to determine the distribution of $\{X_t\}_{t\ge 0}$. For example, it would suffice to consider the class $\mathscr{H}$ of all h of the form

$$h(x_t;\,t\ge 0) \;=\; \prod_{i=0}^{n} I(x_{t_i}\in A_i). \tag{8.5}$$

If $\{X_t\}_{t\ge 0}$ has paths say in $D=D([0,\infty),E)$, then (8.4) for all $h\in\mathscr{H}$ will be equivalent to (8.4) to hold for all bounded measurable $h:D\to\mathbb{R}$. In fact, an easy induction argument shows that it is even sufficient to let $n=0$ in (8.5), and the Markov property in this equivalent formulation then becomes

$$\mathbb{P}(X_{s+t}\in A\,|\,\mathscr{F}_s) \;=\; P^t(X_s,A) \quad\text{where}\quad P^t(x,A)=\mathbb{P}_x(X_t\in A). \tag{8.6}$$

Given a Markov process, it is clear that $P^t(x,A)$ as defined by (8.6) is a transition kernel. Using the Markov property we get

$$P^{t+s}(x,A) \;=\; \mathbb{E}_x\mathbb{P}(X_{s+t}\in A\,|\,\mathscr{F}_s) \;=\; \mathbb{E}P^t(X_s,A) \;=\; \int P^t(y,A)P^s(x,\mathrm{d}y),$$

which in operator notation is written $P^{t+s}=P^tP^s$ and referred to as the *Chapman–Kolmogorov equations* (or the *semi–group property*). Conversely, given a family $\{P^t\}_{t\ge 0}$ satisfying the Chapman–Kolmogorov equations, it is possible to construct a corresponding Markov process. To this end, we proceed as in discrete time: let $X_t:E^{[0,\infty)}\to E$ be the projection and define for $0=t_0<t_1<\cdots<t_n$ a probability on the sub–σ–algebra

$\sigma(X_{t_i};\, i=0,\ldots,n)$ by

$$\begin{aligned}&\mathbb{P}_\mu(X_{t_0}\in A_0, X_{t_1}\in A_1,\ldots,X_{t_n}\in A_n)\\ &\quad= \int_{A_0}\mu(\mathrm{d}x_0)\int_{A_1}P^{t_1-t_0}(x_0,\mathrm{d}x_1)\\ &\qquad\cdots\int_{A_{n-1}}P^{t_{n-1}-t_{n-2}}(x_{n-2},\mathrm{d}x_{n-1})P^{t_n-t_{n-1}}(x_{n-1},A_n).\end{aligned}\tag{8.7}$$

That this defines a semigroup is readily apparent from the Chapman–Kolmogorov equations. Since E is Polish, there thus exists a unique extension to $\mathscr{E}^{[0,\infty)}$, and the Markov property (8.4) with $h\in\mathscr{H}$ is inherent in the definition (8.7).

There are, however, severe difficulties associated with this approach. First, the intuitive description of a particular model is seldom in terms of the P^t. Next, the construction makes $\mathscr{E}^{[0,\infty)}$ the collection of measurable sets, i.e. when $A\notin\mathscr{E}^{[0,\infty)}$ one cannot make sense of $\mathbb{P}_\mu(A)$. But $\mathscr{E}^{[0,\infty)}$ is not very rich since one can easily see that $A\in\mathscr{E}^{[0,\infty)}$ implies that A depends on the X_t for t in a countable collection $T_A\subset[0,\infty)$ of time points. Thus for example sets like

$$\bigl\{\omega:\ X_t(\omega)\text{ is a continuous function of }t\bigr\}$$

is not in $\mathscr{E}^{[0,\infty)}$, and (when say $E=\mathbb{R}$) similarly $\max_{0\le t\le T}X_t$ and $\inf\{t:\ X_t=0\}$ are not measurable. Hence it is necessary to construct versions of the process with sample paths say in D. This requires further properties of the P^t, typically continuity requirements. We shall not go into this since the explicit examples that we shall encounter will almost a priori satisfy such path regularity properties. For example, queues are constructed by simple transformations of sequences of service times and interarrival times, and not starting from semi–groups, consistent families and so on.

Now let σ be a stopping time w.r.t. $\{\mathscr{F}_t\}_{t\ge0}$ amd let $\mathscr{F}_\sigma$ be the stopping time σ–algebra, cf. A10. We say that $\{X_t\}_{t\ge0}$ has the *strong Markov property* w.r.t. σ if a.s. on $\{\sigma<\infty\}$

$$\mathbb{P}\bigl(X_{\sigma+t}\in A\,\big|\,\mathscr{F}_\sigma\bigr)\ =\ P^t(X_\sigma,A);\tag{8.8}$$

again, this implies a functional form

$$\mathbb{E}_\mu\bigl[h(X_{\sigma+t};\,t\ge0)\,\big|\,\mathscr{F}_\sigma\bigr]\ =\ \mathbb{E}_{X_\sigma}h(X_t;\,t\ge0).$$

The process is *strong Markov* if it has the strong Markov property w.r.t. any stopping time σ.

Proposition 8.1 *A Markov process $\{X_t\}_{t\ge0}$ has the strong Markov property w.r.t. any stopping time σ which assumes only a countable number of values, $\sigma\in\{\infty,s_1,s_2,\ldots\}$.*

Proof. We must show that for $A \in \mathscr{E}$, $F \in \mathscr{F}_\sigma$

$$\mathbb{P}_\mu\bigl(X_{\sigma+t} \in A;\, F,\, \sigma < \infty\bigr) \;=\; \mathbb{E}_\mu\bigl[P^t(X_\sigma, A);\, F,\, \sigma < \infty\bigr].$$

However, if $F \subseteq \{\sigma = s_k\}$ this is immediate from the Markov property (8.4). In the general case, decompose $F \cap \{\sigma < \infty\}$ as the disjoint union of the sets $F \cap \{\sigma = s_k\}$ and sum over k. □

As an immediate consequence, we have:

Corollary 8.2 *Any discrete time Markov chain* (with discrete or general state space) *has the strong Markov property.*

Also in continuous time, Proposition 8.1 is greatly helpful in establishing the strong Markov property. A typical example is the following:

Corollary 8.3 *Assume that $\{X_t\}_{t\geq 0}$ has right–continuous paths and that for any bounded continuous $f : E \to \mathbb{R}$ and any s it holds that $\mathbb{E}_x f(X_s)$ is a continuous function of x or, more generally, that the paths of $\mathbb{E}_{X_t} f(X_s)$ are right–continuous functions of t. Then the strong Markov property holds.*

Proof. Let σ be a given stopping time and define $\sigma(k) = n2^{-k}$ on $\bigl\{(n-1)2^{-k} < \sigma \leq n2^{-k}\bigr\}$. Then the $\sigma(k)$ are stopping times and $\sigma(k) \downarrow \sigma$ as $k \to \infty$. By Proposition 8.1 we have furthermore

$$\mathbb{E}_\mu\bigl[f(X_{\sigma(k)+s}) \,\big|\, \mathscr{F}_{\sigma(k)}\bigr] \;=\; \mathbb{E}_{X_{\sigma(k)}} f(X_s). \tag{8.9}$$

If $F \in \mathscr{F}_\sigma$, then $F \in \mathscr{F}_{\sigma(k)}$, and hence (8.9) implies

$$\mathbb{E}_\mu[f(X_{\sigma(k)+s});\, F] \;=\; \mathbb{E}_\mu[\mathbb{E}_{X_{\sigma(k)}} f(X_s);\, F].$$

A check of the assumptions show that the integrands converge pointwise. Thus by dominated convergence,

$$\mathbb{E}_\mu[f(X_{\sigma+s});\, F] \;=\; \mathbb{E}_\mu[\mathbb{E}_{X_\sigma} f(X_s);\, F].$$

The truth of this for all bounded continuous f and all $F \in \mathscr{F}_\sigma$ implies (8.8). □

We next consider the *hitting time* $\tau(A)$ of a Borel subset A, $\tau(A) = \inf\{t > 0 :\, X_t \in A\}$. That $\tau(A)$ is a stopping time is a triviality in discrete time since then obviously

$$\{\tau(A) \leq n\} \;=\; \bigcup_{k=1}^{n} \{X_k \in A\}\,.$$

However, in continuous time some (perhaps unexpected) difficulties arise even for elementary sets like closed and open ones, and this is in fact one of the reasons that one needs to amend and extend the theory that has been discussed so far and which may still appear reasonably simple and intuitive. We discuss these points briefly below, but first state and prove a more elementary result that is sufficient to deal with virtually all the processes to be met and all the questions to be asked in this book.

Proposition 8.4 *Suppose the paths of* $\{X_t\}$ *are piecewise continuous with right limits. Then:*
(a) *the jump times* $0 < \iota(1) < \iota(2) < \cdots$ *are stopping times w.r.t.* $\{\mathscr{F}_t\}$*;*
(b) *if* A *is closed, then* $\tau(A)$ *is a stopping time w.r.t.* $\bigl\{\mathscr{F}_t\bigr\}$*.*

Proof. Let $\mathbb{Q}(t)$ be the set of numbers of the form qt with q rational and $0 \le q \le 1$, and let d be some metric on E. Then the sets $G_1 = \{\iota(1) \le t\}$ and

$$G_2 \;=\; \bigcup_{m=1}^{\infty}\bigcap_{n=1}^{\infty}\bigcup_{u,s\in\mathbb{Q}(t)}\bigl\{|u-s| \le 1/n,\, d(X_u,X_s) > 1/m\bigr\}$$

coincide. In fact, on G_1 we have for some m a jump of size at least m^{-1}, and this easily gives $G_1 \subseteq G_2$. Conversely, the uniform continuity of $\{X_s\}_{s\le t}$ on G_1^c easily shows $G_1^c \subseteq G_2^c$. Since $d(X_u,X_s)$ is $\mathscr{F}_t$–measurable for $u,s \le t$, we have $G_1 = G_2 \in \mathscr{F}_t$, and thus $\iota(1)$ is a stopping time. For $\iota(2)$, just add the requirement $u,s \ge \iota(1)$ in the definition of G_2, and so on.

To prove (b), define $m(S) = \inf\{d(X_u,A);\ u \in S\}$, $S \subseteq [0,\infty)$. If A is closed, we have $X_{\tau(A)} \in A$ by right–continuity, and hence in the special case of continuous paths

$$\{\tau(A) \le t\} \;=\; \bigl\{m\bigl([0,t]\bigr) = 0\bigr\} \;=\; \bigl\{m\bigl(\mathbb{Q}(t)\bigr) = 0\bigr\} \in \mathscr{F}_t. \tag{8.10}$$

But if $I_{k,n} \;=\; \bigl\{u : \iota(k) - 1/n \le u < \iota(k) \le t\bigr\}$, then

$$\{u \in I_{k,n}\} \;=\; \bigl\{u < \iota(k) \le t \wedge (u + 1/n)\bigr\} \;\in\; \mathscr{F}_t.$$

Thus as in (8.10)

$$\begin{aligned}\{\tau(A) \le t\} \;&=\; \lim_{n\to\infty}\Bigl\{\tau(A) \in [0,t]\Bigl\backslash\bigcup_{k=1}^{\infty} I_{k,n}\Bigr\}\\ &=\; \lim_{n\to\infty}\Bigl\{m\Bigl(\mathbb{Q}(t)\Bigl\backslash\bigcup_{k=1}^{\infty} I_{k,n}\Bigr) = 0\Bigr\} \;\in\; \mathscr{F}_t.\end{aligned}$$

□

We conclude with a brief discussion of some more difficult topics which, however, are not essential for the rest of the book. Define $\mathscr{F}_{t+} = \cap_{s>t}\mathscr{F}_s$ and let $\mathscr{G}^{(\mu)}$ denote the $\mathbb{P}_\mu$–completion of $\mathscr{G}$ (some arbitrary σ–field), i.e. the smallest σ–field containing $\mathscr{G}$ and all $\mathbb{P}_\mu$–null sets. Then:

Proposition 8.5 *Suppose that* $\{X_t\}$ *has right–continuous paths. Then:*
(a) *If* A *is open, then* $\tau(A)$ *is a stopping time w.r.t.* $\bigl\{\mathscr{F}_{t+}\bigr\}$*;*
(b) *For any Borel set* A*,* $\tau(A)$ *is a stopping time w.r.t.* $\bigl\{\mathscr{F}_{t+}^{(\mu)}\bigr\}$*.*

Proof of (a). If A is open and $X_u \in A$, then $X_{u+v} \in A$ for all small positive v. Hence the event $\{\tau(A) \le t\}$ may be written as

$$\bigcap_{n=1}^{\infty}\bigcup_{s\le t+1/n}\{X_s \in A\} \;=\; \bigcap_{n=1}^{\infty}\bigcup_{s\in\mathbb{Q}(t+1/n)}\{X_s \in A\}\,,$$

and here the event on the r.h.s. is clearly in $\mathscr{F}_{t+}$.

The proof of (b) is far beyond the present scope (and need!), and we refer, e.g., to Dellacherie and Meyer (1975–93). □

One now defines a *history* of the process as an increasing family $\{\mathscr{G}_t\}_{t\geq 0}$ of σ–fields (a filtration) with $\mathscr{F}_t \subseteq \mathscr{G}_t$ (or equivalenty X_t $\mathscr{G}_t$–measurable) for all t, and say that $\{X_t\}$ is Markov with transition semigroup $\{P^t\}$ w.r.t. $\{\mathscr{G}_t\}$ and some fixed governing probability measure if

$$\mathbb{P}(X_{t+s} \in A \,|\, \mathscr{G}_s) \;=\; P^t(X_s, A). \tag{8.11}$$

Apart from $\{\mathscr{F}_t\}$, some main candidates for the history are $\{\mathscr{F}_{t+}\}$ and $\{\mathscr{F}_{t+}^{(\mu)}\}$. It follows immediately from the chain rule for conditional expectations that if $\{X_t\}$ is Markov w.r.t. some history, then $\{X_t\}$ is Markov w.r.t. $\{\mathscr{F}_t\}$ as well. Conversely:

Proposition 8.6 *Let $\{X_t\}$ be Markov w.r.t. $\{\mathscr{F}_t\}$ and satisfy the regularity conditions of Corollary 8.3. Then:*
(a) *for each μ and each bounded measurable h, we have $\mathbb{P}_\mu$–a.s. that*

$$\begin{aligned}\mathbb{E}_\mu\big[h(X_{s+t}; t\geq 0)\,\big|\,\mathscr{F}_s\big] \;&=\; \mathbb{E}_\mu\big[h(X_{s+t}; t\geq 0)\,\big|\,\mathscr{F}_{s+}\big]\\ &=\; \mathbb{E}_\mu\big[h(X_{s+t}; t\geq 0)\,\big|\,\mathscr{F}_{s+}^{(\mu)}\big];\end{aligned}$$

(b) (BLUMENTHAL'S 0–1 LAW) *if $A \in \mathscr{F}_{0+}$, then for a fixed $x \in E$ either $\mathbb{P}_x(A) = 0$ or $\mathbb{P}_x(A) = 1$.*
(c) *$\{X_t\}$ is Markov w.r.t. $\{\mathscr{F}_{t+}\}$ and $\{\mathscr{F}_{t+}^{(\mu)}\}$ as well.*

Proof. (a) The second identity is just a general property of the completion operator. For the first, arguments similar to those used many times above show that it suffices to take h of the form $h(X_t)$ with $t > 0$ and h continuous and bounded. Since then $h(X_{s+t+1/n}) \overset{\text{a.s.}}{\to} h(X_{s+t})$, it follows from a continuity result for conditional expectations (Chung, 1974, p. 340) that indeed

$$\begin{aligned}\mathbb{E}_\mu\big[h(X_{s+t})\,\big|\,\mathscr{F}_{s+}\big] \;&=\; \lim_{n\to\infty} \mathbb{E}_\mu\big[h(X_{s+t+1/n})\,\big|\,\mathscr{F}_{s+1/n}\big]\\ =\; \lim_{n\to\infty} \mathbb{E}_{X_{s+1/n}} h(X_t) \;&=\; \mathbb{E}_{X_s} h(X_t) \;=\; \mathbb{E}_\mu\big[h(X_{s+t})\,\big|\,\mathscr{F}_s\big],\end{aligned}$$

and the proof of (a) is complete. For (b), let $t = 0$ and $h = I(A)$ in (a) to obtain $\mathbb{P}_x(A|\mathscr{F}_{0+}) = \mathbb{P}_x(A|\mathscr{F}_0)$ a.s. Here the l.h.s. is just $I(A)$ and since $\mathscr{F}_0$ is $\mathbb{P}_x$–trivial, the r.h.s. is constant. Hence $I(A)$ is constant a.s. which is only possible if the probability is either 0 or 1. Finally (c) is an immediate consequence of (a). □

We stop the discussion of the foundations of the general theory of Markov processes at this point. As for the topics discussed in Sections 2–4, classification of states and limit theory will be discussed in Chapter II for a discrete state space and continuous time process. The case of a general E is much more complicated even in discrete time. For example, it is not clear what recurrence should mean since even in simple–minded continuous state space models, $\mathbb{P}_x(\tau(x) < \infty)$ will most often be 0. Some results (more or less the best known ones) are given in VII.3 and can, somewhat surprisingly, be derived as simple consequences of the ergodic theorm for discrete Markov chains. In continuous time, the existing theory is hardly equally satisfying, but a number of special cases will be encountered. For example, the main problem within the whole area of renewal theory (Chapter V) will be seen to be equivalent to the ergodicity question for the continuous–time and –state version of the recurrence time chains in Section 2.

Notes General Markov chains in discrete time are discussed, e.g., in Neveu (1965), Meyn and Tweedie (1993) and Revuz (1984). For up–to–date and readable accounts of the continuous–time case, see Rogers and Williams (1994) or Revuz and Yor (1999).

A topic not treated above but used at a few places in the book is the *generator* $\mathscr{A}$ of a continuous-time Markov process, a certain operator on a subspace $\mathscr{D}_{\mathscr{A}}$ of functions on E. There are many variants of the definition around, but the intutition behind them all is that one should have

$$\mathbb{E}_x f(X_h) \;=\; f(x) + \mathscr{A}f(x)h + \mathrm{o}(h), \quad f \in \mathscr{D}_{\mathscr{A}}. \tag{8.12}$$

The domain $\mathscr{D}_{\mathscr{A}}$ is specified by additional requirements in (8.12), one classical variant (see e.g. Karlin and Taylor, 1981) being that f should be bounded and the convergence in (8.12) uniform. Note that the identification of $\mathscr{D}_{\mathscr{A}}$ in this set–up is tedious even in such a basic case as standard Brownian motion where $\mathscr{A}$ is a restriction of the differential operator $f \to f''/2$. Note also that $\mathscr{D}_{\mathscr{A}}$ actually may contain crucial information on the process. For example, for reflecting Brownian motion with reflection at 0 or absorbtion at 0, $\mathscr{A}f = f''/2$ in both cases, but $f \in \mathscr{D}_{\mathscr{A}}$ requires $f'(0) = 0$ in the reflected case and $f(0) = 0$ in the absorbing case.

Typically, $f(X_t) - \int_0^t \mathscr{A}f(X_s)\,\mathrm{d}s$ is a martingale (the *Dynkin martingale*) for $f \in \mathscr{D}_{\mathscr{A}}$, and a modern variant of the definition is that $f \in \mathscr{D}_{\mathscr{A}}$, $g = \mathscr{A}f$ means that $f(X_t) - \int_0^t g(X_s)\,\mathrm{d}s$ is a local martingale.

The most basic case is a Markov jump process as in Chapter II, where in the finite case it holds for any of the possible definitions that $\mathscr{D}_{\mathscr{A}}$ is the set of all functions on E and $\mathscr{A}$ is the operator $\boldsymbol{f} \to \boldsymbol{\Lambda f}$ where $\boldsymbol{\Lambda}$ is the intensity matrix.

II

Markov Jump Processes

1 Basic Structure

Let E be a discrete (finite or countable) state space and $\{X_t\}_{t\geq 0}$ a Markov process with state space E as defined in I.8, with transition semigroup $\{P\}_{t\geq 0}$. We write $p_{ij}^t = P^t(i,\{j\}) = \mathbb{P}_i(X_t = j)$, and we may identify the transition semigroup by the family $\left\{\boldsymbol{P}^t\right\}_{t\geq 0} = \left\{(p_{ij}^t)\right\}_{t\geq 0}$ of transition matrices. The Chapman–Kolmogorov equations $\boldsymbol{P}^{t+s} = \boldsymbol{P}^t\boldsymbol{P}^s$ may then be interpreted in the sense of usual matrix multiplication.

Problems arising when pursuing the theory without further regularity conditions have already been discussed in I.8. As a further unpleasant possibility, we mention here that some (or even all) states i may be *instantaneous*, i.e. the process jumps out of i immediately after i has been entered. We shall avoid these problems by imposing upon the process a further regularity property, which is inherent in the intuitive picture of any of the models we are concerned with, and which turns out to be sufficient for developing the theory quite smoothly.

The feature that we concentrate on is that of a *pure jump* structure illustrated in Fig. 1.1: the amount of time spent in each state is positive so that the sample paths are piecewise constant. For a pure jump process, we denote the times of jumps by $S_0 = 0 < S_1 < S_2 \cdots$, the sojourn times (or holding times) by $T_n = S_{n+1} - S_n$ and the sequence of states visited by $Y_0, Y_1, \ldots$. Thus the sample paths are constant between consecutive S_n and we define the value at S_n by right–continuity, i.e. $X_{S_n} = Y_n$. Two possible phenomena require some further comment. The process may be absorbed, say at i. In that case there is a last finite S_n (the absorbtion time) and we use the convention $T_n = T_{n+1} = \cdots = \infty$, $Y_n = Y_{n+1} = \cdots = i$. This still yields a very simple structure of paths. More troublesome from that point of view is the possibility of the jumps to accumulate, i.e. of the *explosion time* $\omega(\Delta) = \sup_n S_n$ to be finite (in that case the Y_n and T_n determine the process only up to $\omega(\Delta)$. This seems contrary to intuition in most cases, but is perfectly feasible from the point of view of general theory. We discuss the point in more detail later in Sections 2 and 3, and proceed here to discuss some fundamental properties of a Markov jump process.

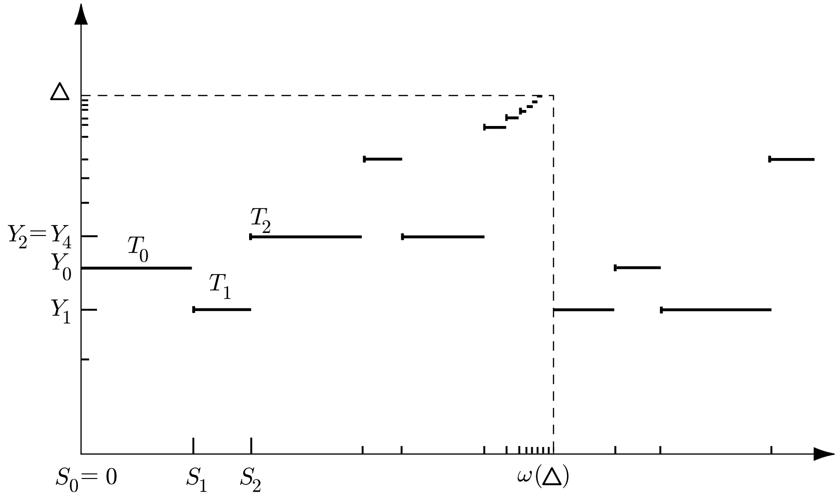

Figure 1.1

Sample path of a pure jump process. The scale of the state space is chosen to illustrate the possibility of explosion within finite time.

Theorem 1.1 *Any Markov jump process has the strong Markov property.*

Proof. This is a trivial consequence of the first part of I.8.3 since when E is discrete, then *any* function g on E (in particular $g(x) = \mathbb{E}_x f(X_s)$) is continuous. □

The next result describes the basic structure of a Markov jump process up to the time of explosion. Consider the exponential distribution with density $\lambda e^{-\lambda x}$, $x > 0$, and denote by the *intensity* (or sometimes *rate*) the parameter λ (by the exponential distribution with intensity $\lambda = 0$ we understand the distribution degenerate at ∞).

Theorem 1.2 *Consider a Markov jump process. Then the joint distribution of the sequences $\{Y_n\}$ of states visited* (before explosion) *and $\{T_n\}$ of holding times is given by:* (i) *$\{Y_n\}$ is a Markov chain;* (ii) *there exist $\lambda(i) \geq 0$ such that, given $\{Y_n\}$, the T_ℓ are independent, with T_k being exponentially distributed with intensity $\lambda(Y_k)$.*

Proof. The joint distribution of the Y_n, T_n is completely specified by probabilities of the form $\mathbb{P}_i F_n$ where $F_n = \{Y_k = i(k), T_{k-1} > t(k),\ k = 1, \dots, n\}$. Letting $i(0) = i$, the assertion of the theorem is equivalent to

$$\mathbb{P}_i F_n \;=\; \prod_{k=1}^{n} q_{i(k-1)i(k)} \exp\bigl\{-\lambda(i(k-1))t(k)\bigr\} \tag{1.1}$$

for some transition matrix $\boldsymbol{Q}$ and suitable intensities $\lambda(i)$. It is clear that the only possible candidate for $\boldsymbol{Q}$ is $q_{ij} = \mathbb{P}_i(Y_1 = j)$. To determine $\lambda(i)$, we let $z(t) = \mathbb{P}_i(T_0 > t)$. Since $X_t = i$ on $\{T_0 > t\}$, the Markov property yields

$$z(t+s) \;=\; \mathbb{E}_i\mathbb{P}_i\bigl(T_0 > t+s \,\big|\, \mathscr{F}_t\bigr) \;=\; \mathbb{E}_i\bigl[\mathbb{P}_i(T_0 > s);\, T_0 > t\bigr] \;=\; z(s)z(t).$$

Since z is nonincreasing with $z(0) = 1$, elementary facts on functional equations yield $z(t) = \mathrm{e}^{-\lambda(i)t}$ for some $\lambda(i) \geq 0$ (the pure jump property implies $z(t) \uparrow 1$ as $t \downarrow 0$ so that $z(t) = I(t = 0)$, i.e. $\lambda(i) = \infty$, is excluded).

Applying the Markov property once more, we get similarly

$$\begin{aligned}\mathbb{P}_i(Y_1 = j, T_0 > t) \;&=\; \mathbb{E}_i\mathbb{P}_i\bigl(Y_1 = j, T_0 > t \,\big|\, \mathscr{F}_t\bigr)\\ &=\; \mathbb{E}_i\bigl[\mathbb{P}_i(Y_1 = j);\, T_0 > t\bigr] \;=\; q_{ij}\mathrm{e}^{-\lambda(i)t}\end{aligned}$$

which is (1.1) for $n = 1$. The case $n > 1$ now follows easily by the strong Markov property and induction. Indeed, evaluating $\mathbb{P}_i F_n$ upon conditioning upon $\mathscr{F}_{S_{n-1}}$ we obtain from $X_{S_{n-1}} = Y_{n-1}$ that

$$\begin{aligned}\mathbb{P}_i F_n \;&=\; \mathbb{E}_i\Bigl[\mathbb{P}_{X_{S_{n-1}}}\bigl(Y_1 = i(n), T_0 > t(n)\bigr);\, F_{n-1}\Bigr]\\ &=\; \mathbb{P}_{i(n-1)}\bigl(Y_1 = i(n), T_0 > t(n)\bigr)\mathbb{P}_i(F_{n-1})\\ &=\; q_{i(n-1)i(n)}\mathrm{e}^{-\lambda(i(n-1))t(n)}\mathbb{P}_i(F_{n-1}).\end{aligned}$$

□

Problems

1.1 Show that the explosion time is a stopping time w.r.t. $\{\mathscr{F}_t\}_{t\geq 0}$.

2 The Minimal Construction

The intuitive description of a practical model is usually given in terms of the intensities $\lambda(i)$ and the jump probabilities q_{ij} rather than in terms of the transition matrices $\boldsymbol{P}^t$ which are difficult to evaluate even in extremely simple cases. The question therefore arises whether *any* set of $\lambda(i)$, q_{ij} leads to a Markov jump process. The construction (given below) is immediately suggested by Theorem 1.2 and the problem becomes to check whether indeed a Markov process comes out. As will be seen, the answer is affirmative.

We assume therefore that we are given a set $\lambda(i) \geq 0$, $i \in E$, and a transition matrix $\boldsymbol{Q}$ on E (the $\boldsymbol{Q}$ of Theorem 1.2 has the property $q_{ii} = 0$ if and only if $\lambda(i) > 0$, but this need not be assumed here). Let $\Delta \notin E$ be some extra state (needed to describe the process after a possible explosion), write $E_\Delta = E \cup \{\Delta\}$ and define $\lambda(\Delta) = 0$, $q_{\Delta\Delta} = 1$. We consider the sample space

$$\Omega \;=\; (0,\infty]^{\mathbb{N}} \times E_\Delta^{\mathbb{N}} \;=\; \{(t_0, t_1, \ldots, y_0, y_1, \ldots) : 0 < t_k \leq \infty, y_k \in E_\Delta\}$$

and let $T_0, T_1, \ldots, Y_0, Y_1, \ldots$ be the obvious coordinate functions (projections) on Ω. It is then a matter of routine to construct probabilities $\mathbb{P}_i$, $i \in E$, on Ω with the following properties:

(i) $\{Y_n\}$ is a Markov chain with transition matrix $\boldsymbol{Q}$ and $\mathbb{P}_i(Y_0 = i) = 1$;

(ii) given $\{Y_n\}$, the T_ℓ are independent, with T_k being exponentially distributed with intensity $\lambda(i)$ on $\{Y_k = i\}$.

We construct $\{X_t\}_{t \geq 0}$ up to the time of explosion simply by reversing the construction of the Y_k, T_k illustrated in Fig. 1.1 (and if needed letting X_t remain in Δ after explosion). That is, we let $S_0 = 0$,

$$\begin{aligned} S_n &= T_0 + \cdots + T_{n-1}, \quad \omega(\Delta) \;=\; \sup S_n \;=\; T_0 + T_1 + \cdots, \\ X_t &= \begin{cases} Y_k & \text{if } S_k \leq t < S_{k+1} \\ \Delta & \text{if } t \geq \omega(\Delta) \end{cases} . \end{aligned}$$

We shall prove the following main result:

Theorem 2.1 *$\{X_t\}_{t \geq 0}$ is a Markov jump process on E_Δ.*

In the proof, we need to study the residual sojourn time (overshoot) R_t, at time t, i.e. $R_t = S_{n(t)} - t$, where $n(t) = \min\{n : S_n > t\}$.

Lemma 2.2 *Given $\mathscr{F}_t = \sigma(X_s;\, s \leq t)$, the conditional distribution of R_t is exponential with intensity $\lambda(X_t)$.*

Proof. The intuitive argument is just that given $\mathscr{F}_t$, the distribution of $T_{n(t)-1}$ is that of T given $T > u$, where $u = t - S_{n(t)-1}$ and T is exponential with intensity $\lambda(Y_{n(t)-1}) = \lambda(X_t)$. To spell out a formal proof we must show that

$$\mathbb{P}_i(R_t > r, A) \;=\; \mathbb{E}_i\bigl[\exp\{-\lambda(X_t) r\}\,;\, A\bigr] \tag{2.1}$$

for all $r < \infty$ and all $A \in \mathscr{F}_t$. If $A \subseteq \{\omega(\Delta) \leq t\}$, then both sides are just $\mathbb{P}_i A$ so we may assume $A \subseteq \{\omega(\Delta) > t\}$ and it then suffices to consider A of the form $\{n(t) - 1 = n,\, F_n\} = \{F_n, S_n \leq t, S_n + T_n > t\}$ where F_n is as in the proof of Theorem 1.2. Thus if we condition upon the Y_k, T_{k-1}, $k = 1, \ldots, n$ and use the formula $\mathbb{P}(T > t + r) = \mathrm{e}^{-\lambda r}\mathbb{P}(T > t)$ for the exponential distribution, we may evaluate the l.h.s. of (2.1) as

$$\mathbb{P}_i\bigl(F_n, S_n \leq t, S_n + T_n > t + r\bigr) \;=\; \mathrm{e}^{-\lambda(i(n))r}\mathbb{P}_i\bigl(F_n, S_n \leq t, S_n + T_n > t\bigr)$$

which is the same as the r.h.s. □

Proof of Theorem 2.1. $\{X_t\}_{t\geq 0}$ is clearly pure jump, so it suffices to show that on $\{X_t = i\}$ the conditional distribution of $\{X_{t+s}\}_{s\geq 0}$ given $\mathscr{F}_t$ is just the $\mathbb{P}_i$–distribution of $\{X_s\}_{s\geq 0}$. Define

$$M_t \;=\; \bigl(T_{n(t)}, T_{n(t)+1}, \ldots, Y_{n(t)-1}, Y_{n(t)}, \ldots\bigr).$$

Then $\{X_{t+s}\}_{s\geq 0}$ is constructed from (R_t, M_t) in just the same way as $\{X_t\}_{t\geq 0}$ is constructed from

$$(R_0, M_0) \;=\; (T_0, M_0) \;=\; \bigl(T_0, T_1, \ldots, Y_0, Y_1, \ldots\bigr).$$

Hence we must show that on $\{X_t = i\}$, the conditional distribution of (R_t, M_t) given $\mathscr{F}_t$ is the $\mathbb{P}_i$–distribution of (R_0, M_0), i.e. that in the conditional distribution (i) R_t, M_t are independent, (ii) R_t has the $\mathbb{P}_i$–distribution of R_0, (iii) M_t has the $\mathbb{P}_i$–distribution of M_0. Now clearly $\{(Y_n, T_n)\}$ is a Markov chain with state space $E_\Delta \times (0, \infty]$ and transition kernel given by

$$\mathbb{P}\bigl(Y_{n+1} = j, T_{n+1} > t \,\big|\, \mathscr{H}_n\bigr) \;=\; q_{Y_n j}\mathrm{e}^{-\lambda(j)t} \tag{2.2}$$

where $\mathscr{H}_n = \sigma(Y_k, T_k \,:\, k \leq n)$. Also $n(t) - 1$ is a stopping time w.r.t. this chain and we shall evaluate the distribution of (R_t, M_t) conditionally upon $\mathscr{F}_t$ by first conditioning upon the larger σ–algebra $\mathscr{H}_{n(t)-1}$. Since $Y_{n(t)-1} = X_t$, the strong Markov property I.8.2 and (2.2) imply that given $\mathscr{H}_{n(t)-1}$, M_t has the $\mathbb{P}_{X_t}$–distribution of M_0, whereas R_t (being $\mathscr{H}_{n(t)-1}$–measurable) is degenerate. These facts and the $\mathscr{F}_t$–measurability of X_t imply (i) and (iii), whereas (ii) is the statement of Lemma 2.2. □

It should be noted, that if the process is explosive (i.e. $\mathbb{P}_i(\omega(\Delta) < \infty) > 0$ for some $i \in E$), then (see the Problems) there are in general several ways of continuing the process after $\omega(\Delta)$ which will lead to a Markov jump process (to use a common phrase, the process "runs out of instructions" at the explosion time). Among such processes, all behaving in the same way up to the explosion time, the one in Theorem 2.1 obviously minimizes $p_{ij}^t = \mathbb{P}_i(X_t = j)$ for any $i, j \in E$, and for this reason it is called the *minimal* one.

Some further discussion of the basic structure of a Markov jump process will be given in Sections 3a, 3b (though essentially this is only a reformulation of what has been shown so far), and we return here to the explosion problem. In most cases this presents an unwanted technicality, and one wants to assert as quickly as possible that a given Markov jump process is nonexplosive (e.g., in the minimal construction one can then restrict the state space to E). Necessary and sufficient conditions are given in the following Proposition 2.3 and in Proposition 3.3 of the next section, whereas Proposition 2.4 gives some sufficient conditions that are easier to work with in many cases.

Proposition 2.3 *Define* $R = \sum_0^\infty \lambda(Y_n)^{-1}$. *Then for any* $i \in E$, *the sets* $\{\omega(\Delta) < \infty\}$ *and* $\{R < \infty\}$ *coincide* $\mathbb{P}_i$*–a.s.*

Proof. Conditionally upon $\{Y_n\}$, $\omega(\Delta) = \sup S_n = \sum_0^\infty T_n$ is distributed as $\sum_0^\infty \lambda(Y_n)^{-1} V_n$, where the V_n are i.i.d. and exponential with intensity 1. The result therefore comes out by standard facts on weighted sums of i.i.d. random variables. Thus $R < \infty$ implies $\omega(\Delta) < \infty$ because of $R = \mathbb{E}(\omega(\Delta) \,|\, Y_0, Y_1, \ldots)$, and the converse may be seen, e.g., by an application of the three–series criterion. □

Proposition 2.4 *Sufficient criteria for* $\mathbb{P}_i(\omega(\Delta) < \infty) = 0$ *for all* $i \in E$ *are:* (i) $\sup_{i\in E} \lambda(i) < \infty$*;* (ii) E *is finite;* (iii) $\{Y_n\}$ *is recurrent.*

Proof. It follows from Proposition 2.3 that $\lambda(Y_n) \to \infty$ on $\{\omega(\Delta) < \infty\}$. Hence the sufficiency of (i) is clear, and (ii) is a consequence of (i). If $\{Y_n\}$ is recurrent, and $X_0 = Y_0 = i$, then $\lambda(i)$ is a limit point of $\{\lambda(Y_n)\}$. Thus $\lambda(Y_n) \to \infty$ cannot hold, so that $R = \infty$ and $\mathbb{P}_i(\omega(\Delta) < \infty) = 0$. □

Problems

2.1 Let $\{X_t\}$ be explosive and modify the process so as to restart in some fixed state i after each explosion. Show that we obtain a Markov jump process.

2.2 Let $E = \mathbb{Z}\setminus\{0\}$ and $\lambda(k) = k^2$, $q_{n(-n-1)} = q_{(-n)(n+1)} = 1/n^2$, $q_{n(n+1)} = q_{(-n)(-n-1)} = 1 - 1/n^2$, $n > 0$. Show that the process is explosive and that $0 < \mathbb{P}F_+ < 1$, $\mathbb{P}F_+ + \mathbb{P}F_- = 1$ where $F_\pm = \left\{\lim_{t\uparrow\omega(\Delta)} X_t = \pm\infty\right\}$. Show that we get a Markov process by letting $X_{\omega(\Delta)} = 1$ on F_+, $= -1$ on F_- (and similarly for the explosions after $\omega(\Delta)$).

2.3 Let $E = \mathbb{Z} \cup \{\Delta\}$ and $\lambda(k) = (k+1)^2$, $q_{k(k+1)} = 1$ for all $k \in \mathbb{Z}$. Show that the process is explosive and (at least heuristically) that there exists a version with Δ as instantaneous state and $X_t \to -\infty$, $t \downarrow \omega(\Delta)$ [such a version cannot be pure jump in the present strict sense, of course].

2.4 Let $E_1 \subset E_2 \subset \cdots \subset E$ be finite sets with $E_n \uparrow E$. Assume that $\lambda(i,j) = 0$ when $i \in E_n$, $j \in E_{n+k}$ with $k > 1$ and that $\lambda(i,j)$ is bounded uniformly in n, $i \in E_n$, $j \in E_{n+1}$. Show that the process is nonexplosive.

3 The Intensity Matrix

3a. Definition and Uniqueness
3b. Reformulations and Examples
3c. Reuter's Explosion Condition
3d. The Forward and Backward Equations

3a Definition and Uniqueness

Assume from now on $q_{ii} = 0$ when $\lambda(i) > 0$ and define the *intensity matrix* $\boldsymbol{\Lambda} = (\lambda(i,j))_{i,j\in E}$ of the process by

$$\lambda(i,j) = \lambda(i) q_{ij}, \; j \neq i, \quad \lambda(i,i) = -\lambda(i). \tag{3.1}$$

Proposition 3.1 *An $E \times E$ matrix $\boldsymbol{\Lambda}$ is the intensity matrix of a Markov jump process $\{X_t\}$ if and only if*

$$\lambda(i,i) \le 0, \quad \lambda(i,j) \ge 0,\ j \ne i, \quad \sum_{j \in E} \lambda(i,j) = 0. \tag{3.2}$$

Furthermore, $\boldsymbol{\Lambda}$ is in one–to–one correspondence with the distribution of the minimal process.

Proof. If $\boldsymbol{\Lambda}$ is an intensity matrix, it follows from (3.1) by considering the cases $\lambda(i) = 0$ and $\lambda(i) > 0$ separately that $\sum_{j \ne i} \lambda(i,j) = \lambda(i)$ and therefore that (3.2) holds. Conversely, if (3.2) is satisfied, then we let $\lambda(i) = -\lambda(i,i)$, define q_{ij} by (3.1) and $q_{ii} = 0$ if $\lambda(i) > 0$, and let $q_{ij} = \delta_{ij}$ otherwise. It is then a matter of routine to check from (3.2) that $\boldsymbol{Q}$ is a transition matrix, and clearly the Markov jump process determined by $\boldsymbol{Q}$ and the $\lambda(i)$ has intensity matrix $\boldsymbol{\Lambda}$. The stated one–to–one correspondence is obvious from Sections 1 and 2. □

3b Reformulations and Examples

It is now possible to give a reinterpretation of the intuitive picture of the evolvement of a Markov jump process which has been developed in Sections 1 and 2. So far, by the well–known interpretation of the intensity parameter of the exponential distribution this picture has been that the process, when in state i at time t, exits from i before $t + \mathrm{d}t$ with probability (risk) $\lambda(i)\mathrm{d}t$. The next value j is selected independently of the time of exit from i and according to q_{ij}. However, we can now instead consider the process as subject to (with a terminology used in survival analysis) *competing risks* with intensities $\lambda(i,j)$, $j \ne i$. That is, after entrance to i the jth type of event has an exponential waiting time Z_{ij} and the Z_{ij} are independent. Physically only the first (say $J = j$) of the events occur at time $Z_i = \min_j Z_{ij}$ and the process then jumps to j. That this yields the given transition mechanism is checked as follows:

$$\begin{aligned}
\mathbb{P}(Z_i > z, J = j) &= \mathbb{P}(Z_{ik} > Z_{ij} > z,\ k \ne j) \\
&= \lambda(i,j) \int_z^\infty \mathbb{P}(Z_{ik} > y,\ k \ne j)\mathrm{e}^{-\lambda(i,j)y}\,\mathrm{d}y \\
&= \lambda(i,j) \int_z^\infty \prod_{k \ne j} \mathrm{e}^{-\lambda(i,k)y}\mathrm{e}^{-\lambda(i,j)y}\,\mathrm{d}y \\
&= \lambda(i,j) \int_z^\infty \mathrm{e}^{-\lambda(i)y}\,\mathrm{d}y \;=\; \frac{\lambda(i,j)}{\lambda(i)}\mathrm{e}^{-\lambda(i)z} \;=\; q_{ij}\mathrm{e}^{-\lambda(i)z}.
\end{aligned}$$

This means in infinitesimal terms that the probability of a transition to j before $t + \mathrm{d}t$ is $\lambda(i,j)\mathrm{d}t$. In standard $\mathrm{o}(\cdot)$, $\mathrm{O}(\cdot)$ notation, the meaning is that the probability of a transition to j before $t + h$ is $\lambda(i,j)h + \mathrm{o}(h)$.

A description along these lines is usually the most natural in a given practical situation, and the intensity matrix is therefore the parameter in terms of which the process is usually specified. An obvious example is a queue where arrivals occur at rate β and service is completed at rate δ (the $M/M/1$ queue, cf. III.1b). Here $E = \mathbb{N}$ and

$$\mathbf{\Lambda} \;=\; \begin{pmatrix} -\beta & \beta & 0 & 0 & 0 & \dots \\ \delta & -\beta-\delta & \beta & 0 & 0 & \dots \\ 0 & \delta & -\beta-\delta & \beta & 0 & \dots \\ \vdots & & & & & \ddots \end{pmatrix}.$$

When started at state $i > 0$ at time $t = 0$, we may think of $Z_{i(i-1)}$ as the service time of the customer being presently handled by the server, and of $Z_{i(i+1)}$ as the waiting time until the next arrival. In contrast, the holding time $T_0 = Z_i$ is the time until either an arrival occurs or service is completed, and is not quite as intuitive as the Z_{ij}.

In some situations it may also be convenient to extend the sample space of the minimal construction in order that certain random variables naturally associated with the process are well defined. An example is a linear birth–death process, i.e. $E = \mathbb{N}$ and

$$\mathbf{\Lambda} \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \dots \\ \delta & -\beta-\delta & \beta & 0 & 0 & \dots \\ 0 & 2\delta & -2\beta-2\delta & 2\beta & 0 & \dots \\ \vdots & & & & & \ddots \end{pmatrix},$$

where one may think of X_t as the total size at time t of a population with individuals who (independently of one another) terminate their lives with intensity δ and give birth with intensity β. Here quantities like the individual lifetimes or the number of children of an individual are not recognizable from the minimal construction and a more natural construction proceeds as follows: represent each individual by its *life*, i.e. the pair of its lifetime Z (exponential with intensity δ) and an independent Poisson process with intensity β whose events in $[0, Z)$ correspond to the birth times. Construct the process, started from say $X_0 = 1$, from a sequence of i.i.d. lives by letting the first correspond to the ancestor, the second to his first child, ..., the nth to the nth individual being born; see Fig. 3.1. Such variants of the minimal construction will sometimes be used without further notice.

As a by–product and further illustration of the above discussion, we shall also show an important property of the exponential distribution (which is also easily proved by a direct analytical argument; cf. Problem 3.1):

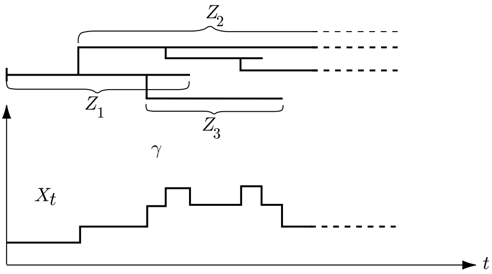

Figure 3.1

Lemma 3.2 *Let $T_0, T_1, \ldots$ be i.i.d. and exponential with intensity δ, and let N be independent of the T_n and geometric, $\mathbb{P}(N = n) = (1-\rho)\rho^{n-1}$, $n = 1, 2, \ldots$. Then $S = T_0 + \cdots + T_{N-1}$ is exponential with intensity $\eta = \delta(1-\rho)$.*

Proof. Let the intensities of a three–state process be specified by Fig. 3.2 where $\beta = \rho\delta$.

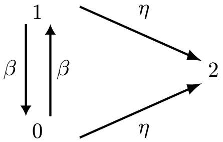

Figure 3.2

Then if we start the process in 1, the sojourn times $T_0, T_1, \ldots$ and $N = \inf\{n \geq 1 : Y_n = 2\}$ satisfy the given assumptions because of $\beta/(\beta+\eta) = \rho$ and $\beta + \eta = \delta$, and S is just the entrance time $\omega(2) = \inf\{t > 0 : X_t = 2\}$ of 2. On the other hand, the symmetry between 0 and 1 ensures that the distribution of $\omega(2)$ is left unchanged if we collapse 0, 1 into the single state 1 according to $1 \xrightarrow{\eta} 2$. This makes it clear that $\mathbb{P}(\omega(2) > s) = \mathrm{e}^{-\eta s}$. □

Problems

3.1 Show Lemma 3.2 (a) using Laplace transforms, (b) by showing that S has failure rate η.

3c Reuter's Explosion Condition

The following result is of a similar form as the transience criterion I.5.2 for Markov chains and gives a necessary and sufficient condition (known as *Reuter's condition*) for a Markov jump process to be explosive; for a nontrivial application (birth–death processes), see III.2.2.

Proposition 3.3 *A Markov jump process is nonexplosive if and only if the only nonnegative bounded solution $\boldsymbol{k} = (k_i)_{i\in E}$ to the set of equations $\boldsymbol{\Lambda k} = \boldsymbol{k}$ is $\boldsymbol{k} = \boldsymbol{0}$.*

Proof. Suppose first that the process is explosive and define $k_i = \mathbb{E}_i \mathrm{e}^{-\omega(\Delta)}$. Then $\boldsymbol{k} = (k_i)$ is bounded and $k_i > 0$ at least for one i. Conditioning upon the time $T_0 = y$ of the first jump we get

$$k_i \;=\; \int_0^\infty \sum_{j\neq i} \lambda(i,j)\mathbb{E}_j \mathrm{e}^{-y-\omega(\Delta)}\mathrm{e}^{-\lambda(i)y}\,\mathrm{d}y \;=\; \sum_{j\neq i}\frac{\lambda(i,j)k_j}{1+\lambda(i)}.$$

Using $\lambda(i) = -\lambda(i,i)$, this implies $k_i = \sum_{j\in E}\lambda(i,j)k_j$ and $\boldsymbol{k} = \boldsymbol{\Lambda k}$. Suppose, conversely, the process is nonexplosive, and define $h_i^{(n)} = \mathbb{E}_i \exp\{-T_0 - \cdots - T_{n-1}\}$, $h_i^{(0)} = 1$. Then just as above

$$h_i^{(n+1)} \;=\; \int_0^\infty \sum_{j\neq i}\lambda(i,j)\mathrm{e}^{-y}h_j^{(n)}\mathrm{e}^{-\lambda(i)y}\,\mathrm{d}y \;=\; \sum_{j\neq i}\frac{\lambda(i,j)h_j^{(n)}}{1+\lambda(i)}. \qquad (3.3)$$

Now let $\boldsymbol{k} \geq \boldsymbol{0}$ be bounded (w.l.o.g. $k_i \leq 1$) with $\boldsymbol{\Lambda k} = \boldsymbol{k}$. Then $k_i = \sum_{j\neq i}\lambda(i,j)k_j/(1+\lambda(i))$, and since $1 = h_i^{(0)} \geq k_i$, it follows by induction from (3.3) that $h_i^{(n)} \geq k_i$ for all n. But $T_0 + \cdots + T_n \uparrow \omega(\Delta) = \infty$ implies $h_i^{(n)} \to 0$. Hence $k_i = 0$ for all i. □

Problems

3.2 Consider a pure birth process ($E = \mathbb{N}$, $\lambda(i,i+1) = \lambda(i) = \beta_i$). Show that the process is nonexplosive if and only if $\sum_0^\infty \beta_n^{-1} = \infty$, and check that Propositions 2.3 and 3.3 yield the same result.

3d The Forward and Backward Equations

We now turn to one of the most celebrated classical topics in Markov process theory:

Theorem 3.4 *Let* $\boldsymbol{\Lambda}$ *be an intensity matrix on* E *and* $\{X_t\}$ *the corresponding minimal Markov jump process on* E *constructed in Theorem 2.1,* $p_{ij}^t = \mathbb{P}_i(X_t = j)$. *Then the* $E \times E$*–matrices* $\boldsymbol{P}^t = (p_{ij}^t)$*satisfy the backward equation* $(\mathrm{d}/\mathrm{d}t)\boldsymbol{P}^t = \boldsymbol{\Lambda}\boldsymbol{P}^t$, *i.e.*

$$\frac{\mathrm{d}p_{ij}^t}{\mathrm{d}t} \;=\; \sum_{k\in E}\lambda(i,k)p_{kj}^t \qquad (3.4)$$

and the forward equation $(\mathrm{d}/\mathrm{d}t)\boldsymbol{P}^t = \boldsymbol{P}^t\boldsymbol{\Lambda}$, *i.e.*

$$\frac{\mathrm{d}p_{ij}^t}{\mathrm{d}t} \;=\; \sum_{k\in E}p_{ik}^t\lambda(k,j)\,. \qquad (3.5)$$

Proof. Conditioning upon $T_0 = s$ yields

$$p_{ij}^t \;=\; \mathbb{P}_i(T_0 > t)\delta_{ij} \;+\; \int_0^t \lambda(i)\mathrm{e}^{-\lambda(i)s}\sum_{k\neq i}q_{ik}p_{kj}^{t-s}\,\mathrm{d}s$$

$$= \quad \mathrm{e}^{-\lambda(i)t}\left[\delta_{ij} + \int_0^t \sum_{k\neq i} \lambda(i,k)\mathrm{e}^{\lambda(i)s} p_{kj}^s \,\mathrm{d}s\right].$$

The integrand $f(s) = \sum_{k\neq i}\cdots$ is well defined with $\sup_{s\leq T} f(s)$ for all $T < \infty$ since $\sum|\lambda(i,k)| = 2\lambda(i) < \infty$. This shows first that p_{ij}^t (and similarly all p_{kj}^t) is continuous and thereafter that $f(s)$ is continuous. Therefore p_{ij}^t is differentiable with derivative

$$\begin{aligned} & -\lambda(i)\mathrm{e}^{-\lambda(i)t}\left[\delta_{ij} + \int_0^t f(s)\,\mathrm{d}s\right] + \mathrm{e}^{-\lambda(i)t} f(t) \\ = \quad & -\lambda(i)p_{ij}^t + \sum_{k\neq i}\lambda(i,k)p_{kj}^t \;=\; \sum_{k\in E}\lambda(i,k)p_{kj}^t. \end{aligned}$$

The proof of the forward equation is more involved and *will only be given subject to the assumption*

$$\sup_{i\in E}\lambda(i) < \infty \tag{3.6}$$

which will be used to infer that the $(p_{kj}^s - \delta_{kj})/s$ are bounded uniformly in s, j, k. This follows since

$$\begin{aligned} 0 \;&\leq\; p_{kj}^s \;\leq\; \lambda(k)\int_0^s \mathrm{e}^{-\lambda(k)u}\,\mathrm{d}u, \quad k\neq j, \\ 0 \;&\leq\; 1 - p_{kk}^s \;\leq\; \lambda(k)\int_0^s \mathrm{e}^{-\lambda(k)u}\,\mathrm{d}u, \end{aligned}$$

and (3.5) comes out by dominated convergence (using $\sum p_{ik}^t < \infty$) from

$$\frac{p_{ij}^{t+s} - p_{ij}^t}{s} \;=\; \sum_{k\in E} p_{ik}^t \frac{p_{kj}^s - \delta_{kj}}{s} \;\to\; \sum_{k\in E} p_{ik}^t \lambda(k,j). \qquad \square$$

In the case of a finite E, standard results on existence and uniqueness of systems of linear differential equations yield together with $\boldsymbol{P}^0 = \boldsymbol{I}$ yield:

Corollary 3.5 *If E is finite, then* $\boldsymbol{P}^t \;=\; \mathrm{e}^{\boldsymbol{\Lambda}t} \;=\; \displaystyle\sum_{n=0}^{\infty}\frac{t^n}{n!}\boldsymbol{\Lambda}^n,\ t\geq 0.$

Example 3.6 Suppose that E has just $p = 2$ states 1, 2 and, to avoid trivialities, that $\lambda(1)$ and $\lambda(2)$ are not both zero. Then $\boldsymbol{\Lambda}$ has eigenvalues 0 and $\lambda = -\lambda(1) - \lambda(2)$ with corresponding right eigenvectors $(1\ 1)^{\mathsf{T}}$, $\big(\lambda(1)\ \ -\lambda(2)\big)^{\mathsf{T}}$. Hence

$$\begin{aligned} \boldsymbol{\Lambda} \;&=\; \begin{pmatrix} -\lambda(1) & \lambda(1) \\ \lambda(2) & -\lambda(2) \end{pmatrix} \;=\; \boldsymbol{B}\begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}\boldsymbol{B}^{-1} \quad \text{where} \\ \boldsymbol{B} \;&=\; \begin{pmatrix} 1 & \lambda(1) \\ 1 & -\lambda(2) \end{pmatrix}, \quad \boldsymbol{B}^{-1} \;=\; \frac{1}{\lambda(1)+\lambda(2)}\begin{pmatrix} \lambda(2) & \lambda(1) \\ 1 & -1 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned}\boldsymbol{P}^t &= \mathrm{e}^{\boldsymbol{\Lambda}t} = \sum_{n=0}^{\infty}\frac{t^n}{n!}\boldsymbol{B}\begin{pmatrix}0 & 0\\ 0 & \lambda^n\end{pmatrix}\boldsymbol{B}^{-1} = \boldsymbol{B}\begin{pmatrix}1 & 0\\ 0 & \mathrm{e}^{\lambda t}\end{pmatrix}\boldsymbol{B}^{-1}\\ &= \frac{1}{\lambda(1)+\lambda(2)}\begin{pmatrix}\lambda(2)+\lambda(1)\mathrm{e}^{\lambda t} & \lambda(1)-\lambda(1)\mathrm{e}^{\lambda t}\\ \lambda(2)-\lambda(2)\mathrm{e}^{\lambda t} & \lambda(1)+\lambda(2)\mathrm{e}^{\lambda t}\end{pmatrix}.\end{aligned}$$

□

For purposes like those of the present book, the backward and forward equations are of quite limited utility. This is so in particular for an infinite state space, but even for $p < \infty$ states the Jordan canonical form and hence the algebra corresponding to Example 3.6 becomes much more cumbersome for $p > 2$ due to the possibility of eigenvalues which are complex or of multiplicity > 1. One common application is to look for a stationary probability distribution ($\boldsymbol{\pi}\boldsymbol{P}^t = \boldsymbol{\pi}$) by means of $(\boldsymbol{\pi}\boldsymbol{P}^t)'|_{t=0} = \boldsymbol{0}$, i.e. $\boldsymbol{\pi}\boldsymbol{\Lambda} = \boldsymbol{0}$. This equation comes out, however, quite easily by a direct argument in the next section. Also the time–dependent solution (i.e. the p_{ij}^t for $t < \infty$) can be found explicitly only in very special cases when E is infinite and is even then frequently easier to obtain by different means. Examples are the linear birth–death process (see e.g. Harris, 1963) and the $M/M/1$ queue to be discussed in III.8.

4 Stationarity and Limit Results

4a. Classification of States
4b. Stationary Measures
4c. Ergodicity Criteria and Limit Results
4d. Spectral Properties, the Fundamental Matrix and the CLT

4a Classification of States

When defining concepts such as irreducibility, recurrence or transience in continuous time, one may either mimic the discrete time definition or refer to the jump chain $\{Y_n\}$. We consider in the following a minimal process and look first at irreducibility. Then:

Proposition 4.1 *The following properties are equivalent:* (a) *$\{Y_n\}$ is irreducible;* (b) *for any $i, j \in E$ we have $p_{ij}^t > 0$ for some $t > 0$;* (c) *for any $i, j \in E$ we have $p_{ij}^t > 0$ for all $t > 0$.*

Proof. Denote here and in the following

$$\omega(i) = \inf\bigl\{t > 0 : X_t = i, \lim_{s \uparrow t} X_s \neq i\bigr\} \tag{4.1}$$

($\omega(i) = \infty$ if no such t exists) so that $\omega(i)$ is the hitting time of i if $X_0 \neq i$ and the recurrence time of i if $X_0 = i$. Since j has an exponential holding time, it is clear that $p_{ij}^t > 0$ if and only if $\mathbb{P}_i(\omega(j) \leq t) > 0$ (similarly we always have $p_{ii}^t > 0$). Now $\mathbb{P}_i(\omega(i) \leq t) > 0$ if and only if some path $ii_1 \cdots i_n j$ from i to j is possible for $\{Y_n\}$, and in that case we may evaluate the conditional distribution F of $\omega(j)$ given $\omega(j) < \infty$ by conditioning on the various paths. Thus F is a mixture of convolutions of exponential distributions with intensities $\lambda(i, i_1)$, $\lambda(i_1, i_2)$, ... and hence has a density > 0 on $(0, \infty)$. Thus $\mathbb{P}_i(\omega(j) \leq t) > 0$ if and only if $\{Y_n\}$ can reach j from i, proving the proposition. □

Accordingly we define $\{X_t\}$ to be irreducible if one of properties (a), (b), (c) hold. Similarly (but easier), it is seen that we can define i to be transient (recurrent) for $\{X_t\}$ if either (a) the set $\{t : X_t = i\}$ is bounded (unbounded) $\mathbb{P}_i$–a.s., (b) i is transient (recurrent) for $\{Y_n\}$ or (c) $\mathbb{P}_i(\omega(i) < \infty) < 1$ $(= 1)$. As will be seen in the following, the distinction between null recurrence and positive recurrence cannot, however, be related to $\{Y_n\}$ alone. Note also that we do not pay attention to periodicity. This is due to the fact that even though $\{Y_n\}$ may be periodic, the exponential holding times smooth away any such behaviour in continuous time.

4b Stationary Measures

A measure $\boldsymbol{\nu} \neq \mathbf{0}$ is *stationary* if $0 \leq \nu_j < \infty$, $\boldsymbol{\nu}\boldsymbol{P}^t = \boldsymbol{\nu}$ for all t.

Theorem 4.2 *Suppose that $\{X_t\}$ is irreducible and recurrent on E. Then there exists one, and up to a multiplicative factor only one, stationary measure $\boldsymbol{\nu}$. This $\boldsymbol{\nu}$ has the property $\nu_j > 0$ for all j and can be found in either of the following ways:*
(i) *for some fixed but arbitrary state i, ν_j is the expected time spent in j between successive entrances to i. That is, with $\omega(i)$ given by* (4.1)

$$\nu_j \;=\; \mathbb{E}_i \int_0^{\omega(i)} I(X_t = j)\,\mathrm{d}t; \tag{4.2}$$

(ii) $\nu_j = \mu_j/\lambda(j)$, *where $\boldsymbol{\mu}$ is stationary for $\{Y_n\}$;*
(iii) *as solution of $\boldsymbol{\nu}\boldsymbol{\Lambda} = \mathbf{0}$.*

Proof. We first prove uniqueness by considering the Markov chain $X_0, X_1, \ldots$. This is irreducible since all $p_{ij}^t > 0$, and any $\boldsymbol{\nu}$ stationary for $\{X_t\}$ is also stationary for $\{X_n\}$, so in order to apply I.3.4 we just have to show that $\{X_n\}$ is recurrent. But for any i, the sequence $U_1, U_2, \ldots$ of holding times of i is nonterminating since i is recurrent. The U_k being i.i.d. with $\mathbb{P}(U_k > 1) > 0$, we have $U_k > 1$ i.o. and therefore also $X_n = i$ i.o.

For (i) we show the stationarity of (4.2) by evaluating the jth component of $\boldsymbol{\nu}\boldsymbol{P}^h$ in a somewhat similar manner as in the proof of I.3.2. First note that $\{X_t\}_{0\leq t\leq h}$ and $\left\{X_{\omega(i)+t}\right\}_{0\leq t\leq h}$ have the same $\mathbb{P}_i$–distribution because

$X_{\omega(i)} = i$. Hence

$$\begin{aligned}\nu_j &= \mathbb{E}_i\Bigl[\int_0^h + \int_h^{\omega(i)} I(X_t = j)\,\mathrm{d}t\Bigr] \;=\; \mathbb{E}_i\Bigl[\int_{\omega(i)}^{\omega(i)+h} + \int_h^{\omega(i)} I(X_t = j)\,\mathrm{d}t\Bigr]\\ &= \mathbb{E}_i\Bigl[\int_h^{\omega(i)+h} I(X_t = j)\,\mathrm{d}t\Bigr] \;=\; \mathbb{E}_i\int_0^{\omega(i)} I(X_{t+h} = j)\,\mathrm{d}t\end{aligned}$$

(note that the first equality is valid also if $\omega(i) < h$). Thus

$$\begin{aligned}\nu_j &= \mathbb{E}_i\int_0^\infty \mathbb{P}\bigl(X_{t+h} = j, \omega(i) > t \,\big|\, \mathscr{F}_t\bigr)\,\mathrm{d}t \;=\; \mathbb{E}_i\int_0^\infty p^h_{X_t j} I(\omega(i) > t)\,\mathrm{d}t\\ &= \sum_{k\in E} p^h_{kj}\mathbb{E}_i\int_0^\infty I(X_t = k, \omega(i) > t)\,\mathrm{d}t \;=\; \sum_{k\in E}\nu_k p^h_{kj},\end{aligned}$$

proving $\boldsymbol{\nu}\boldsymbol{P}^h = \boldsymbol{\nu}$ and (i). With $\tau(i) = \inf\{n : Y_n = i\}$, we then get

$$\begin{aligned}\nu_j &= \mathbb{E}_i\int_0^{\omega(i)} I(X_t = j)\,\mathrm{d}t \;=\; \mathbb{E}_i\sum_{n=0}^{\tau(i)-1} T_n I(Y_n = j)\\ &= \sum_{n=0}^{\infty}\mathbb{E}_i\mathbb{E}_i\bigl[T_n;\, Y_n = j, \tau(i) > n \,\big|\, \{Y_n\}_0^\infty\bigr]\\ &= \frac{1}{\lambda(j)}\mathbb{E}_i\sum_{n=0}^{\infty} I(Y_n = j, \tau(i) > n) \;=\; \frac{1}{\lambda(j)}\frac{\mu_j}{\mu_i},\end{aligned}$$

using I.(3.1) in the last step. That is, ν_j is proportional to $\mu_j/\lambda(j)$, showing (ii).

For (iii) we note that according to (ii) $\boldsymbol{\nu}$ is stationary for $\{X_t\}$ if and only if $(\nu_j\lambda(j))_{j\in E}$ is stationary for $\{Y_n\}$, i.e. if and only if $\sum_{i\in E}\nu_i\lambda(i)q_{ij} = \nu_j\lambda(j)$ for all $j \in E$, or, since $q_{ii} = 0$, if and only if

$$0 \;=\; -\nu_j\lambda(j) + \sum_{i\neq j}\nu_i\lambda(i,j) \;=\; \sum_{i\in E}\nu_i\lambda(i,j)\,.$$

Finally, $0 < \nu_j < \infty$ follows easily say by (ii), since in the recurrent case $0 < \lambda(j) < \infty$. □

4c Ergodicity Criteria and Limit Results

An irreducible recurrent process with the stationary measure having finite mass is called *ergodic*, and we have:

Theorem 4.3 *An irreducible nonexplosive Markov jump process is ergodic if and only if one can find a probability solution* $\boldsymbol{\pi}$ $(\boldsymbol{\pi}\mathbf{1} = 1,\, 0 \le \pi_i \le 1)$ *to* $\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0}$. *In that case* $\boldsymbol{\pi}$ *is the stationary distribution.*

Proof. That a solution exists and is stationary in the ergodic case follows immediately from Theorem 4.2. Suppose conversely that a solution exists and define

$$p_{ij}^{t;n} \;=\; \mathbb{P}_i(X_t = j, T_0 + \cdots + T_n > t), \quad n = 0, 1, 2, \ldots .$$

Now a path starting from i contributes to $p_{ij}^{t;n}$ either if it has no jumps before t (and $i = j$) or it has a last jump, say from k to j at time $s \le t$, and at most $n-1$ jumps before s. Thus collecting terms we get

$$p_{ij}^{t;n} \;=\; \delta_{ij}\mathrm{e}^{-\lambda(i)t} + \int_0^t \sum_{k \neq j} p_{ik}^{s;n-1}\lambda(k,j)\mathrm{e}^{-\lambda(j)(t-s)}\,\mathrm{d}s\,,$$

$$\sum_{i\in E}\pi_i p_{ij}^{t;n} \;=\; \pi_j \mathrm{e}^{-\lambda(i)t} + \int_0^t \mathrm{e}^{-\lambda(j)(t-s)} \sum_{k\neq j}\lambda(k,j)\sum_{i\in E}\pi_i p_{ik}^{s;n-1}\,\mathrm{d}s. \quad (4.3)$$

Obviously

$$\sum_{i\in E}\pi_i p_{ij}^{t;0} \;=\; \pi_j \mathrm{e}^{-\lambda(j)t} \;\le \pi_j, \quad \text{i.e. } \boldsymbol{\pi}\boldsymbol{P}^{t;0} \le \boldsymbol{\pi}.$$

It thus follows by induction from (4.3) that $\boldsymbol{\pi}\boldsymbol{P}^{t;n} \le \boldsymbol{\pi}$ since then

$$\begin{aligned}\sum_{i\in E}\pi_i p_{ij}^{t;n+1} \;&\le\; \pi_j \mathrm{e}^{-\lambda(j)t} + \int_0^t \mathrm{e}^{-\lambda(j)(t-s)}\sum_{k\neq j}\lambda(k,j)\pi_k\,\mathrm{d}s \\ &=\; \pi_j \mathrm{e}^{-\lambda(j)t} + \pi_j\lambda(j)\int_0^t \mathrm{e}^{-\lambda(j)s}\,\mathrm{d}s \;=\; \pi_j .\end{aligned}$$

But since the process is nonexplosive, we have $p_{ij}^{t;n} \to p_{ij}^t$ and $\sum_{j\in E} p_{ij}^t = 1$. Hence $\sum_{i\in E}\pi_i p_{ij}^t \le \pi_j$, and since summing both sides over j yields 1, equality must hold so that $\boldsymbol{\pi}\boldsymbol{P}^t = \boldsymbol{\pi}$. Thus $\boldsymbol{\pi}$ is a stationary distribution. This implies recurrence (since in the transient case $\mathbb{P}_{\boldsymbol{\pi}}(X_t = j) \to 0$) and $\boldsymbol{\pi}\mathbf{1} = 1$ then finally shows ergodicity. □

As noted in Section 3, the equation $\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0}$ is the same as that which comes out by formal manipulations with the differential equations. In the literature one occasionally proves ergodicity by checking irreducibility and finding a probability solution to $\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0}$. *This procedure is, however, not valid without having excluded explosion.* To see this, consider for example a *transient* $\{Y_n\}$ with a stationary measure $\boldsymbol{\mu}$ (for an example, see Problem I.3.2) and choose the $\lambda(j)$ such that $\pi_j = \mu_j/\lambda(j)$ has mass 1. Then as in the proof of Theorem 4.2(iii), it holds that $\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0}$, and clearly the transience of $\{Y_n\}$ excludes recurrence of $\{X_t\}$ (it follows from Theorem 4.3 that $\{X_t\}$ must even be explosive). However:

Corollary 4.4 *A sufficient condition for ergodicity of an irreducible process is the existence of a probability $\boldsymbol{\pi}$ that solves $\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0}$ and has the additional property $\sum \pi_j\lambda(j) < \infty$ (which is automatic if $\sup_{i\in E}\lambda(i) < \infty$).*

Proof. Letting $\mu_j = \pi_j \lambda(j)$, it follows as in the proof of part (iii) of Theorem 4.2 that $\boldsymbol{\mu}$ is stationary for $\{Y_n\}$. Since $\boldsymbol{\mu}$ has finite mass, $\{Y_n\}$ is positive recurrent, in particular recurrent, hence $\{X_t\}$ is nonexplosive and Theorem 4.3 applies. □

Exactly as in I.3.1(ii) one also has:

Proposition 4.5 *If the process is ergodic, then there exists a strictly stationary version* $\{X_t\}_{-\infty<t<\infty}$ *with doubly infinite time.*

We next turn to the limiting behaviour of the p_{ij}^t and have as expected:

Theorem 4.6 *If* $\{X_t\}$ *is ergodic and* $\boldsymbol{\pi}$ *the stationary distribution, then* $p_{ij}^t \to \pi_j$, $t \to \infty$, *for all* i, j.

Proof. As noted above in the case $\delta = 1$, $\{X_{n\delta}\}$ is an irreducible recurrent aperiodic Markov chain for each δ. It is ergodic since $\boldsymbol{\pi}$ is stationary, and hence $p_{ij}^{n\delta} \to \pi_j$ as $n \to \infty$. The continuity of the p_{ij}^t being straightforward to verify, the assertion thus follows by the method of discrete skeletons, A11.2. Alternatively, we may apply the more elementary A11.1. The required uniform continuity follows say from the backward equation (3.4) which in conjunction with $\sum |\lambda(i,k)| < \infty$ shows that $\mathrm{d}p_{ij}^t/\mathrm{d}t$ exists and is bounded in t. □

As in discrete time, I.(4.3), time–average properties like

$$\frac{1}{T}\int_0^T f(X_t)\,\mathrm{d}t \overset{\text{a.s.}}{\to} \pi(f) \;=\; \mathbb{E}_{\boldsymbol{\pi}} f(X_t) = \sum_{i\in E} \pi_i f(i) \qquad (4.4)$$

hold under suitable conditions on f; see VI.3.

Exactly the same argument as for Theorem 4.6 yields

Corollary 4.7 *If* $\{X_t\}$ *is irreducible recurrent but not ergodic* (i.e. $\boldsymbol{\nu}\mathbf{1} = \infty$), *then* $p_{ij}^t \to 0$ *for all* $i, j \in E$.

Corollary 4.8 *For any minimal Markov jump process* (irreducible or not), *the limits* $\lim_{t\to\infty} p_{ij}^t$ *exist* [recall that in discrete time periodicity might cause an exception to the parallel result].

Proof. Clearly $p_{ij}^t \to 0$ if j is transient. If j is in a recurrent class C, let $\boldsymbol{\nu}^{(C)}$ be stationary for the process restricted to C. Then by Theorem 4.6 and Corollary 4.7

$$p_{ij}^t \;\to\; \mathbb{P}_i(\text{some } X_t \in C)\frac{\nu_j^{(C)}}{\boldsymbol{\nu}^{(C)}\mathbf{1}}\,. \qquad □$$

Problems

4.1 Consider a Markov jump process $\{X_t\}$ with bounded intensities, say $\lambda(i) \le \lambda < \infty$. Show that $\widetilde{\boldsymbol{Q}} = \boldsymbol{\Lambda}/\lambda + \boldsymbol{I}$ is a transition matrix. Now consider a Poisson process $\{N_t\}$ with intensity λ and a process $\{\widetilde{X}_t\}$ which jumps according to $\widetilde{\boldsymbol{Q}}$ at the jumps of $\{N_t\}$, say $\widetilde{X}_t = \widetilde{Y}_n$ on $\{N_t = n\}$ where $\{\widetilde{Y}_n\}$ is a Markov chain governed by $\widetilde{\boldsymbol{Q}}$ and independent of $\{N_t\}$. Show that $\{\widetilde{X}_t\}$ is a version of $\{X_t\}$, that $\{X_t\}$ is ergodic if and only if $\{\widetilde{Y}_n\}$ is positive recurrent and that then the stationary distributions are the same (this procedure is known as *uniformization*; see e.g. Keilson, 1979).

Notes For further discussion of the equation $\boldsymbol{\pi}\boldsymbol{\Lambda} = \boldsymbol{0}$ is the explosive case, see Kelly (1983).

4d Spectral Properties, the Fundamental Matrix and the CLT

Corollary 4.9 *If $\boldsymbol{\Lambda}$ is an irreducible $p \times p$ intensity matrix, then 0 is eigenvalue with left and right eigenvectors $\boldsymbol{\pi}$, resp. $\boldsymbol{1}$. Any other eigenvalue has strictly negative real part.*

Proof. The first statement is obvious. For the second, let $a > 0$ be larger than any $|\lambda(i,i)|$ and consider $\boldsymbol{A} = \boldsymbol{\Lambda}/a + \boldsymbol{I}$. Then $\boldsymbol{A}$ is an ergodic transition matrix, so that by I.6 one of the eigenvalues s, say, is $s = 1$ and all others have $|s| < 1$. Since the eigenvalues of $\boldsymbol{\Lambda}$ are precisely the numbers of the form $a(s-1)$, the assertion follows. □

Define the *fundamental matrix* by

$$\boldsymbol{Z} \;=\; \int_0^\infty \left(\mathrm{e}^{\boldsymbol{\Lambda}t} - \boldsymbol{1}\boldsymbol{\pi}\right)\mathrm{d}t \;=\; (\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{\Lambda})^{-1}(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\pi}). \tag{4.5}$$

Note that the existence of the integral as well as the inverse is ensured by Corollary 4.9. Note also that $\boldsymbol{Z}$ is singular ($\boldsymbol{\pi}\boldsymbol{Z} = \boldsymbol{0}$ and $\boldsymbol{Z}\boldsymbol{1} = \boldsymbol{0}$) in contrast to the discrete time case in I.7. The second expression for $\boldsymbol{Z}$ follows from

$$(\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{\Lambda})\int_0^T \left(\mathrm{e}^{\boldsymbol{\Lambda}t} - \boldsymbol{1}\boldsymbol{\pi}\right)\mathrm{d}t \;=\; -\int_0^T \boldsymbol{\Lambda}\mathrm{e}^{\boldsymbol{\Lambda}t}\,\mathrm{d}t \;=\; \boldsymbol{I} - \mathrm{e}^{\boldsymbol{\Lambda}T} \;\to\; \boldsymbol{I} - \boldsymbol{1}\boldsymbol{\pi}.$$

Poisson's equation in continuous time is $\boldsymbol{\Lambda}\boldsymbol{g} = -\boldsymbol{f}$.

Proposition 4.10 *Assume $\pi(f) = 0$. Then $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{f}$ solves Poisson's equation and is the unique solution with $\pi(g) = 0$.*

Proof. Using the second expression in (4.5) and $\boldsymbol{\pi}\boldsymbol{Z} = \boldsymbol{0}$, we get

$$-\boldsymbol{\Lambda}\boldsymbol{g} \;=\; (-\boldsymbol{1}\boldsymbol{\pi} + \boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{\Lambda})\boldsymbol{g} \;=\; -\boldsymbol{1}\boldsymbol{\pi}\boldsymbol{Z}\boldsymbol{f} + (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\pi})\boldsymbol{f} \;=\; -\boldsymbol{0} + \boldsymbol{f} - \boldsymbol{0}.$$

Uniqueness follows since the difference $\boldsymbol{d}$ between two solutions satisfies $\boldsymbol{\Lambda}\boldsymbol{d} = \boldsymbol{0}$. Hence $\boldsymbol{d} = c\boldsymbol{1}$, and $\pi(d) = 0$ then gives $c = 0$. □

Theorem 4.11 *Let $f : E \to \mathbb{R}$ and define $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{f}$, $\sigma^2(f) = 2\pi(fg)$. Then*

$$\frac{1}{T^{1/2}}\left(\int_0^T f(X_t)\,\mathrm{d}t \;-\; T\pi(f)\right) \;\to\; N\big(0, \sigma^2(f)\big)\,.$$

Proof. Assume first $\pi(f) = 0$. Then $\int_0^T f(X_t)\,\mathrm{d}t = -g(X_T) + M_T$ where $M_T = g(X_T) - \int_0^T \Lambda g(X_t)\,\mathrm{d}t$ is a martingale (see the Notes to I.8). Here

$$\begin{aligned}
\mathbb{V}ar_i(M_h) &= \mathbb{E}_i\Bigl(g(X_h) - g(i) - \int_0^h \Lambda g(X_t)\,\mathrm{d}t\Bigr)^2 \\
&= h\sum_{j\neq i}\lambda(i,j)\bigl(g(j)-g(i)\bigr)^2 + \mathrm{o}(h) \\
&= h\sum_{j\in E}\lambda(i,j)\bigl(g(j)-g(i)\bigr)^2 + \mathrm{o}(h) \\
&= h\bigl(\Lambda g^2(i) - 2g(i)\Lambda g(i)\bigr) + \mathrm{o}(h),
\end{aligned}$$

$$\mathbb{V}ar_i(M_{t+h} - M_t \mid \mathscr{F}_t) \;=\; h\bigl(\Lambda g^2(X_t) - 2g(X_t)\Lambda g(X_t)\bigr) + \mathrm{o}(h).$$

Thus the quadratic variation of $\{M_t\}$ is

$$Q_T \;=\; \int_0^T \bigl(\Lambda g(X_t)^2 - 2g(X_t)\Lambda g(X_t)\bigr)\,\mathrm{d}t.$$

Since $\boldsymbol{\pi\Lambda g}^2 = \boldsymbol{0}$, $\boldsymbol{\Lambda g} = -\boldsymbol{f}$, the LLN (4.4) yields $Q_T/T \overset{\text{a.s.}}{\to} 2\pi(fg) = \sigma^2(f)$. The rest of the proof is a straightforward application of the martingale CLT as in I.7.

For the case $\pi(f) \neq 0$, see Problem 4.2. □

Problems

4.2 For $\pi(f) \neq 0$, let $\widetilde{\boldsymbol{f}} = \boldsymbol{f} - \pi(f)\boldsymbol{1}$, $\widetilde{\boldsymbol{g}} = \boldsymbol{Z}\widetilde{\boldsymbol{f}}$. Check that $\pi(fg) = \pi(\widetilde{f}\widetilde{g})$.

4.3 Verify using the diagonalization formulas in Example 3.6 that for $p = 2$ states one has

$$\begin{aligned}
\boldsymbol{Z} &= \boldsymbol{B}\begin{pmatrix} 0 & 0 \\ 0 & -\lambda^{-1}\end{pmatrix}\boldsymbol{B}^{-1} = (\lambda(1)+\lambda(2))^{-2}\begin{pmatrix} -\lambda(1) & \lambda(1) \\ \lambda(2) & -\lambda(2)\end{pmatrix}, \\
\sigma^2(f) &= \bigl(f(1) - f(2)\bigr)^2\frac{\lambda(1)\lambda(2)}{|\lambda(1)+\lambda(2)|^3}\,.
\end{aligned}$$

Notes A matrix with nonpositive off–diagonal elements is called a *Z–matrix*, and a matrix whose eigenvalues have nonnegative real parts a *M–matrix*. Thus if $\boldsymbol{\Lambda}$ is an intensity matrix, $-\boldsymbol{\Lambda}$ is a Z–matrix and (by Corollary 4.9) a M–matrix in the irreducible case. For further discussion of Z– and M–matrices and their spectral properties, see Berman and Plemmons (1994).

5 Time Reversibility

Time reversibility (or just *reversibility*) of a process means loosely that the process evolves in just the same way irrespective of whether time is read forward (as usual) or backward. The concept is studied here mainly for the purpose of certain queueing applications in Chapter IV, but its scope is in fact rather more general. For example, it could be mentioned that time reversibility of processes occuring in physics is considered a property of intrinsic physical interest.

Our main interest is in Markov jump processes, but we start by the Markov chain case in order to motivate the definition to follow and to make some simple observations.

Proposition 5.1 *Let $X_0,\ldots,X_N$ be a time–homogeneous Markov chain with transition matrix $\boldsymbol{P}$ and define the time–reversed chain $\widetilde{X}_0,\ldots,\widetilde{X}_N$ by $\widetilde{X}_n = X_{N-n}$. Consider some fixed Markov probability $\mathbb{P}$, define $\pi_i(n) = \mathbb{P}(X_n = i)$ and assume that $\pi_i(n) > 0$ for all $i \in E$, $n = 0,\ldots,N$. Then:*
(a) $\widetilde{X}_0,\ldots,\widetilde{X}_N$ is a time–inhomogeneous Markov chain with transition matrices $\widetilde{\boldsymbol{P}}(n) = \big(\widetilde{p}_{ij}(n)\big)$ given by

$$\widetilde{p}_{ij}(n) \;=\; \mathbb{P}\big(\widetilde{X}_{n+1} = j \,\big|\, \widetilde{X}_n = i\big) \;=\; \frac{\pi_j(N-n-1)p_{ji}}{\pi_i(N-n)}. \tag{5.1}$$

If furthermore all $p_{ij} > 0$, then:
(b) $\widetilde{X}_0,\ldots,\widetilde{X}_N$ is time–homogeneous, i.e. $\widetilde{\boldsymbol{P}}(n)$ independent of n, if and only if $X_0,\ldots,X_N$ is stationary, i.e. $\pi_i(n) = \pi_i$ independent of n;
(c) $\widetilde{X}_0,\ldots,\widetilde{X}_N$ has the same distribution as $X_0,\ldots,X_N$ if and only if $X_0,\ldots,X_N$ is stationary and $\pi_j p_{ji} = \pi_i p_{ij}$.

Proof. (a) Letting $\widetilde{\pi}_i(n) = \mathbb{P}(\widetilde{X}_n = i) = \pi_i(N-n)$, we must show that

$$\mathbb{P}(\widetilde{X}_0 = i_0,\ldots,\widetilde{X}_N = i_N) \;=\; \widetilde{\pi}_{i_0}(0)\widetilde{p}_{i_0i_1}(0)\widetilde{p}_{i_1i_2}(1)\ldots\widetilde{p}_{i_{N-1}i_N}(N-1). \tag{5.2}$$

But the l.h.s. of (5.2) is

$$\mathbb{P}(X_0 = i_N,\ldots,X_N = i_0) \;=\; \pi_{i_N}(0)p_{i_Ni_{N-1}}p_{i_{N-1}i_{N-2}}\cdots p_{i_1i_0}. \tag{5.3}$$

Inserting the definition of $\widetilde{\boldsymbol{P}}(n)$ in the r.h.s. of (5.2), the π_j telescope and the r.h.s. of (5.3) comes out.

(b) It is clear that $\pi_i(n) = \pi_i$ implies that $\widetilde{\boldsymbol{P}}(n)$ is independent of i. For the converse, first let $i = j$ in (5.1). It then follows that $\pi_i(n) = \pi_i\rho_i^n$ for suitable π_i, ρ_i. Since $p_{ji} > 0$, the independence of $\widetilde{p}_{ij}(n)$ of n yields $\rho_i = \rho_j$. Hence all ρ_i are equal, say $\rho_i = \rho$. But then $1 = \sum \pi_i(n) = \rho^n \sum \pi_i$ implies $\rho = 1$, i.e. $\pi_i(n) = \pi_i$ and stationarity. In (c) stationarity is necessary by (b); $\pi_j p_{ji} = \pi_i p_{ij}$ is then equivalent to $\tilde{p}_{ij} = p_{ij}$ by (5.1). □

This result does not cover the Markov chain case in full generality since all $\pi_i(n)$ and all p_{ij} being nonzero is a restriction. However, if $X_0,\ldots,X_N$ is obtained by observing an irreducible Markov jump process at times $\delta, 2\delta, \ldots, (N+1)\delta$, this assumption is automatic. Since time reversibility in continuous time should imply reversibility of such discrete skeletons, it follows by (b) that we can safely restrict attention to stationary versions of ergodic processes.

Consider thus, as in Proposition 4.5, a stationary version $\{X_t\}_{t\in\mathbb{R}}$ of an ergodic Markov jump process with doubly infinite time (we assume for convenience that $\{X_t\}$ is nonexplosive). Define the time–reversed process $\{\widetilde{X}_t\}_{t\in\mathbb{R}}$ by $\widetilde{X}_t = X_{-t-} = \lim_{s\uparrow -t} X_s$ [the reason for not simply letting $\widetilde{X}_t = X_{-t}$ is to obtain right–continuous paths; of course, this is immaterial for the distribution of finite–dimensional sets (since the probability of a jump at t is always zero) and therefore of the whole process].

Proposition 5.2 *$\{\widetilde{X}_t\}$ is a stationary Markov jump process with $\boldsymbol{\pi}$ as stationary distribution and intensities $\widetilde{\lambda}(i,j) = \pi_j\lambda(j,i)/\pi_i$.*

Proof. By nonexplosiveness, $\{\widetilde{X}_t\}$ is a pure jump process which is time–homogeneous Markov by Proposition 5.1 (consider discrete skeletons). Thus all that remains to be shown is that the asserted expression for $\widetilde{\lambda}(i,j)$ holds for $i \neq j$. But

$$\begin{aligned}\widetilde{\lambda}(i,j) &= \lim_{h\downarrow 0}\frac{\widetilde{p}^h_{ij}}{h} = \lim_{h\downarrow 0}\frac{\mathbb{P}(\widetilde{X}_0 = j, \widetilde{X}_{-h} = i)}{h\mathbb{P}(\widetilde{X}_{-h} = i)}\\ &= \lim_{h\downarrow 0}\frac{\mathbb{P}(X_0 = j, X_h = i)}{h\mathbb{P}(X_h = i)} = \lim_{h\downarrow 0}\frac{\pi_j p^h_{ji}}{h\pi_i} = \frac{\pi_j\lambda(j,i)}{\pi_i}. \qquad \square\end{aligned}$$

Call $\{X_t\}$ *time–reversible* if $\{\widetilde{X}_t\}$ has the same distribution as $\{X_t\}$.

Corollary 5.3 *Let $\boldsymbol{\pi}$ be the ergodic distribution. Then a necessary and sufficient condition for time reversibility is $\pi_i\lambda(i,j) = \pi_j\lambda(j,i)$ for all $i \neq j$.*

The term $\pi_i\lambda(i,j)$ is the rate at which transitions $i \to j$ occur in stationarity and is often denoted as the *probability flow* from i to j. Thus the reversibility condition means that the flow from i to j is the same as the flow from j to i, and for this reason it is called the condition of *local* or *detailed balance* in contrast to the equilibrium equation $\boldsymbol{\pi\Lambda} = \mathbf{0}$ which is the condition of *full balance*. More precisely, rewriting $\boldsymbol{\pi\Lambda} = \mathbf{0}$ in the form $\pi_i\lambda(i) = \sum_{j\neq i}\pi_j\lambda(j,i)$, the l.h.s. is the total flow out of state i and the r.h.s. the total flow into state i.

Corollary 5.4 *Let $\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*$ be nonexplosive intensity matrices and $\boldsymbol{\pi}$ a distribution such that $\boldsymbol{\pi}_i\lambda(i,j) = \pi_j\lambda^*(j,i)$ for all $i,j \in E$. Then $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ are ergodic with stationary distribution $\boldsymbol{\pi}$ for both, and further $\boldsymbol{\Lambda}^*$ coincides with the intensity matrix $\widetilde{\boldsymbol{\Lambda}}$ of the time–reversed process $\{\widetilde{X}_t\}$.*

Proof. Summing $\pi_i\lambda(i,j) = \pi_j\lambda^*(j,i)$ over i and using that the row sums of $\boldsymbol{\Lambda}^*$ are zero immediately yields $\boldsymbol{\pi\Lambda} = \mathbf{0}$. Theorem 4.3 then gives ergodicity, and $\boldsymbol{\Lambda}^* = \widetilde{\boldsymbol{\Lambda}}$ then follows by Proposition 5.2. □

Problems

5.1 Show that a nonexplosive intensity matrix $\boldsymbol{\Lambda}$ is time reversible if and only if there is a function $\gamma(i,j)$ such that $\gamma(i,j) = \gamma(j,i)$ and $\lambda(i,j)/\gamma(i,j)$ is independent of j.

5.2 Show *Kolmogorov's loop criterion*: an ergodic Markov process is time reversible if and only if for each chain $i_0, i_1, \ldots, i_n$ of states with $i_0 = i_n$ it holds that $\prod_1^n \lambda(i_{k-1}, i_k) = \prod_1^n \lambda(i_k, i_{k-1})$. [*Hint*: For "if," show that local balance holds if one takes $\pi_i = \prod_1^m \lambda(j_{k-1}, j_k)$ where j_0 is arbitrary but fixed and $j_0, j_1, \ldots, j_m = i$ is an arbitrary chain of states with positive transitions rates connecting j_0 and i.]

5.3 Consider a *circular birth–death process*. That is, $E = \{\mathrm{e}^{\mathrm{i}2\pi k/n} : k = 1, \ldots, n\}$ and the transition rate from $\mathrm{e}^{\mathrm{i}2\pi k/n}$ to $\mathrm{e}^{\mathrm{i}2\pi(k+1)/n}$ is β_k and the rate from $\mathrm{e}^{\mathrm{i}2\pi k/n}$ to $\mathrm{e}^{2\pi(k-1)/n}$ is δ_k (all other transitions have rate 0). Show that the process is reversible if and only if $\prod_1^n \beta_k = \prod_1^n \delta_k$.

Notes Time reversibility is studied for example in Kelly (1979), Keilson (1979) and Serfozo (1999). Corollary 5.4 is often referred to as *Kelly's lemma* and will be used repeatedly in Chapter IV.

III
Queueing Theory at the Markovian Level

1 Generalities

1a. Queueing Theory and Some of Its Daily Life Motivations
1b. Classification of Simple Queues
1c. The Queue Discipline
1d. Queue Lengths, Waiting Times and Other Functionals
1e. Measures of Performance. The Traffic Intensity
1f. Equilibrium Theory versus Time-Dependence
1g. Queueing Theory in This Book

1a Queueing Theory and Some of Its Daily Life Motivations

Though the general field of applied probability has by now developed into a diversity of subareas, queueing theory is not only one of the oldest, but also one of the most notable and prominent. Queueing problems come up in a variety of situations in the real world and have stimulated an enormous literature which, though in part quite mathematical and abstract, is not of a purely academic nature. In fact, there has been a considerable interaction between the developments at the various degrees of abstraction in the field. Thus, though the more theoretical–orientated part of the literature (incorporating this book) tends to deal with models and problems too simplified to be of any great direct practical applicability, the notions and techniques that are studied are also important for the practical worker in the field. Conversely, the call for solutions to particular problems has of course stimulated not only the theory of queueing but also that of probability as a whole, fields like Markov processes, renewal theory and random walks owing their present state and importance to a large extent to the impact from queueing theory. Queueing problems present a great challenge to the probabilist and a *memento mori* to probability theory as a whole. The development of abstract probability theory may be of great beauty, but seldom sheds much light on how to come up with the numbers the practical worker asks for. The crux is more often a thorough understanding of the particular features of the model combined with a few basic mathematical techniques, and it is a feeling for this that the present treatment aims at bringing the reader.

Queueing situations from daily life are almost too obvious, but we shall list a few anyway: customers queueing up before the m cashiers in a supermarket; telephone callers waiting for one of the lines of an exchange to become available; aircraft circling over the airport before a runway becomes free; machines under care of a repairman who can handle only one at a time; and so on. Of more recent date than these classical examples are a number of problems connected with computer organization or networks in teletraffic theory or data transmission: in a time–sharing computer, we may think of the jobs as customers who are served by the central processor unit (CPU) and possibly input/output facilities. At each of these units queues may form, and in particular the queue at the CPU has some rather specific features (feedback, simultaneous service). In telephone networks there is a hierarchy of exchanges, so that, for example, local calls need only to pass an exchange of the lowest level, whereas long–distance calls may be directed among one of several possible paths connecting exchanges at various levels. Queues may form at the exchanges and are highly interactive.

We finally mention that a number of other situations may either directly be formulated in queueing terms or at least are closely related. Examples occur in inventory processes and insurance risk. For example, in a store with items placed from time to time and taken out as demand arises, we may think of items as customers and of the removals as service events.

1b Classification of Simple Queues

The great diversity of queueing problems gives rise to an enormous variety of models each with their specific features. Incorporating more than one or two such features usually makes the model not only complicated but also analytically intractable. Therefore a substantial part of the literature deals with models of a very simple structure.

Without attempting anything near a classification of all queueing situations, one might tentatively single out the following relevant features for the description of a queue of reasonably simple structure: (a) the *input* or *arrival process*, i.e. the way in which the customers arrive to the queue; (b) the *service facilities*, i.e. the way in which the system handles a given input stream. Logically incorporated in (b) but treated separately in Section 1c is (c) the *queue discipline*, i.e. the algorithm determining the order in which the customers are served. The descriptions of these features may be quite complicated and are, at least in their verbal form, always lengthy. A convenient shorthand notation system was suggested by D.G. Kendall in 1953 and has to a large extent become standard since then. It enables one to replace phrases such as "the single–server queue with completely random arrivals and general service times" with symbolic notation such as "$M/G/1$". The notation covers some simple and basic queueing systems (but by no means all important ones) which have the following characteristics:

(i) Customers arrive one at a time according to a renewal process in discrete or continuous time. That is, the intervals between successive arrivals of customers are i.i.d. and governed by a distribution A on $\mathbb{N}$ or $(0,\infty)$. We number the customers $0,1,2,\ldots$ and assume most often that customer 0 arrives at time 0. Thus, if T_n denotes the interval between the arrival of customers n and $n+1$, the T_n are i.i.d. governed by A and the arrival instants are $0, T_0, T_0+T_1, \ldots$.

(ii) The service times of different customers are i.i.d. and independent of the arrival process. We denote the governing distribution (concentrated on $(0,\infty)$) by B and the service time of customer n by U_n. Thus $U_0, U_1, \ldots$ are i.i.d. governed by B and independent of the T_n.

In Kendall's notation, a queueing system of this type is denoted by a string of the type $\alpha/\beta/m$, where α refers to the form of the interarrival distribution, β to the form of the service time distribution and m is the number of servers. The most common values of α, β are as follows:

M The exponential distribution. (M = Markovian. Other terms are "completely random" and "Poissonian.")

D The distribution degenerate at some point $d \in (0,\infty)$, frequently $d = 1$. (D = deterministic. Also, the term "regular" is used.)

E_k The Erlang distribution with k stages; see Section 4.

H_k The hyperexponential distribution with k parallel channels; see Section 4.

PH A more general phase–type distribution; see Section 4.

GI **or** G No restrictions on the form of the distribution. (GI = General Independent, G = General; we shall here follow the tradition to use GI when referring to the interarrival distribution and G for the service time distribution.)

Thus examples of the particular queueing models become $M/D/1$, $GI/G/\infty$, $E_k/M/1$, $M/H_k/m$ etc., with, for example, $M/D/1$ denoting the single–server queue with Poisson arrivals and deterministic service times.

The notation is widely accepted, but notice should be taken that variants and extensions abound in the literature. One variant has a different distinction between GI and G than the (usual) one given here. This is motivated from the considerable attention that has in recent years been given to queues where the independence assumptions are replaced by the sequences $\{T_n\}$ and $\{U_n\}$ being only strictly stationary. One then writes $G/G/1$ etc. and uses GI to denote the classical independent case (e.g. in $G/GI/1$ service times will then be independent but interarrival times not). Other extensions (that will not be used in the present book) are for example $M^X/D/m$ and $M/M/m/n$. The first case refers to customers arriving in batches, distributed as the random variable X, at the epochs of a Poisson process. The second may be used for an $M/M/m$ queue with a finite waiting room of size n, a finite population of n customers or even other models.

1c The Queue Discipline

We start with a list of some of the main types of queue disciplines.

FIFO First In, First Out. Also denoted FCFS = First Come, First Served. The customers are served in the order of arrival. Apparently this is the usual procedure at an ordered queue and therefore the predominant assumption in the literature. *Unless otherwise stated, this is the queue discipline in force throughout this book.*

LIFO Last in, First out. Also denoted LCFS = Last Come, First Served. After having completed a service the server turns to the latest arrived customer. This would occur, for example, in inventories where the items (customers) are stacked and all in–out operations occur at the top of the pile.

SIRO Service in Random Order. After having served a customer, the server picks the next at random among the remaining ones. This would occur, for example, in technical systems such as telephone exchanges where the system does not remember when the customers arrived.

PS Processor Sharing. The customers share the server, i.e. when n customers are present, the server devotes $1/n$ of his capacity to each. Equivalently, the customers attain service at rate $1/n$ and leave the system once the attained service reaches the service time. The situation is illustrated in Fig. 1.1. The main example is a computer with several jobs running simultaneously. Here PS is really only an approximation to what physically occurs, namely the next discipline in the list:

RR Round Robin. Here the server works on the customers one at a time in a fixed time quantum δ. A customer not having completed service within this time is put back in the queue, and before he can retain service the other customers are each allowed their quantum of δ (or less, if service is completed). The situation is illustrated in Figs. 1.2, 1.3. As δ becomes infinitely small, PS is obtained as a limiting case of RR.

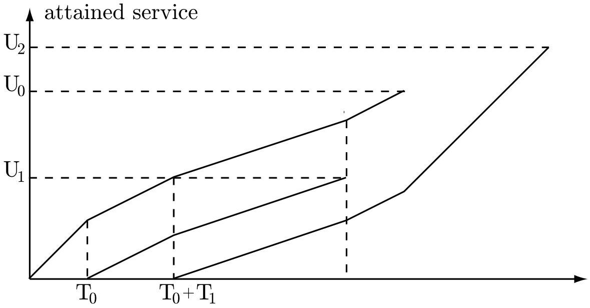

Figure 1.1

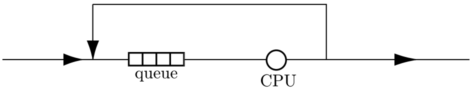

Figure 1.2

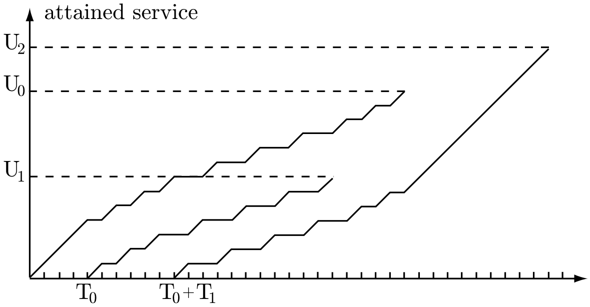

Figure 1.3

This list is by no means complete and does not cover all aspects. For example, it is not quite clear what is meant by a FIFO $GI/G/s$ queue since the customers may either queue up in one line (what we shall assume in the following) or in some way form s separate waiting lines. Further examples of queue disciplines are found above all in the area of *priority queueing*. Here the customers are divided into priority classes $1, 2, \ldots, K$, a customer from a lower class having priority before one from a higher class. The system may be *preemptive* or *nonpreemptive.* In the first case the customer being served is interrupted in his service if a new one of higher priority arrives, in the second case not. In the preemptive case an interrupted customer may then either have attained some service or not, and so on, a great number of variations being possible.

1d Queue Lengths, Waiting Times and Other Functionals

In connection with a given queueing system, a great variety of stochastic processes and functionals arise. The main ones that we shall study are the following three (defined for $GI/G/s$, but with obvious generalizations to many other models):

Q_t The *queue length* at time t (denoted X_t in the present chapter where $\{Q_t\}_{t\geq 0}$ is a Markov jump process). Also denoted as the *number in system* to stress that the customer being presently handled by the server is included.

W_n The *actual waiting time* (or just *waiting time*) of customer n, i.e. the time from arrival to the system until service starts.

V_t The *workload* in the system at time t, i.e. the total time the m servers have to work to clear the system. Thus V_t is the sum of the residual service times of customers being presently served and the customers awaiting service. In the case $m = 1$ of a single server, this is simply the time needed for the server to clear the system provided that no new customers arrive, i.e. the waiting time of a hypothetical customer arriving just after t. For this reason V_t is sometimes denoted the *virtual waiting time* at time t for $m = 1$.

The connection is illustrated in Fig. 1.4. It is simplest to visualize for $m = 1$, where the actual waiting time of customer $n+1$ is the virtual waiting time $V_{\sigma(n)-}$ just before the time $\sigma(n) = T_0 + \cdots + T_n$ of his arrival. Thus on the figure, $W_0 = W_1 = W_4 = 0$ and $W_2 > 0$, $W_3 > 0$. For other aspects, see the Problems.

There are two points worth noting when concentrating interest around these processes: (a) the processes or functionals of interest are not always of one of these three types, but have frequently very close relations; (b) it depends very much on the practical situation whether it is the queue length, the actual or the virtual waiting time or some other functionals that are of interest.

An obvious example of (a) is the *sojourn time* of customer n, i.e. the total time he spends in the system. This is the waiting time followed by the service time, i.e. $W_n + U_n$, and since W_n, U_n are clearly independent, the sojourn time distribution is a simple functional of the waiting time distribution, namely the convolution with the service time distribution B. Another example is busy and idle times, which in $GI/G/1$ can be described by the time intervals where $Q_t > 0$ (or equivalently $V_t > 0$) and $Q_t = V_t = 0$, respectively. We also mention that the interest in $\{V_t\}$ is due to a large

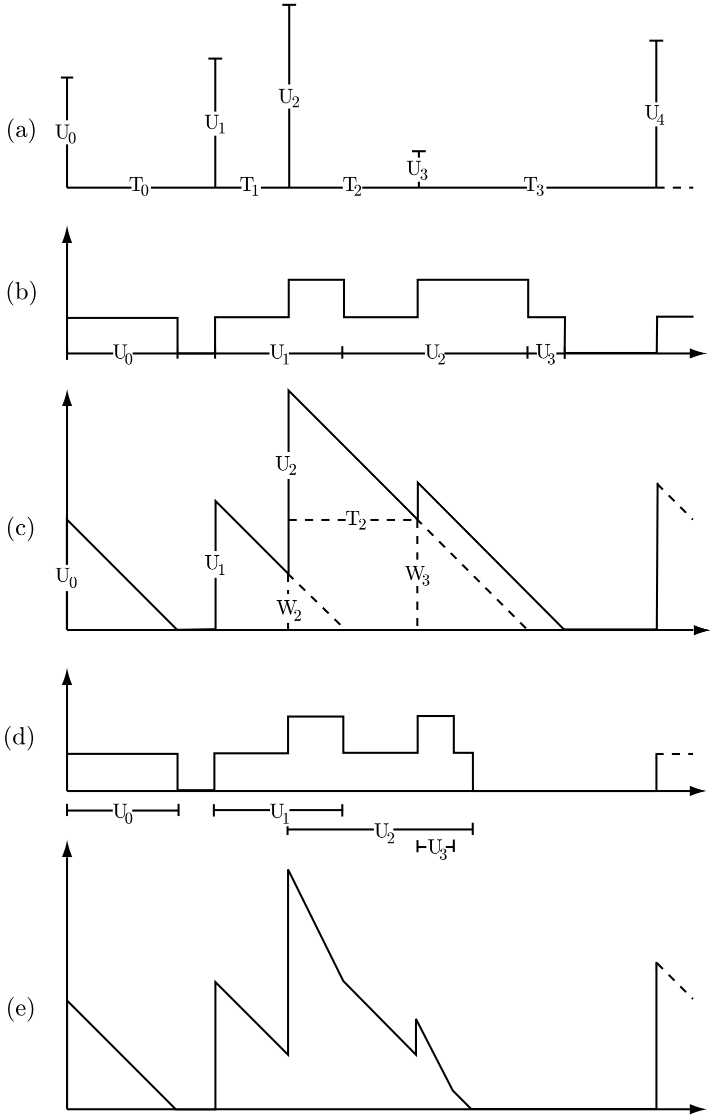

Figure 1.4

(a) Input of service times and interarrival times; (b) the corresponding single–server queue length process; (c) the single–server workload (virtual waiting time) process; (d) the two–server queue length process; (e) the two–server workload process.

extent to the reinterpretations of this process within the areas of storage, dams and insurance risk; see Chapter XIV.

For (b), note that, for example, sometimes the interest centres around the workload put on the system itself, in other cases around the inconvenience caused to the customers by exceedingly long waiting times (more typically, the aim is to balance these points of view). A typical example would be the design problem for the cash system in a supermarket: say for simplicity that we have m identical servers and want to choose the best value of m. If m is large, we expect the system to be idle for a considerable amount of time and thereby be insufficiently utilized compared to the cost of running. If, on the other hand, m is small, then we expect long waiting times for the customers, which will encourage them to use instead a less congested competing shop nearby. The quantitative evaluation of this effect of discouragement is of course a matter of management judgement and not mathematics. However, once this has been settled we need to say something about both idle times and waiting times for a given arrival rate. Possibly the discouragement could be an effect of the visible queue length and not the related but unobservable waiting time. Therefore, the queue length is also of potential interest here. It is certainly so in other situations such as telephone exchanges with a limited number K of lines, where queue lengths $\geq K$ mean the possibility of calls being lost.

1e Measures of Performance. The Traffic Intensity

Seen from a practical point of view, the purpose of theoretical analysis is to shed some light on the queueing situation in question. The meaning of this may be rather vague and, for example, it may be argued that just formulating a simplified mathematical model is helpful since it necessitates thinking through and properly clarifying which features of the system are the basic ones. Having passed this point, however, interest centers on evaluating the performance of a given system (and possibly some related ones, for the purpose of assessing the effect of a change). That is, the first step is to define some appropriate measures of performance.

In rather general terms, we want to describe the properties of the basic processes of queue length and waiting times. A main step in that direction is the study of one–dimensional distributions, say for example $\mathbb{P}(W_n \leq t)$. Now this is difficult to compute in most situations and the dependence on t is a complicating factor for the sake of comparisons (so is the dependence on n, but we defer the discussion of this to Section 1f). Therefore, it becomes appropriate to consider some simple characteristics of a distribution F on $[0, \infty)$, and some of the main ones that are usually considered relevant are the following:

(i) The *mean* $\mu = \int_0^\infty x\, F(\mathrm{d}x)$, measuring the average values.

(ii) Possibly some of the *higher order moments* $\mu^{(k)} = \int_0^\infty x^k\, F(\mathrm{d}x)$.

(iii) The *variance* $\sigma^2 = \mu^{(2)} - \mu^2$, measuring the dispersion around the mean; possibly also some higher order *cumulants*.

(iv) The squared *coefficient of variation* σ^2/μ^2, giving a scale invariant measure of dispersion.

(v) The *tail characteristics* describing the asymptotic behaviour of the tail $\overline{F}(x) = 1 - F(x)$. For example, the relation $\overline{F}(x) \sim C\mathrm{e}^{-\gamma x}$ holds for many distributions in queueing theory (cf. XIII.5), and appropriate tail characteristics are then C and (in particular) γ.

Which of these characteristics is appropriate depends on the situation.

There is, however, one measure of performance of a queueing system that is of universal interest. This is the so–called *traffic intensity* ρ, which we define here for $GI/G/m$ by

$$\rho = \frac{\mathbb{E}U_k}{m\mathbb{E}T_k} = \frac{\int_0^\infty x\,B(\mathrm{d}x)}{m\int_0^\infty x\,A(\mathrm{d}x)} \tag{1.1}$$

(there are appropriate generalizations for most other queueing systems) and the interpretation is as follows. Suppose that for a very large amount of time t the system is working at full capacity, i.e. that all servers are busy. Then by the LLN there will be about $t/\mathbb{E}T_k$ arrivals and a total of about $mt/\mathbb{E}U_k$ services ($t/\mathbb{E}U_k$ for each server). Thus ρ is about the ratio, i.e. when $\rho > 1$ the number of arrivals exceeds the number of services so that we expect the queue to grow indefinitely. In contrast, when $\rho < 1$ then eventually even a very long initial queue will be cleared (in the sense that not all servers are busy; after that the queue may build up again, but will again be cleared up for the same reason, and so on, the system evolving in cycles). Thus the behaviour should be like transience when $\rho > 1$, and like recurrence when $\rho < 1$. This will be made more precise later in the various models. Also, results will be shown stating that the behaviour for $\rho = 1$ is like null recurrence.

1f Steady State Theory versus Time–Dependence

The notion of *steady state* is within setting of Markov processes just what so far has been called stationarity: a Markov chain or Markov jump process is in steady state if it is ergodic and stationary (another common term is *equilibrium*, inspired from statistical mechanics).

The results developed in Chapters I and II state that after a long period of time an ergodic process attains the steady state (settles in equilibrium). A similar behaviour is, on intuitive grounds, to be expected far beyond the Markovian setting: if the capacity of the queueing system is sufficient to deal with the arriving workload, say the traffic intensity is less than 1, one expects the system to alternate between being busy and idle, and that the initial conditions will be smoothed away by the stochastic variation in the length of the cycles. Thus, under appropriate conditions there should exist limiting distributions of Q_t, V_t as $t \to \infty$ and of W_n as $n \to \infty$, and there is then an apparent possibility of studying the characteristics of the queueing system by means of these limiting distributions. More generally, when studying functionals of the whole process such as departure processes, one could restrict attention to a stationary, i.e. steady–state, version. This will be represented either by a governing probability distribution $\mathbb{P}_{\rm e}$ (e for equilibrium) or by notation like W to denote a random variable having the limiting steady–state waiting–time distribution). Thus, for an ergodic Markov chain $\{X_n\}$ with stationary distribution $\boldsymbol{\pi}$, we have

$$\pi_j \;=\; \mathbb{P}_{\rm e}(X_n = j) \;=\; \mathbb{P}(X = j) \;=\; \lim_{n\to\infty} \mathbb{P}_i(X_n = j),$$

and $\mathbb{P}_{\rm e}$ is the same as the $\mathbb{P}_{\boldsymbol{\pi}}$ of Chapter I.

The idea of passing from the study of say $\mathbb{P}(W_n \le x)$, $n = 0, 1, 2, \ldots$, to $\mathbb{P}_{\rm e}(W_n \le x) = \mathbb{P}(W \le x)$ is clearly convenient, if nothing else, by eliminating the dependence on n. The motivations are in fact deeper than just this, with the two following points as the cornerstones: (1) a queueing system will frequently be operating for such long periods of time that the steady state is entered rather early in that period; (2) in addition to its limiting interpretation, $\mathbb{P}_{\rm e}$ also describes the long–term behaviour in terms of time averages. For example, one has for a Markovian queue in continuous time that subject to suitable conditions $T^{-1}\int_0^T Q_t\,{\rm d}t \approx \mathbb{E}_{\rm e}Q$ (cf. II.4), and this average is frequently an appropriate characteristic of the whole segment $\{Q_t\}_{0\le t\le T}$.

The overwhelming majority of queueing theory (and also the material presented in this book) is concerned with the steady–state properties of the systems rather than finding *time–dependent* quantities like p_{ij}^n in a Markov chain (instead of time–dependent, frequently the somewhat unfortunate term "transient" is used). The reasons for this are most often motivated by (1) and (2) above. However, without any doubt the fact that time–dependent solutions are exceedingly more difficult to come by than the steady–state ones also plays an important role in practice. Thus, it seems clear that in many situations it is not clear a priori what a long time period in (1) means. Hence it is necessary to have at least some estimate on the rate of convergence to the steady state, i.e. some ideas on the time–dependent behaviour. Also, it is clear that in other situations such as the presence of a rushhour where the queue suddenly builds up after having behaved stably, the steady–state point of view is not adequate at all.

1g Queueing Theory in This Book

A particular practical problem will usually exhibit a considerable number of the great variety of aspects presented so far, and most likely some further specific ones. Comprehensive mathematical models will therefore tend to be complicated and usually intractable: possibly the existence of a limiting steady state can be proved, but the derivation of its properties in a form suitable for numerical calculations is usually out of the question. Therefore the practitioner may have to use either empirical or semi–empirical methods such as simulation, approximations or bounds, or to trust that solutions of greatly simplified models have something to say about his problem as well.

It is not our aim here to present queueing theory in a form ready for practical implementation, but rather to study some of the basic mathematical problems and techniques. In this and the next chapter, a rather broad class of problems are studied within the Markovian setting, and after having developed the necessary mathematical tools in Part B, we then deal with a more narrow class of problems in Part C, assuming either general distributions of interarrival times and service times or, as is in fact better motivated, Markovdependence.

The Markovian assumptions greatly simplify the modelling and solution. They are therefore also frequently the first step when faced with a new type of problem, and they will be used here to look into phenomena requiring considerable effort in more general settings. Examples are queueing networks, time dependence, the busy–period distribution, the effect of queue disciplines other than the FIFO one and also some finite models (clearly, many important models are not touched upon at all). The Markovian set–up has its drawbacks, however. One is that queue lengths as discrete variables are more naturally incorporated than the continuous waiting and sojourn times. For example, in a network we can study the length of the various waiting lines, but not the presumably more interesting total sojourn time of a customer. Another deficit is the reliance on assumptions such as Poisson arrivals and (probably more seriously) exponential service times. The phase method (to be developed in Section 4) presents a partial solution by extending the Markovian set–up to a class of models that is in a certain sense dense.

Finally, we mention that one of the classical topics in Markovian queueing theory, imbedded Markov chains, has been deferred to X.5. A Markov chain is imbedded in a (typically non–Markovian) queue if it is obtained by observing the queue length at certain random times. Main examples are $M/G/1$ just after departure times and $GI/M/1$ just before arrival times. However, in $M/G/1$ in particular, the imbedded Markov chain is only of limited intrinsic interest, and it requires the more advanced tools of Part B to relate it to the queue length in continuous time and the waiting times.

Problems

1.1 Consider the LIFO single–server queue. Show that the waiting times corresponding to the input in Fig. 1.4(a) are the same as for the FIFO case, and draw a different figure where this is not the case.

1.2 Draw a figure of the PS single–server queue length process corresponding to the input in Fig. 1.4(a). Find (graphically) the sojourn times of the customers.
1.3 Find (graphically) the waiting times of the customers in the $GI/G/2$ case of Fig. 1.4(a).
1.4 Consider $GI/G/1$ with $\rho < 1$. Show heuristically that the server is idle or busy in an average proportion ρ, resp. $1-\rho$, of the time.
1.5 Consider $GI/G/1$ with $\rho < 1$. Show by heuristic time–average considerations *Little's formula* $\ell = \lambda w$. Here ℓ, w denote the steady–state mean of the queue length and the sojourn time, and $\lambda = 1/\mathbb{E}T_k$ the average arrival rate. [*Hint:* Evaluate $\int_0^T Q_t \, \mathrm{d}t$ in terms of the sojourn times of the customers that arrived in $[0,T]$, neglecting boundary effects; a formal proof is in X.4.] Is the FIFO assumption essential?

Notes Queueing theory as a whole is an enormous area. Most of the standard textbooks are listed in the Bibliography, and for the current development of research in the area, some of the main journals to consult are the *Advances in Applied Probability*, *Annals of Applied Probability*, *Journal of Applied Probability*, *Mathematics of Operations Research*, *Operations Research*, *Probability in the Engineering and Information Sciences*, *Queueing Systems*, *Stochastic Models* and *Stochastic Processes and Their Applications.*

Obviously, it is not possible to cover all special models in a single book. Of topics not treated, we mention in particular polling systems (a single server switches between several queues), fork–join or split–and–match queues (applied in manufacturing), negative customers (see Chao *et al*, 1999, for references), join–the–shortest–queue disciplines (e.g. Foley and McDonald, 2001), retrial queues (Falin and Templeton, 1997; Artalejo, 1999) and queues with vacations (the server is temporalily unavailable). All of these areas are currently active, and the reader interested in one or more is advised to perform a database search for references.

2 General Birth–Death Processes

By a *birth–death process* we understand a Markov jump process $\{X_t\}_{t\geq 0}$ on $E = \mathbb{N}$ which is *skip–free*, i.e. from state n it can only move to $n-1$ or $n+1$ (from 0 even only to 1). That is, the intensity matrix is of the form

$$\boldsymbol{\Lambda} = \begin{pmatrix} -\beta_0 & \beta_0 & 0 & 0 & \dots \\ \delta_1 & -\beta_1-\delta_1 & \beta_1 & 0 & \dots \\ 0 & \delta_2 & -\beta_2-\delta_2 & \beta_2 & \dots \\ \vdots & & & \ddots & \end{pmatrix}.$$

We denote the β_n as *birth intensities* and the δ_n as *death intensities.* In this terminology, one thinks of the process as the total size of a population and the most well–known example is the *linear birth–death process* $\beta_n = n\beta$, $\delta_n = n\delta$ which corresponds to the individuals giving birth and dying independently of one another, with rates independent of the population size. In our applications we interpret instead X_t as the number of customers in a queue at time t: a jump upward corresponds to a customer arriving at the queue and a jump downward to a customer having completed service and leaving the system. Thus in this generality the arrival rate β_n and the service rate δ_n depend in an unspecified manner on the number n of customers present. For example β_n could be decreasing in n, corresponding to customers being discouraged by long queues, (*balking* or *reneging*) and δ_n increasing, corresponding to the server working more rapidly when faced with a long queue. However, the main interest in birth–death processes is due to the more concrete interpretation of the models associated with the specific choices of β_n, δ_n to be presented in Sections 3a–3g. We proceed here to develop the general theory.

The jump chain $\{Y_n\}$ is clearly skip–free as well and may be viewed as a state–dependent Bernoulli random walk (i.e. the increments are ±1), with reflection at zero. The transition matrix is

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ q_1 & 0 & p_1 & 0 & \dots \\ 0 & q_2 & 0 & p_2 & \dots \\ \vdots & & \ddots & \vdots & \end{pmatrix}$$

where $p_n = \beta_n/(\beta_n+\delta_n)$, $q_n = 1-p_n = \delta_n/(\beta_n+\delta_n)$. We assume for a while that no p_n can take the values 0 or 1. This obviously implies irreducibility.

Proposition 2.1 *Recurrence of* $\{X_t\}_{t\geq 0}$ *or equivalently* $\{Y_n\}$ *is equivalent to*

$$\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\beta_1 \cdots \beta_n} = \sum_{n=1}^{\infty} \frac{q_1 \cdots q_n}{p_1 \cdots p_n} = \infty. \tag{2.1}$$

Proof. We apply the transience criterion I.5.2 with $i = 0$ to $\{Y_n\}$ and have to look for $h(k)$, $k \geq 1$, satisfying $h(j) = \sum_{k\neq 0} q_{jk}h(k)$, $j \neq 0$, i.e.

$$\begin{aligned} h(1) &= p_1 h(2), \\ h(2) &= q_2 h(1) + p_2 h(3), \\ &\vdots \\ h(n) &= q_n h(n-1) + p_n h(n+1), \\ &\vdots \end{aligned}$$

If on the l.h.s. we write $h(n) = (p_n + q_n)h(n)$ and solve for $h(n) - h(n-1)$, we get

$$\begin{aligned} h(2) - h(1) &= q_1 h(1)/p_1, \\ h(n+1) - h(n) &= \frac{q_n}{p_n}\bigl(h(n) - h(n-1)\bigr) = \cdots \\ &= \frac{q_n q_{n-1} \cdots q_2}{p_n p_{n-1} \cdots p_2}\bigl(h(2) - h(1)\bigr) = \frac{q_n \cdots q_1}{p_n \cdots p_1} h(1), \end{aligned}$$

and it is clear that there is one, and up to proportionality only one, nonzero solution that is bounded if and only if

$$\sup_n h(n) \;=\; h(1) + \sum_{n=1}^{\infty} \bigl[h(n+1) - h(n)\bigr] \;=\; h(1)\Bigl\{1 + \sum_{n=1}^{\infty} \frac{q_n \cdots q_1}{p_n \cdots p_1}\Bigr\}$$

is finite. Thus transience is equivalent to (2.1) to fail. □

The criterion (2.1) states loosely that the q_n in some average sense should be as large as the p_n, i.e. that there is no drift to infinity. Assume, for example, some smooth behaviour such as the existence of $\sigma = \lim p_n/q_n$. Then if $\sigma < 1$, (2.1) is infinite and we have recurrence, whereas (2.1) is finite for $\sigma > 1$ and we have transience (for $\sigma = 1$ both possibilities may occur, cf. Problem 2.2).

Proposition 2.2 *A birth–death process is nonexplosive if and only if $R < \infty$ where*

$$R \;=\; \sum_{n=1}^{\infty} r_n, \qquad r_n \;=\; \sum_{k=0}^{n} \frac{\delta_{k+1} \cdots \delta_n}{\beta_k \cdots \beta_n}\,.$$

Proof. We apply Reuter's condition II.3.3, which states that the process is nonexplosive if and only if any nonnegative solution $\boldsymbol{x} = (x_n)_{n\in\mathbb{N}}$ to $\boldsymbol{\Lambda x} = \boldsymbol{x}$ is trivial, $\boldsymbol{x} = \boldsymbol{0}$. Equations 0 and $n \geq 1$ of $\boldsymbol{\Lambda x} = \boldsymbol{x}$ are

$$-\beta_0 x_0 + \beta_0 x_1 \;=\; x_0, \quad \delta_n x_{n-1} - (\beta_n + \delta_n)x_n + \beta_n x_{n+1} \;=\; x_n,$$

which, letting $\Delta_n = x_n - x_{n-1}$, $f_n = 1/\beta_n$, $g_n = \delta_n/\beta_n$, can be rewritten as

$$\Delta_1 \;=\; f_0 x_0, \quad \Delta_{n+1} \;=\; f_n x_n + g_n \Delta_n.$$

This shows that the solution with $x_0 = 0$ is $\boldsymbol{x} = \boldsymbol{0}$ and that the solution with $x_0 > 0$ (say $x_0 = 1$) is strictly increasing. Iterating and noting that $r_n = \sum_0^n f_k g_{k+1} \cdots g_n$ yields

$$\Delta_{n+1} \;=\; \sum_{k=0}^{n} f_k g_{k+1} \cdots g_n x_k \left\{ \begin{array}{ll} \geq & r_n x_0 \\ \leq & r_n x_n \end{array} \right. .$$

Summing and using the lower bound shows that $R < \infty$ is necessary for $\boldsymbol{x}$ to be bounded. The upper bound yields

$$x_{n+1} \;\leq\; (1 + r_n)x_n \;\leq\; \cdots \;\leq\; \prod_{k=0}^{n} (1 + r_k),$$

so that if conversely $R < \infty$ and hence $\prod_0^\infty (1 + r_k) < \infty$, then $\boldsymbol{x}$ is bounded. □

Lemma 2.3 *Irrespective of recurrence or transience, there is one, and up to proportionality only one, solution* $\boldsymbol{\nu} = (\nu_n)_{n\in\mathbb{N}}$ *to* $\boldsymbol{\nu\Lambda} = \mathbf{0}$*, given by*

$$\nu_n \;=\; \frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n}\nu_0. \tag{2.2}$$

Proof. That $\boldsymbol{\nu\Lambda} = \mathbf{0}$ means

$$\beta_0\nu_0 = \delta_1\nu_1, \quad (\beta_n+\delta_n)\nu_n = \beta_{n-1}\nu_{n-1} + \delta_{n+1}\nu_{n+1},\ n \ge 1.$$

It is clear that given ν_0, these equations uniquely determine $\boldsymbol{\nu}$, and insertion shows that (2.2) is indeed a solution. □

Corollary 2.4 *In the recurrent case, the stationary measure* $\boldsymbol{\mu} = (\mu_n)_{n\in\mathbb{N}}$ *for* $\{Y_n\}$ *is given by*

$$\mu_n \;=\; \frac{p_1\cdots p_{n-1}}{q_1\cdots q_n}\mu_0, \quad n = 1,2\cdots. \tag{2.3}$$

Proof. Take $\boldsymbol{\mu}$ as in II.4.2(ii),(iii), $\mu_n = \nu_n\lambda(n)$. Then $\mu_0 = \nu_0\beta_0$ and for $n = 1,2,\ldots$

$$\begin{aligned}\mu_n \;&=\; \nu_n\lambda(n) \;=\; \frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n}(\beta_n+\delta_n)\nu_0\\ &=\; \frac{p_1\cdots p_{n-1}}{q_1\cdots q_n}\frac{\beta_0 q_n}{\delta_n}(\beta_n+\delta_n)\nu_0 \;=\; \frac{p_1\cdots p_{n-1}}{q_1\cdots q_n}\mu_0.\end{aligned}$$

□

Now define

$$S \;=\; 1 + \sum_{n=1}^{\infty}\frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n}\,.$$

Corollary 2.5 $\{X_t\}_{t\ge 0}$ *is ergodic if and only if* (2.1) *holds and* $S < \infty$*, in which case the ergodic distribution* $\boldsymbol{\pi} = (\pi_n)_{n\in\mathbb{N}}$ *is given by*

$$\pi_0 = \frac{1}{S}, \quad \pi_n = \frac{1}{S}\frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n}, \quad n = 1,2,\ldots. \tag{2.4}$$

Proof. Recurrence is equivalent to (2.1), and in that case the total mass of (2.2) is $|\boldsymbol{\nu}| = S\nu_0$ so that according to II.4.3 ergodicity is equivalent to $S < \infty$. In that case, $\boldsymbol{\pi} = \boldsymbol{\nu}/|\boldsymbol{\nu}|$. □

We conclude with some formulas for the case of a finite state space $\{0,\ldots,K\}$. This occurs if $\beta_K = 0$ since then $\{0,\ldots,K\}$ is a closed set. Irreducibility and hence ergodicity will hold if

$$\beta_0 > 0,\ldots,\beta_{K-1} > 0,\ \beta_K = 0,\ \delta_1 > 0,\ldots,\delta_K > 0, \tag{2.5}$$

and in just the same manner as in Lemma 2.3 and Corollary 2.5 one obtains the stationary distribution as

$$\pi_0 = \frac{1}{S}, \quad \pi_n = \frac{1}{S}\frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n}, \quad n = 1,\ldots,K \tag{2.6}$$

where $S = 1 + \sum_1^K (\beta_0 \cdots \beta_{n-1})/(\delta_1 \cdots \delta_n)$.

Remark 2.6 In many examples, the finite case arises as a modification of an infinite model by letting some $\beta_K = 0$. If the stationary distributions are $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^{(K)}$, respectively, it is seen from (2.6) that $\boldsymbol{\pi}^{(K)}$ is simply obtained by conditioning (or truncation) of $\boldsymbol{\pi}$ to $\{0, \ldots, K\}$, $\pi_n^{(K)} = \pi_n/(\pi_0 + \cdots + \pi_K)$, $n \le K$. Compare also I.3.9 (or rather the continuous–time analogue), and see further XIV.3. □

Problems

2.1 Show that recurrence holds if $\beta_n \le \delta_n$ for all large n.
2.2 Suppose $\beta_n = 1$, $\delta_n = (1-1/2n)^\gamma$ where $\gamma \ge 0$. Show that there is transience for $\gamma > 2$ and null recurrence for $\gamma \le 2$.
2.3 Suppose $\beta_n = 1$, $\delta_n = (1+1/n)^\gamma$ where $\gamma \ge 0$. Show that there is ergodicity for $\gamma > 1$ and null recurrence for $\gamma \le 1$.
2.4 Show that there is transience if $\delta_n = 1$ for all n, $\beta_{2^k} = k$, $\beta_n = 1$ for all other n.
2.5 Consider for $k = 0, 1, 2 \ldots$ birth–death processes with $\beta_n^{(0)} = 1$, $\delta_n^{(0)} = 2$ and, for $k \ge 1$, $\delta_n^{(k)} = 2$, $\beta_k^{(k)} = 2^k$, all other $\beta_n^{(k)} = 1$. Show that there is ergodicity for $k \ge 0$, that $\beta_n^{(k)} \to \beta^{(0)}$, $\delta_n^{(k)} \to \delta^{(0)}$ as $k \to \infty$ but that $\boldsymbol{\pi}^{(k)} \to \boldsymbol{\pi}$ fails.
2.6 Let $\boldsymbol{\pi}$ be a distribution on $\mathbb{N}$ satisfying $\pi_n > 0$ for all n. Show that there exists an ergodic birth–death process with $\boldsymbol{\pi}$ as stationary distribution. Are the β_n, δ_n unique? Are they unique up to proportionality?

Notes The more refined theory of birth–death processes owes much to a series of papers by Karlin and McGregor in the 1950s; see Anderson (1991). For further examples of results beyond the present (standard) ones, see Keilson (1979), van Doorn (1980) and Ball and Stefanov (2001).

3 Birth–Death Processes as Queueing Models

3a. The $M/M/1$ Queue
3b. The $M/M/\infty$ Queue
3c. The $M/M/m$ Queue
3d. The $M/M/1$ Queue with Finite Waiting Room
3e. Erlang's Loss System
3f. Engseth's Loss System
3g. Palm's Machine Repair Problem

3a The $M/M/1$ Queue

The $M/M/1$ queue length process as defined in Section 1d clearly corresponds to a birth–death process with $\beta_n = \beta$ and $\delta_n = \delta$ independent of n. This is by far the conceptually most simple queueing system, the one of the greatest analytical tractability (at least for an infinite state space), and therefore it plays a prominent role in the literature.

The traffic intensity as defined in Section 1e is $\rho = \beta/\delta$. Thus the recurrence criterion (2.1) becomes $\sum_0^\infty \rho^{-n} = \infty$ and we have at once:

Proposition 3.1 *The $M/M/1$ queue with traffic intensity ρ is recurrent if and only if $\rho \leq 1$.*

This is intuitively reasonable, at least if $\rho \leq 1$, by recalling the interpretation of ρ as the ratio (1.1), and will be seen to hold for more general queues (e.g. $GI/G/m$). Similary, the ergodicity conditions come out immediately from Corollary 2.5. We get $S = \sum_0^\infty \rho^{-n} = (1-\rho)^{-1}$ for $\rho \leq 1$ and thus:

Proposition 3.2 *The $M/M/1$ queue with traffic intensity ρ is ergodic if and only if $\rho < 1$. In that case, the steady state distribution $\boldsymbol{\pi}$ of the queue length is geometric, $\pi_n = \mathbb{P}_e(X_t = n) = (1-\rho)\rho^n$, $n = 0, 1, 2, \ldots$.*

This permits us immediately to calculate a number of interesting quantities. For example, the probability that the server is idle or busy in steady state is

$$\mathbb{P}_e(X_t = 0) \;=\; \pi_0 \;=\; 1-\rho, \quad \text{resp. } \mathbb{P}_e(X_t > 0) \;=\; 1-\pi_0 \;=\; \rho, \tag{3.1}$$

whereas by standard formulas for the geometric distribution we have

$$\mathbb{E}_e X_t \;=\; \frac{\rho}{1-\rho}, \quad \mathbb{V}ar_e X_t \;=\; \frac{\rho}{(1-\rho)^2}, \quad \mathbb{P}_e(X_t \geq N) \;=\; \rho^N. \tag{3.2}$$

These formulas show among other things that as $\rho \uparrow 1$, then (not unexpectedly) with high probability ρ the server is busy and the mean queue length $\rho/(1-\rho)$ is large. Again, these properties are qualitatively (but not quantitatively) typical of more general queues, cf. X.7.

3b The $M/M/\infty$ Queue

This corresponds clearly to the case $\beta_n = \beta$, $\delta_n = n\delta$. We may think of each customer being handled by his own server so that his sojourn time in the system is exponential with intensity δ and independent of all other customers. A different interpretation is therefore an *immigration–death process* with immigration according to a Poisson process and each individual dying after an exponential time.

The definition (1.1) of the traffic intensity yields $\rho = 0$. Instead, the interesting parameter is $\eta = \beta/\delta$ and we get

$$\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\beta_1 \cdots \beta_n} \;=\; \sum_{n=1}^{\infty} n! \frac{1}{\eta^n} = \infty, \quad S \;=\; 1 + \sum_{n=1}^{\infty} \frac{\eta^n}{n!} \;=\; \mathrm{e}^{\eta}.$$

Thus Corollary 2.4 yields:

Proposition 3.3 *The $M/M/\infty$ queue is ergodic for all values of η. The steady state distribution $\boldsymbol{\pi}$ is Poisson with mean η, $\pi_n = \mathrm{e}^{-\eta}\eta^n/n!$.*

Notes For more advanced aspects of $M/M/\infty$, see Robert (2000) and Preater (2002). A different case of a Poisson $\boldsymbol{\pi}$ is in Problem 3.1.

3c The M/M/m Queue

Here $\beta_n = \beta$ and $\delta_n = m(n)\delta$, where $m(n)$ is the number of busy servers in state n, i.e. $m(n) = m \wedge n$. The traffic intensity is $\rho = \beta/m\delta$ and we have $\beta_n/\delta_n = \rho$, $n \geq m$. Thus, as in the case $m = 1$, (2.1) and recurrence hold if and only if $\sum \rho^{-n} = \infty$, i.e. $\rho \leq 1$. Similarly, with $\eta = \beta/\delta$

$$S \;=\; 1+\sum_{n=1}^{\infty}\frac{\beta_0\cdots\beta_{n-1}}{\delta_1\cdots\delta_n} \;=\; \sum_{n=0}^{m-1}\frac{\eta^n}{n!}+\frac{\eta^m}{m!}\sum_{n=0}^{\infty}\rho^n \;=\; \sum_{n=0}^{m-1}\frac{\eta^n}{n!}+\frac{\eta^m}{m!}(1-\rho)^{-1}$$

is finite if and only if $\rho < 1$, and we get

Proposition 3.4 *The $M/M/m$ queue with traffic intensity ρ is ergodic if and only if $\rho < 1$. In that case the ergodic distribution $\boldsymbol{\pi}$ is given by*

$$\pi_n \;=\; \frac{1}{S}\frac{\eta^n}{n!},\; n = 0,\ldots,m, \quad \pi_n \;=\; \frac{1}{S}\frac{\eta^m}{m!}\rho^{n-m},\; n = m, m+1,\ldots.$$

This solution is analytically slightly more complicated than those encountered so far since the functional form of π_n is not the same for $n < m$ and $n > m$, and also S is more complicated. The probabilistic interpretation is, however, quite interesting: $\boldsymbol{\pi}$ is a combination of the $M/M/\infty$ solution and the $M/M/1$ solution, with the $M/M/\infty$ solution on the states $\{0,\ldots,m\}$ with full server availability (no customers awaiting service) and the $M/M/1$ solution on the states $\{m, m+1,\ldots\}$ where some customers must await service.

Again it is straightforward to evaluate functionals. For example, the probability that all servers are busy and the mean queue length are

$$\mathbb{P}_e(X_t \geq m) \;=\; \pi_m + \pi_{m+1} + \cdots \;=\; \frac{1}{S}\frac{\eta^m}{m!}\frac{1}{1-\rho}, \quad \text{resp.}$$

$$\mathbb{E}_e X_t \;=\; \sum_{n=0}^{\infty} n\pi_n \;=\; \frac{1}{S}\left\{\sum_{n=1}^{m-1}\frac{\eta^n}{(n-1)!}+\frac{\eta^m}{m!}\left[\frac{\rho}{(1-\rho)^2}+\frac{m}{1-\rho}\right]\right\}.$$

3d The M/M/1 Queue with Finite Waiting Room

So far we have had the infinite state space $\{0, 1, 2, \ldots\}$ in all examples. However, clearly in many practical situations there is a limited capacity of the system so that the queue may not be arbitrarily long, and examples of this will now be given here and in the following subsections.

A simple basic case is the $M/M/1$ queue with waiting room of size K. That is, at most K customers at a time can be present in the system (including the one being served) and customers arriving to a full system are lost. Thus $\beta_n = \beta$, $n < K$, $\beta_K = 0$, $\delta_n = \delta$, $n = 1, \ldots, K$. Referring to Remark 2.6, we get for $\rho = \beta/\delta < 1$ the stationary distribution by conditioning (or *truncation*) of the geometric $M/M/1$ solution,

$$\pi_n \;=\; \frac{\rho^n}{1+\rho+\cdots+\rho^K} \;=\; \frac{1-\rho}{1-\rho^{K+1}}\rho^n, \quad n = 0, \ldots, K.$$

It can be immediately checked from (2.6) that this also holds for $\rho > 1$, whereas all $\pi_n = (1+K)^{-1}$ when $\rho = 1$.

3e Erlang's Loss System

A well–known and historically important example was considered by Erlang in connection with design problems for telephone exchanges. Suppose we have an exchange of K lines, that calls arrive at rate β and have exponential durations with rate δ, and that calls arriving while all lines are busy are lost. Let $\eta = \beta/\delta$. What is (in steady state) $E_K(\eta)$, the fraction of calls that are lost?

To solve this problem, we may model the number of busy lines as a birth–death process on $\{0, \ldots, K\}$ with $\beta_0 = \cdots = \beta_{K-1} = \beta$, $\delta_k = k\delta$, $k = 1, \ldots, K$. This corresponds to letting $\beta_K = 0$ in a $M/M/\infty$ (or $M/M/K$) queue so that by Remark 2.6 the stationary distribution is conditional Poisson,

$$\pi_n \;=\; \frac{\eta^n/n!}{1+\eta+\cdots+\eta^K/K!}, \quad n = 0, \ldots, K.$$

The probability of a particular call being lost in equilibrium is now simply the probability π_K of arriving at a full system so that

$$E_K(\eta) \;=\; \frac{\eta^K/K!}{1+\eta+\cdots+\eta^K/K!}\,. \tag{3.3}$$

This is the well–known *Erlang's loss formula* (also referred to as *Erlang's first formula* or *Erlang's B–formula*) and of considerable interest in teletraffic theory. The formula is *insensitive* to the distribution of the duration of calls, i.e. it holds also in a $M/G/1$ setting with δ^{-1} replaced by the mean duration of a call, see IV.3.

3f Engseth's Loss System

All examples considered so far have Poisson arrivals, i.e. $\beta_n = \beta$. This is adequate if we have a finite but large population of customers. Here "large" also means large compared to the sizes of the queues building up, so that even with queues of rather unlikely lengths the proportion of customers in the system is vanishing, i.e. the intensity of the source does not decrease significantly. Clearly this is not the case in all practical situations, and here and in Section 3g we shall consider two of the models that have been suggested in specific situations.

The first example is a teletraffic model considered by Engseth, essentially by just modifying Erlang's loss system to a finite population of N subscribers. Let K be the total number of lines and X_t the number of busy lines at time t. Any call is assumed to involve only one subscriber. Assuming that $N > K$, we then have a birth–death process on $\{0,\dots,K\}$ with $\beta_n = (N-n)\beta$, $\delta_n = n\delta$. Thus $\beta_0 \cdots \beta_{n-1} = N^{(n)}\beta^n$ (descending factorial), $\delta_1 \cdots \delta_n = n!\delta^n$, and letting $\eta = \beta/\delta$ we obtain the stationary distribution as

$$\pi_n \;=\; \frac{\binom{N}{n}\eta^n}{1 + N\eta + \cdots + \binom{N}{K}\eta^K}, \quad n = 0,\dots,K.$$

Choosing $p \in (0,1)$ such that $p/(1-p) = \eta$ (i.e. $p = \eta/(1+\eta) = \beta/(\beta+\delta)$), we see that this is the binomial distribution with parameters (N,p) conditioned to be in $\{0,\dots,K\}$, i.e. a *truncated binomial* or *Engseth distribution.*

3g Palm's Machine Repair Problem

Consider a population of K machines that each break down with intensity β and is immediately taken care of by one of N repairmen working at rate δ, as soon as one becomes available. Thus if X_t is the number of machines under repair or awaiting repair, we have

$$\beta_n = (K-n)\beta,\ n = 0,\dots,K-1, \quad \delta_n = (N \wedge n)\delta,\ n = 1,\dots,K.$$

One might wish to study the way in which the production loss due to stoppages and repairs depend on N, for the purpose of allocating the optimal number N of servers. To this end we need the wage expenses per unit time which are N times a known constant, and the average number of stopped machines per unit time, i.e. the equilibrium mean $\sum_0^K n\pi_n$. Obviously $\boldsymbol{\pi}$ can be immediately computed by means of (2.6). We shall not spell out the formulas, but mention only an important reinterpretation in the case $N = 1$. Considering the number $\widetilde{X}_t = K - X_t$ of working machines instead of X_t, the intensities change to

$$\widetilde{\beta}_n = \delta_{K-n} = \delta,\ n = 0,\dots,K-1, \quad \widetilde{\delta}_n = \beta_{K-n} = n\beta,\ n = 1,\dots,K.$$

This is of the same form as for Erlang's loss system, and we conclude immediately that $\widetilde{\boldsymbol{\pi}}$ is truncated Poisson in equilibrium.

This model has recently received renewed attention due to a computer system interpretation, where one thinks of the customers as terminals and the repairman as the computer handling requests from the terminals (with or without processor sharing). Thus each terminal generates requests with intensity β, and X_t is the number of requests being presently handled by the computer.

Problems

3.1 Consider the case $\beta_n = \beta/(n+1)$, $\delta_n = \delta$ of the customers being discouraged by long queue lengths (reneging). Show that the ergodic distribution exists and is Poisson.

3.2 Show the recursion formula $E_{K+1}(\eta) = \eta E_K(\eta)/[K+1+\eta E_K(\eta)]$; cf. (3.3).

3.3 Let $\beta_n = \beta(N+n)$, $\delta_n = n\delta$. Show that there is ergodicity for $\beta/\delta < 1$ and that $\boldsymbol{\pi}$ is negative binomial,

$$\pi_n = \binom{-N}{n}(-\rho^n)(1-\rho)^N.$$

Give a demographic interpretation of the model.

3.4 Consider Erlang's loss system and let $H(k)$ denote the probability that k lines are busy. Show the *Palm–Jacobæus* formula $H(k) = E_K(\eta)/E_{K-k}(\eta)$.

3.5 Consider the same model as in Engseth's loss system except that now $N \le K$. Show that $\boldsymbol{\pi}$ is binomial (N,p) where $p = \beta/(\beta+\delta)$.

3.6 Consider the $M/M/2$ queue with *heterogeneous servers*, i.e. servers 1, 2 have intensities $\delta^{(1)} > \delta^{(2)}$. There are many ways to model the system behaviour if one or both servers are idle, but we assume here that if server 1 becomes idle, the customer served by 2 switches to 1. Explain that this corresponds to $\beta_n = \beta$, $\delta_1 = \delta^{(1)}$, $\delta_n = \delta^{(1)} + \delta^{(2)}$, $n = 2, 3, \ldots$. Show that there is ergodicity if and only if $\beta < \delta^{(1)} + \delta^{(2)}$ and find the stationary distribution.

3.7 Same questions as in Problem 3.6, but the model is now modified such that the customers cannot switch to 1 and an arriving customer always joins 1 if the system is idle. [*Hint:* Look first at the system restricted to $\{2, 3, \ldots\}$ as a birth–death process, and next split state 1 into two states indicating which server is busy.]

4 The Phase Method

The amenability of Markovian models to analysis should by now have become apparent, and further examples are given in the following sections and Chapter IV. However, the Markovian set–up puts some restriction on the modelling and one of the most serious ones is that whereas it will frequently be very reasonable to assume that interarrival times are exponential (i.e. we have Poisson arrivals), then this is not the case for service times.

The first idea on how to overcome this apparent difficulty was the so–called *method of stages* due to Erlang. The idea is to think of the customer as being composed of k stages each having an exponential service time, say with intensity δ. The stages are then served one at a time, and the customer completes service when all stages are served. That is, the service time of the customer himself is the *Erlang distribution* with k stages, namely a convolution of k exponentials with the same intensity δ so that the density is

$$\delta^k \frac{x^{k-1}}{(k-1)!} \mathrm{e}^{-\delta x}, \quad x > 0. \tag{4.1}$$

The point is now that if we count stages instead of customers and the arrival process is say Poisson, then we get a Markov process $\{X_t\}_{t\geq 0}$. Indeed, since the arrival of a customer corresponds to the arrival of k stages, the nonzero off–diagonal intensities are $\lambda(n, n+k) = \beta$, $\lambda(n, n-1) = \delta$, $n \geq 1$. The queue length process $\{Q_t\}$ is then obtained simply by summing the stages out. For example, if we want to determine $\mathbb{P}_e(Q_t = n)$, we solve for the stationary distribution π for $\{X_t\}$ and have

$$\mathbb{P}_e(Q_t = n) \;=\; \pi_{(n-1)k+1} + \pi_{(n-1)k+2} + \cdots + \pi_{nk}.$$

Apparently what we have just described is, in the Kendall notation, the queueing system $M/E_k/1$. It should be stressed that this way of imbedding an apparently non–Markovian queue into a Markovian set–up is essentially an artifice: the stages themselves usually can be given no physical interpretation. The gain is the greater flexibility in the choice of the service time distribution.

A related classical idea is to use instead a service time distribution that is a mixture of exponentials corresponding to a density of the form $\sum_1^k \alpha_r \delta_r \mathrm{e}^{-\delta_r x}$, $x > 0$, where $0 < \alpha_r < 1$, $\sum_1^k \alpha_r = 1$; such a distribution is denoted by H_k, *the hyperexponential distribution with k parallel channels.* The state space of the Markov process describing the $M/H_k/1$ queue is $\mathbb{N} \times \{1, \ldots, k\}$, the first component describing the number in system and the second the channel in which the server is currently operating (will serve the next customer when the system is idle), and the nontrivial intensities are

$$\lambda(nr, (n+1)r) = \beta, \quad \lambda(nr, (n-1)s) = \delta_r \alpha_s.$$

In Fig. 4.1, some examples of E_k– and H_k densities are given. The plots illustrate among other things the behaviour of the squared coefficient of variation (s.c.v.) η. This is $1/k$ for E_k, i.e. $\eta \in (0, 1]$ with 1 attained for the exponential distribution E_1 and 0 in the limit $k \to \infty$ (i.e. E_k approaches D). For H_k, always $\eta > 1$ (provided at least two δ_r are different) and the range is $(1, \infty)$ for all k. To derive these properties, use for E_k either standard moment formulas for the gamma distribution or the representation as a sum S_k of i.i.d. exponentials $Y_1, \ldots, Y_k$, yielding

$$\eta \;=\; \frac{\mathbb{V}ar S_k}{(\mathbb{E} S_k)^2} \;=\; \frac{k \mathbb{V}ar Y_1}{(k \mathbb{E} Y_1)^2} \;=\; \frac{k/\delta^2}{(k/\delta)^2} \;=\; \frac{1}{k}.$$

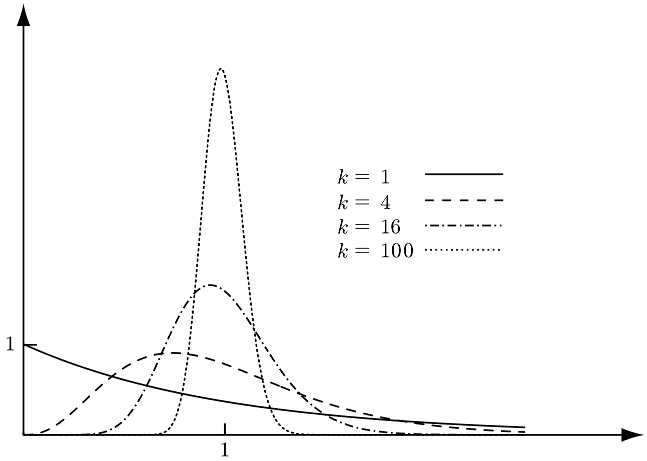

Figure 4.1(a)

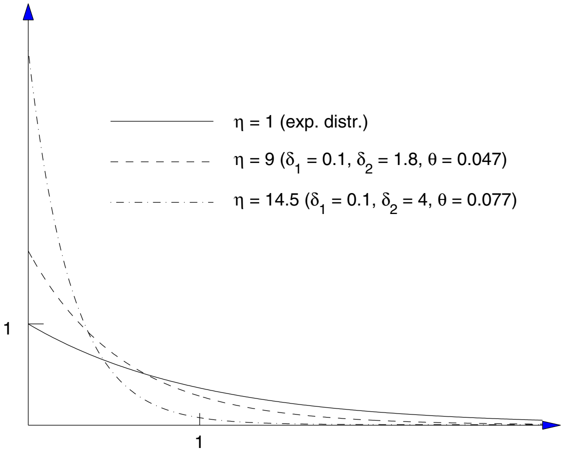

Figure 4.1(b)

For H_k, we may use the representation Y_τ where $\mathbb{P}(\tau = r) = \alpha_r$, Y_r has density $\delta_r e^{-\delta_r x}$ (i.e. mean $\mu_r = \delta_r^{-1}$ and variance $\sigma_r^2 = \delta_r^{-2}$) and is independent of τ. Then conditioning upon τ we get

$$\eta = \frac{\mathbb{V}ar Y_\tau}{(\mathbb{E}Y_\tau)^2} = \frac{\mathbb{V}ar \mu_\tau + \mathbb{E}\sigma_\tau^2}{(\mathbb{E}\mu_\tau)^2} = \frac{\mathbb{E}\mu_\tau^2 - (\mathbb{E}\mu_\tau)^2 + \mathbb{E}\mu_\tau^2}{(\mathbb{E}\mu_\tau)^2} = \frac{2\mathbb{E}\mu_\tau^2}{(\mathbb{E}\mu_\tau)^2} - 1$$

which is > 1 provided $\mathbb{V}ar\mu_\tau > 0$, i.e. at least two δ_r are different.

In early literature, the discussion sometimes stopped at this point, the argument being that for a given distribution one can always choose an appropriate H_k or E_k with a good fit to the s.c.v. However, not only is this point of view very rigid (two densities with the same s.c.v. may be very different), but in fact it turns out that usually no additional difficulties arise when working with general *phase–type distributions* and that in fact these approximate any distribution arbitrarily well. Such a distribution F is defined in terms of a Markov jump process $\{J_t\}_{t\geq 0}$ with finite state space $E\cup\{\Delta\}$, such that Δ is absorbing and the states in E transient, and an initial distribution $\boldsymbol{\alpha}$, such that F is the distribution of the time $\zeta = \inf\{t>0: J_t=\Delta\}$ to absorption, $F(t)=\mathbb{P}_{\boldsymbol{\alpha}}(\zeta\leq t)$. It is usually assumed that $\boldsymbol{\alpha}$ has mass 0 at Δ such that we can write $\boldsymbol{\alpha}$ as an E–vector. Further, the intensity matrix $\boldsymbol{Q}$ partitioned according to states in E, resp. the single state Δ, must have the form

$$\boldsymbol{Q} \;=\; \begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \boldsymbol{0} & 0 \end{pmatrix} \tag{4.2}$$

(the form of the last row follows from Δ being absorbing) where $\boldsymbol{t}=-\boldsymbol{T}\boldsymbol{1}$ with $\boldsymbol{1}=(1\ \ldots\ 1)^{\mathsf{T}}$ since the rows sum to 0. We refer to the $E\times E$ matrix $\boldsymbol{T}$ as the *phase generator*, to the E–column vector $\boldsymbol{t}$ as the *exit vector*, to $(E,\boldsymbol{\alpha},\boldsymbol{T})$ or sometimes just $(\boldsymbol{\alpha},\boldsymbol{T})$ as the *representation* of F and write $F\in\mathscr{PH}$.

Proposition 4.1 *Let $F\in\mathscr{PH}$ have representation $(E,\boldsymbol{\alpha},\boldsymbol{T})$. Then:*
(i) *For $x\geq 0$, the c.d.f. is $F(x)=1-\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{1}$ and the density is $f(x)=\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{t}$.*
(ii) *The nth moment is $(-1)^n n!\boldsymbol{\alpha}\boldsymbol{T}^{-n}\boldsymbol{1}$.*
(iii) *The Laplace transform $\hat{F}[s]=\int_0^\infty \mathrm{e}^{-sx}f(x)\,\mathrm{d}x$ is $\hat{F}[s]=\boldsymbol{\alpha}(s\boldsymbol{I}-\boldsymbol{T})^{-1}\boldsymbol{t}$ and is rational* [a ratio between two polynomials].

Proof. It follows easily by induction from (4.2) that the upper left corner of $\boldsymbol{Q}^n$ is $\boldsymbol{T}^n$. Hence the upper left corner of $\mathrm{e}^{\boldsymbol{Q}x}$ is $\mathrm{e}^{\boldsymbol{T}x}$ and therefore

$$\begin{aligned} 1-F(x) &= \mathbb{P}_{\boldsymbol{\alpha}}(\zeta>x) \;=\; \mathbb{P}_{\boldsymbol{\alpha}}(J_x\in E) \;=\; \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{1}, \\ f(x) &= -\frac{\mathrm{d}}{\mathrm{d}x}\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{1} \;=\; -\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{T}\boldsymbol{1} \;=\; \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{t}. \end{aligned}$$

For (iii), note that according to II.4d all eigenvalues λ for $\boldsymbol{T}$ have negative real part. Hence so is the case for $\boldsymbol{A}=-s\boldsymbol{I}+\boldsymbol{T}$ when $\Re(s)\geq 0$, and the matrix analogue of the formula $\int_0^\infty \mathrm{e}^{ax}\,\mathrm{d}x=-1/a$, $\Re(a)<0$, then yields

$$\hat{F}[s] \;=\; \boldsymbol{\alpha}\Bigl(\int_0^\infty \mathrm{e}^{-sx}\mathrm{e}^{\boldsymbol{T}x}\,\mathrm{d}x\Bigr)\boldsymbol{t} \;=\; \boldsymbol{\alpha}\Bigl(\int_0^\infty \mathrm{e}^{\boldsymbol{A}x}\,\mathrm{d}x\Bigr)\boldsymbol{t} \;=\; -\boldsymbol{\alpha}\boldsymbol{A}^{-1}\boldsymbol{t};$$

that $\hat{F}[s]$ is rational then follows since all elements of $(s\boldsymbol{I}-\boldsymbol{T})^{-1}$ are so because the determinant and all subdeterminants of $s\boldsymbol{I}-\boldsymbol{T}$ are polynomials.

Part (ii) follows by differentiating the m.g.f., which yields the nth moment as

$$\begin{aligned} \frac{\mathrm{d}^n}{\mathrm{d}s^n}\boldsymbol{\alpha}(-s\boldsymbol{I}-\boldsymbol{T})^{-1}\boldsymbol{t}\Big|_{s=0} &= (-1)^{n+1}n!\boldsymbol{\alpha}(s\boldsymbol{I}+\boldsymbol{T})^{-n-1}\boldsymbol{t}\Big|_{s=0} \\ = (-1)^{n+1}n!\boldsymbol{\alpha}\boldsymbol{T}^{-n-1}\boldsymbol{t} &= (-1)^n n!\boldsymbol{\alpha}\boldsymbol{T}^{-n-1}\boldsymbol{T}\boldsymbol{1} \;=\; (-1)^n n!\boldsymbol{\alpha}\boldsymbol{T}^{-n}\boldsymbol{1}\,. \end{aligned}$$

□

A major property of the class $\mathscr{PH}$ is that any distribution F on $[0,\infty)$ can be approximated arbitrarily well by a phase–type distribution. In the proof of this fact, we will employ the class $\mathscr{PH}_{\mathrm{ME}}$ of (finite) mixtures of Erlang distributions with the same intensity, i.e. with densities of the form

$$\sum_{i=1}^{k} \alpha_i \delta^{n(i)+1} \frac{x^{n(i)}}{n(i)!} e^{-\delta x}, \quad x > 0, \tag{4.3}$$

where $n(i) \in \mathbb{N}$, $\alpha_i > 0$ for all i and $\sum_1^k \alpha_i = 1$; this class is obviously a subclass of the class $\mathscr{PH}_{\mathrm{S/P}}$ of *exponential distributions in series and/or parallel*, that is, with a phase representation as in Fig. 4.2 or equivalently with a Laplace transform of the form

$$\sum_{i=1}^{k} \alpha_i \prod_{j=1}^{n(i)} \frac{\delta_{ij}}{\delta_{ij} + s}. \tag{4.4}$$

Clearly, $\mathscr{PH}_{\mathrm{ME}}$ corresponds to $\delta_{ir} \equiv \delta$ and $\mathscr{PH}_{\mathrm{S/P}} \subseteq \mathscr{PH}$.

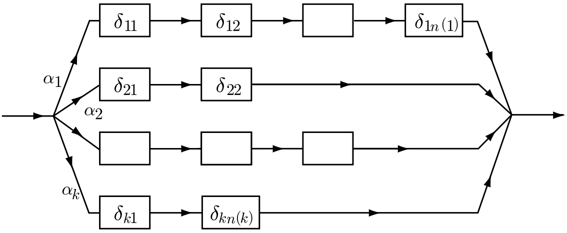

Figure 4.2

Theorem 4.2 *The class $\mathscr{PH}$ is dense* (in the sense of weak convergence) *in the set $\mathscr{P}$ of all probability distributions on* $(0,\infty)$*. More generally, to any $F \in \mathscr{P}$ with finite pth moment $\mu_F^{(p)}$ there are $F_k \in \mathscr{PH}$ with $F_k \xrightarrow{w} F$ and $\mu_{F_k}^{(q)} \to \mu_F^{(q)}$ for all $q \le p$.*

Proof. Let d be a metric for weak convergence in $\mathscr{P}$ and define

$$d_p(F,G) \;=\; d(F,G) + \left|\mu_F^{(p)} - \mu_G^{(p)}\right|, \quad p \ge 0$$

(so that $d_0 = d$). Since $F_k \xrightarrow{w} F$ and $\mu_{F_k}^{(p)} \to \mu_F^{(p)}$ implies $\mu_{F_k}^{(q)} \to \mu_F^{(q)}$ for $q \le p$ (uniform integrability!), it is sufficient to show that $\mathscr{PH}_{\mathrm{ME}}$ is dense w.r.t. d_p in $\mathscr{P}_p = \left\{P \in \mathscr{P} : \mu_F^{(p)} < \infty\right\}$. Letting F_A be F truncated at A (i.e $F_A(x) = F(x \wedge A)/F(A)$), it is easily seen that $d_p(F, F_A) \to 0$ as $A \to \infty$. Hence if $\mathscr{P}^{(A)}$ is the set of distributions supported by $(0, A]$, $\cup_{A<\infty} \mathscr{P}^{(A)}$ is dense in $\mathscr{P}$. Further, it is standard that the subset $\widetilde{\mathscr{P}}^{(A)}$ of $\mathscr{P}^{(A)}$ consisting of distributions with a finite support is dense w.r.t. d, and hence w.r.t. d_p since d, d_p are equivalent on $\mathscr{P}^{(A)}$. Now let $G \in \widetilde{\mathscr{P}}^{(A)}$ have atoms $t_1, \ldots, t_k$ with weights $\alpha_1, \ldots, \alpha_k$. For each i, choose integers $n_{m,i}$ such that $n_{m,i}/m \to t_i$, $m \to \infty$. Consider the Erlang distribution $G_{m,i}$

with $n_{m,i}$ stages and intensity $\delta_m = m$. The mean is $n_{m,i}/m$ and the s.c.v. is $n_{m,i}^{-1}$, hence the weak limit as $m \to \infty$ is the distribution degenerate at t_i. Thus with $G_m = \sum_1^k \alpha_i G_{m,i}$ we have $d(G_{m,i}, G) \to 0$. An easy calculation shows that also all moments converge. In particular, $d_p(G_{m,i}, G) \to 0$ so that $\overline{\mathscr{PH}_{\mathrm{ME}}} \supseteq \widetilde{\mathscr{P}}^{(A)}$. Taking first the union over A and next the closure shows that $\mathscr{PH}_{\mathrm{ME}}$ is dense in $\mathscr{P}_p$. Hence $\mathscr{PH}$ is so. □

The denseness of $\mathscr{PH}$ is illustrated in Fig. 4.3, where it is shown how fits $F_p \in \mathscr{PH}$ with p phases (produced by maximum likelihood for $p = 2, 3$ and 6) provide a convergent sequence of approximations of a given distribution F (in this case an inverse Gaussian having density XIII.(4.3) with $\xi = 1$, $c = 2$).

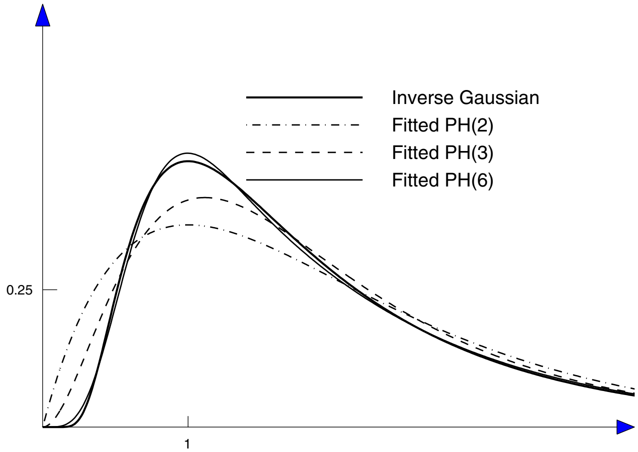

Figure 4.3

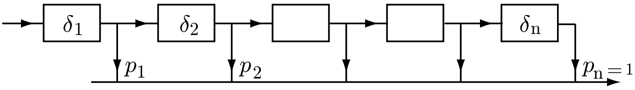

Figure 4.4

In much of the older literature, one works within the class $\mathscr{RLT}$ of distributions with rational Laplace transforms. In addition to the special classes of phase–type distributions discussed above, the class $\mathscr{PH}_{\mathrm{C}}$ of *Coxian distributions* is also frequently encountered. This is defined by a representation as in Fig. 4.4 or equivalently by a Laplace transform of the form

$$\sum_{k=1}^{n} q_1 \ldots q_{k-1} p_k \prod_{i=1}^{k} \frac{\delta_i}{\delta_i + s} \tag{4.5}$$

where $q_k = 1 - p_k$. One has:

Theorem 4.3 *$\mathscr{PH}_{\mathrm{ME}} \subset \mathscr{PH}_{\mathrm{C}} = \mathscr{PH}_{\mathrm{S/P}} \subset \mathscr{PH} \subset \mathscr{RLT}$.*

Proof. Assume w.l.o.g. that $n(i) = i - 1$ in (4.3) (take some $\alpha_i = 0$). Then letting $n = n(k)$, $\delta_i = \delta$, $p_1 = \alpha_1$, $p_2 = \alpha_2/(1-\alpha_1)$, $p_3 = \alpha_3/(1-\alpha_1-\alpha_2)$, ... in Fig. 4.4 shows that $\mathscr{PH}_{\mathrm{ME}} \subseteq \mathscr{PH}_{\mathrm{C}}$ (that the inclusion is strict is obvious; e.g. $\mathscr{PH}_{\mathrm{C}}$ contains the convolution of two exponentials with different intensities which is clearly not in $\mathscr{PH}_{\mathrm{ME}}$). That $\mathscr{PH}_{\mathrm{C}} \subseteq \mathscr{PH}_{\mathrm{S/P}}$ follows (say) from the expressions (4.5), (4.4) for the Laplace transforms.

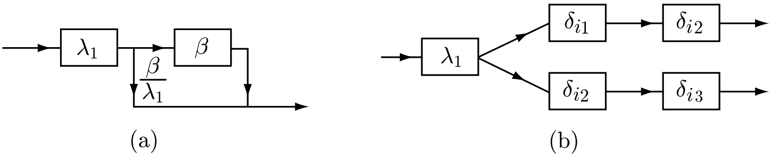

Figure 4.5

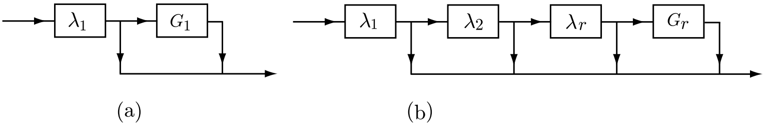

Figure 4.6

For the converse, consider $G \in \mathscr{PH}_{\mathrm{S/P}}$ represented as in Fig. 4.2. We may clearly assume that $\delta_{i1} \geq \delta_{i2} \geq \cdots$ for all i, and define λ_1 as the largest δ_{i1}. Now for any $\beta < \lambda_1$ a simple calculation shows that the distribution in Fig. 4.5(a) is simply the exponential distribution with intensity β. Applying this to a channel with $\beta = \delta_{i1} < \lambda_1$ yields the representation in Fig. 4.5(b), and altogether we may represent G as in Fig. 4.6(a) where G_1 has the same intensities as G, except that λ_1 has been removed. In any case, the maximal number of occurences of λ_1 in any channel has been reduced by 1, and continuing in this manner we end up with the situation in Fig. 4.6(b) where G_r has a representation with possibly many channels but only one intensity. That is, G_r is an exponential distribution so that indeed $G \in \mathscr{PH}_{\mathrm{C}}$. That $\mathscr{PH}_{\mathrm{S/P}} \subseteq \mathscr{P}H$ is trivial; that the inclusion is strict is shown in Problem 4.6. Finally, $\mathscr{PH} \subseteq \mathscr{RLT}$ follows from Proposition 4.1(iii), and the strict inclusion from Problem 4.7. □

Noting that the proof of Theorem 4.2 only used the class $\mathscr{PH}_{\mathrm{ME}}$, we get:

Corollary 4.4 *The conclusion of Theorem* 4.2 *holds true if* $\mathscr{PH}$ *is replaced by any of the classes* $\mathscr{PH}_{\mathrm{ME}}$, $\mathscr{PH}_{\mathrm{C}}$, $\mathscr{PH}_{\mathrm{S/P}}$ *or* $\mathscr{RLT}$.

When faced with a queueing problem on say $M/G/1$, we may now approximate (say by maximum likelihood) the service time distribution B by some $\widetilde{B} \in \mathscr{PH}$ and think of the server as moving in E in the same way as $\{J_t\}$ during services and being restarted according to $\boldsymbol{\alpha}$ at each service completion. Exactly as for $M/E_k/1$ or $M/H_k/1$, this yields a Markovian representation of the approximating queue, and the steady–state solution may then be used as an approximation to the steady–state solution of the given $M/G/1$ queue. However:

1. From the theoretical point of view, it must be proved that approximation of B with $\widetilde{B}$ also implies approximation of the corresponding steady–state characteristics of the queue. This is a question of so–called *continuity* or *robustness*, and far from trivial or without pitfalls; see Problem 2.5 and X.6.

2. From the practical point of view, a good approximation of B may for certain types of distributions require a large E (note the slow convergence to 1 in Fig. 4.1(a)!), so that the solution of the approximating $M/PH/1$ queue will be computationally demanding.

3. Modelling by a Markov jump process is not the only way to exploit phase–type distributions in queueing theory. Another is to use the probabilistic interpretation to provide solutions to certain fundamental random walk problems; see further VIII.5. Still another point of view is taken in the literature based upon transform methods, where the class $\mathscr{RLT}$ is exploited in a purely analytical way.

Problems

4.1 Write up the appropriate state space and intensities for some queueing system such as $H_k/E_\ell/1$, $M/H_k/c$, etc.

4.2 Show that the s.c.v. in H_2 can attain any value in $(1,\infty)$.

4.3 Let $F_1, F_2 \in \mathscr{PH}$. Show that the convolution $F_1 * F_2$ and a convex combination $\theta F_1 + (1-\theta)F_2$ are again in $\mathscr{PH}$.

4.4 Show that if X has distribution $F \in \mathscr{PH}$ with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$, then the overshoot distribution $F^{(z)}$ (the distribution of $X - z$ given $X > z$, i.e. $\overline{F}^{(z)}(y) = \overline{F}(y+z)/\overline{F}(z)$) is phase–type with representation $(\boldsymbol{\alpha}^{(z)}, \boldsymbol{T})$ for some $\boldsymbol{\alpha}^{(z)}$, and give an expression for $\boldsymbol{\alpha}^{(z)}$.

4.5 Give an alternative derivation of the form of the mean of $F \in \mathscr{PH}$ by deriving equations for the $\mathbb{E}_i\zeta$ by conditioning upon the first jump. Do the same for the Laplace transform.

4.6 Show that (a) if $F \in \mathscr{PH}_{\mathrm{S/P}}$ has the Laplace transform Q/R with Q, R polynomials without common roots, then R cannot have complex roots; (b) if $F = (1-\theta)\sum_1^\infty \theta^{n-1} G^{*n}$ with $G \in \mathscr{PH}$, then $F \in \mathscr{PH}$; (c) if $G = E_3$ in (b), then $F \notin \mathscr{PH}_{\mathrm{S/P}}$.

4.7 Show that (a) the density $f(x)$ of $F \in \mathscr{PH}$ satisfies $f(x) > 0$, $x > 0$. [*Hint: F* contains a component of exponentials in series.] Show that (b) the distribution with density proportional to $(1+\sin x)\mathrm{e}^{-x}$ is in $\mathscr{RLT} \setminus \mathscr{PH}$.

Notes The modern revival of the class $\mathscr{PH}$ is due to a large extent to M.F. Neuts, a main source being his 1981 book; some later textbooks with more extensive treatments are Wolff (1989), Rolski *et al.* (1999) and Asmussen (2000). Statistical fitting is treated via maximum likelihood in Asmussen *et al.* (1996)

and via a Bayesian Markov chain Monte Carlo approach in Bladt *et al.* (2003). A survey of the class $\mathscr{RLT}$ is in Asmussen and O'Cinneide (1999); there is also much discussion, in part at the more heuristical level, in Lipsky (1992).

5 Renewal Theory for Phase–Type Distributions

We consider a point process on $[0,\infty)$ with epochs $0 = S_0 < S_1 < S_2$ such that the interarrival times $Y_k = S_k - S_{k-1}$ are i.i.d. with common distribution F of phase–type, say with representation $(E, \boldsymbol{\alpha}, \boldsymbol{T})$.

The key idea in studying such a process is to piece together the phase processes governing the individual Y_k to a Markov process $\{X_t\}_{t\geq 0}$. Namely, assume given i.i.d. Markov processes $\{J_t^{(1)}\}$, $\{J_t^{(2)}\}$, ... on $E\cup\{\Delta\}$ with the same distribution as $\{J_t\}$ in Section 4. We can then represent Y_k as the absorbtion time of $\{J_t^{(k)}\}$ and define $X_t = J_t^{(1)}$, $0\leq t< Y_1$, $X_t = J_t^{(2)}$, $Y_1\leq t< Y_1+Y_2$ and so on; cf. Fig. 5.1 where there are two Markov states 1 =thin, 2 =thick.

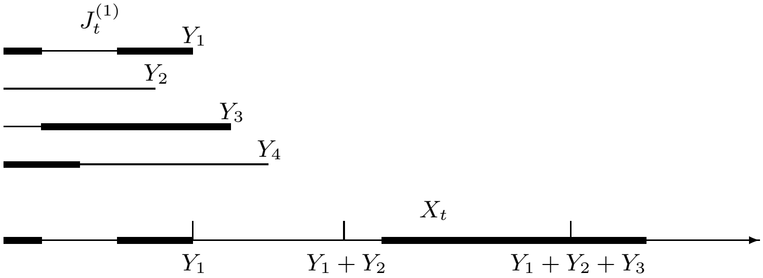

Figure 5.1

Proposition 5.1 *$\{X_t\}$ is Markov on E with intensity matrix $\boldsymbol{\Lambda} = \boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha}$.*

Proof. Let $i,j\in E$, $i\neq j$. Then a jump from i to j occurs if the $\{J_t^{(k)}\}$ currently in operation is in state i and either makes a jump to j (the intensity is t_{ij}), or if it jumps to the absorbing state (intensity t_i) and $J_0^{(k+1)} = j$ which occurs w.p. α_j. Hence the intensity is $t_{ij}+t_i\alpha_j$ independently of the past, which shows the Markov property and that the off–diagonal elements of $\boldsymbol{\Lambda}$ are as asserted. It only remains to check that the rows sum to zero, which follows since $\boldsymbol{\alpha}\boldsymbol{1} = 1$ implies $(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha})\boldsymbol{1} = -\boldsymbol{t}+\boldsymbol{t} = \boldsymbol{0}$. □

We define as in I.2 the forward recurrence time B_t as the waiting time until the next renewal, i.e. $B_t = S_n - t$ if $S_{n-1}\leq t< S_n$. The *renewal density* $u(t)$ is defined as the density of the intensity measure of the renewal point process $\{S_n\}$ (cf. A3).

Corollary 5.2 (a) *The distribution of B_t is phase–type with representation $(\boldsymbol{\alpha}_t, \boldsymbol{T})$ where $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}\mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha})t}$;* (b) *the renewal density exists and is given by $u(t) = \boldsymbol{\alpha}_t\boldsymbol{t} = \boldsymbol{\alpha}\mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha})t}\boldsymbol{t}$.*

Proof. Since $\{X_t\}$ has initial distribution $\boldsymbol{\alpha}$ and intensity matrix $\boldsymbol{T}+\boldsymbol{t\alpha}$, $\boldsymbol{\alpha}_t$ is simply the distribution of X_t. Part (a) is then immediately clear from the probabilistic interpretation. For (b), just note that $u(t)$ must be the same as the density of B_t at 0, which is $\boldsymbol{\alpha}_t \mathrm{e}^{\boldsymbol{T}\cdot 0}\boldsymbol{t} = \boldsymbol{\alpha}_t\boldsymbol{t}$. □

Corollary 5.3 *Assume that* $\boldsymbol{\Lambda}$ *is irreducible. Then the stationary distribution of* $\{X_t\}$ *is* $\boldsymbol{\pi} = -\boldsymbol{\alpha T}^{-1}/\mu_F$ *where* $\mu_F = -\boldsymbol{\alpha T}^{-1}\mathbf{1}$ *is the mean of* F*. Further,* B_t *has a limiting distribution that is phase–type with representation* $(\boldsymbol{\pi},\boldsymbol{T})$*.*

Proof. The expression for μ_F (Proposition 4.1(ii)) shows immediately that $\boldsymbol{\pi}\mathbf{1} = 1$, so that stationarity follows from

$$\mu_F\boldsymbol{\pi\Lambda} \;=\; -\boldsymbol{\alpha T}^{-1}(\boldsymbol{T}+\boldsymbol{t\alpha}) \;=\; -\boldsymbol{\alpha}+\boldsymbol{\alpha}\mathbf{1}\boldsymbol{\alpha} = \; -\boldsymbol{\alpha}+\boldsymbol{\alpha} \;=\; \mathbf{0}.$$

The last statement follows since by ergodicity $\boldsymbol{\alpha}_t \to \boldsymbol{\pi}$. □

Remark 5.4 Since $\mathrm{e}^{\boldsymbol{\Lambda}t} \to \mathbf{1}\boldsymbol{\pi}$, it follows that

$$u(t) \;\to\; \boldsymbol{\alpha}\mathbf{1}\boldsymbol{\pi t} \;=\; \boldsymbol{\pi t} \;=\; \mu_F^{-1}(-\boldsymbol{\alpha T}^{-1})(-\boldsymbol{T}\mathbf{1}) \;=\; \mu_F^{-1}\boldsymbol{\alpha}\mathbf{1} \;=\; \mu_F^{-1}\,.$$

This is a continuous time and nonlattice version of I.2.2, and the generalization to a completely general (rather than phase–type) distribution F is equivalent (in the absolutely continuous case) to the renewal theorem to be presented in V.4. But note that this generalization has no easy proof using the denseness of phase–type distributions. □

Example 5.5 Consider the case of two phases where

$$\boldsymbol{\Lambda} \;=\; \boldsymbol{T}+\boldsymbol{t\alpha} \;=\; \begin{pmatrix} t_{11}+t_1\alpha_1 & t_{12}+t_1\alpha_2 \\ t_{12}+t_2\alpha_1 & t_{22}+t_2\alpha_2 \end{pmatrix} \;=\; \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix} \quad \text{(say)}.$$

Here $\boldsymbol{\pi} = (\pi_1\ \pi_2) = (\lambda_2/(\lambda_1+\lambda_2)\ \lambda_1/(\lambda_1+\lambda_2))$. The nonzero eigenvalue of $\boldsymbol{\Lambda}$ is $\lambda = -\lambda_1-\lambda_2$, and by standard diagonalization techniques (see II.3.6 for details) we get the renewal density as

$$\begin{aligned} u(t) \;=\; \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{\Lambda}t}\boldsymbol{t} \;&=\; (\alpha_1\ \alpha_2)\left\{\begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix} + \mathrm{e}^{\lambda t}\begin{pmatrix} \pi_2 & -\pi_2 \\ -\pi_1 & \pi_1 \end{pmatrix}\right\}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \\ &=\; (\pi_1\ \pi_2)\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \;+\; \mathrm{e}^{\lambda t}\,(\alpha_1\ \alpha_2)\begin{pmatrix} \pi_2(t_1-t_2) \\ \pi_1(t_2-t_1) \end{pmatrix} \\ &=\; \pi_1t_1+\pi_2t_2+\mathrm{e}^{\lambda t}\,(\alpha_1\pi_2-\alpha_2\pi_1)\,(t_1-t_2) \\ &=\; \frac{1}{\mu_F}+\mathrm{e}^{\lambda t}\,(\alpha_1\pi_2-\alpha_2\pi_1)\,(t_1-t_2) \end{aligned}$$

□

Example 5.6 Consider the Erlang distribution with n stages and, w.l.o.g., mean n, i.e. $\delta = 1$. Here $\boldsymbol{\Lambda} = \boldsymbol{T} + \boldsymbol{t\alpha}$ is the matrix

$$\begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 1 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix}$$

(this form of $\boldsymbol{\Lambda}$ is also probabilistically obvious since $\{X_t\}$ is cyclic). The characteristic equation of this matrix is $(1+\lambda)^n = 1$, with roots $\lambda_k = \theta^k - 1$, $k = 0, \ldots, n-1$, where $\theta = \mathrm{e}^{\mathrm{i}2\pi/n}$ is the nth root of unity. The corresponding left and right eigenvectors are

$$\boldsymbol{l}_k = \begin{pmatrix} \theta^{-k} & \theta^{-2k} & \theta^{-3k} & \cdots & \theta^{-(n-1)k} & \theta^{-nk} \end{pmatrix}, \quad \boldsymbol{r}_k = \frac{1}{n} \begin{pmatrix} \theta^k \\ \theta^{2k} \\ \theta^{3k} \\ \vdots \\ \theta^{(n-1)k} \\ \theta^{nk} \end{pmatrix}$$

(here $\theta^{-nk} = \theta^{nk} = 1$) where $1/n$ occurs to obtain $\boldsymbol{l}_k \boldsymbol{r}_k = 1$. Thus

$$\begin{aligned} \boldsymbol{\Lambda} &= \sum_{k=0}^{n-1} \lambda_k \boldsymbol{r}_k \boldsymbol{l}_k, \quad \mathrm{e}^{\boldsymbol{\Lambda}x} = \sum_{k=0}^{n-1} \mathrm{e}^{\lambda_k x} \boldsymbol{r}_k \boldsymbol{l}_k, \\ u(x) &= \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{\Lambda}x}\boldsymbol{t} = \sum_{k=0}^{n-1} \mathrm{e}^{\lambda_k x} (\boldsymbol{\alpha r}_k) \cdot (\boldsymbol{l}_k \boldsymbol{t}) = \sum_{k=0}^{n-1} \mathrm{e}^{\lambda_k x} \frac{\theta^k}{n} \cdot 1 \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \exp\Big\{ \Big(\cos\frac{2\pi k}{n} - 1\Big)x + \mathrm{i}\Big(\sin\frac{2\pi k}{n}x + \frac{2\pi k}{n}\Big) \Big\} \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \exp\Big\{ -\Big(1 - \cos\frac{2\pi k}{n}\Big)x \Big\} \cos\Big(\sin\frac{2\pi k}{n}x + \frac{2\pi k}{n}\Big) \end{aligned}$$

(using $u(x) \in \mathbb{R}$ in the last step). □

Returning to the general theory, consider a doubly infinite stationary version $\{X_t\}_{-\infty<t<\infty}$ of $\{X_t\}$. Let $\{\widetilde{X}_t\}$ denote the time–reversed process ($\widetilde{X}_t = X_{-t-}$) and $\widetilde{\boldsymbol{\Lambda}}$ its intensity matrix with rsth element $\pi_s \lambda(s,r)/\pi_r$, and define

$$\widetilde{\alpha}_r = \mu_F t_r \pi_r, \quad \widetilde{t}_{rs} = \frac{\pi_s t_{sr}}{\pi_r}, \quad \widetilde{t}_r = \frac{\alpha_r}{\mu_F \pi_r}.$$

Then $\widetilde{\boldsymbol{\alpha}}$ is a probability vector since

$$\widetilde{\boldsymbol{\alpha}}\mathbf{1} \;=\; \mu_F \boldsymbol{\pi}\boldsymbol{t} \;=\; (-\boldsymbol{\alpha}\boldsymbol{T}^{-1})(-\boldsymbol{T}\mathbf{1}) \;=\; \boldsymbol{\alpha}\mathbf{1} \;=\; 1.$$

Further, just the same argument as in II.5.2 shows that transitions $s \to r$ of $\{X_t\}$ not taking place at a renewal epoch (i.e. governed by the $\boldsymbol{T}$–part of $\boldsymbol{\Lambda}$) correspond to transitions $r \to s$ of $\{\widetilde{X}_t\}$ governed by $\widetilde{\boldsymbol{T}}$. Since

$$\widetilde{\lambda}(r,s) \;=\; \frac{\pi_s}{\pi_r}(t_{sr} + t_s \alpha_r) \;=\; \widetilde{t}_{rs} + \widetilde{t}_r \widetilde{\alpha}_s,$$

it follows that transitions $s \to r$ ($r = s$ is included) of $\{X_t\}$ at renewal epochs (i.e., governed by the $\boldsymbol{t\alpha}$–part of $\boldsymbol{\Lambda}$) correspond to transitions $r \to s$ of $\{\widetilde{X}_t\}$ governed by $\widetilde{\boldsymbol{t}}\widetilde{\boldsymbol{\alpha}}$. Thus the time–reversed renewal point process must be a phase–type renewal process with representation $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{T}})$ of the interarrival distribution. However, the long–run distribution of interarrival times is the same no matter whether time is read forward or backward, and hence:

Proposition 5.7 *Let F be phase–type with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$, and assume that $\boldsymbol{T} + \boldsymbol{t\alpha}$ is irreducible. Then $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{T}})$ is again a representation of F.*

For an algebraic proof, let $\boldsymbol{\Delta}$ be the diagonal matrix with the π_r on the diagonal. Then

$$\begin{aligned}
\widetilde{\boldsymbol{\alpha}} \mathrm{e}^{\widetilde{\boldsymbol{T}}t} \widetilde{\boldsymbol{t}} \;&=\; (\mu_F \boldsymbol{\Delta t})^{\mathsf{T}} \mathrm{e}^{\boldsymbol{\Delta}^{-1}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{\Delta}t} (\boldsymbol{\alpha}\boldsymbol{\Delta}^{-1}/\mu_F)^{\mathsf{T}} \\
&=\; (\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\Delta})(\boldsymbol{\Delta}^{-1}\mathrm{e}^{\boldsymbol{T}^{\mathsf{T}}t}\boldsymbol{\Delta})(\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}^{\mathsf{T}}) \\
&=\; \boldsymbol{t}^{\mathsf{T}}\mathrm{e}^{\boldsymbol{T}^{\mathsf{T}}t}\boldsymbol{\alpha}^{\mathsf{T}} \;=\; (\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{t})^{\mathsf{T}} \;=\; \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{t}
\end{aligned}$$

(any 1×1 matrix is symmetric!) shows that the two representations lead to the same density.

For obvious reasons, we refer to $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{T}})$ as the *time–reversed representation* of F. It becomes important in IV.3–4.

Problems

5.1 Doublecheck using Examples 5.5, 5.6 that the renewal density for E_2 with density $\delta^2 x \mathrm{e}^{-\delta x}$ is $u(t) = \frac{\delta}{2}\left(1 - \mathrm{e}^{-2\delta t}\right)$.

5.2 Show that the renewal density for H_2 with density $\alpha_1 \mathrm{e}^{-\delta_1 x} + \alpha_2 \mathrm{e}^{-\delta_2 x}$ is

$$u(t) \;=\; \frac{\delta_1\delta_2}{\delta_1\alpha_2 + \delta_2\alpha_1} + \mathrm{e}^{-(\delta_1\alpha_2+\delta_2\alpha_1)t}\frac{(\delta_1 - \delta_2)^2\alpha_1\alpha_2}{\delta_1\alpha_2 + \delta_2\alpha_1}\;.$$

5.3 Let F be phase–type. Show that deleting states $r \in E$ such that $\mathbb{P}_{\boldsymbol{\alpha}}(J_t = r \text{ for some } t < \zeta) = 0$ leads to a representation with $\boldsymbol{T} + \boldsymbol{t\alpha}$ irreducible.

5.4 Verify algebraically that $\widetilde{\boldsymbol{t}} = -\widetilde{\boldsymbol{T}}\mathbf{1}$.

Notes For an additional explicit example in renewal theory, see V.2.8; references to the area can be found in Asmussen and Bladt (1996).

6 Lindley Processes

By a *Lindley process*, we understand a discrete time process of the form

$$W_0 = w, \quad W_{n+1} = (W_n + X_n)^+, \; n = 0, 1, \ldots, \tag{6.1}$$

(x^+ means $\max(x, 0)$) where $w \geq 0$ and $X_0, X_1, \ldots$ are i.i.d. (in general having both positive and negative values), say with common distribution F. Equivalently, the process may be described as a Markov chain on $E = [0, \infty)$ with transition kernel given by

$$\begin{aligned} P(w, [0, m]) &= \mathbb{P}\big(W_1 \leq m \,\big|\, W_0 = w\big) = \mathbb{P}\big((w + X_0)^+ \leq m\big) \\ &= \mathbb{P}(w + X_0 \leq m) = F(m - w), \quad w, m \geq 0. \end{aligned} \tag{6.2}$$

If F is lattice concentrated on $\{0, \pm h, \pm 2h, \ldots\}$ and we consider only initial values of the form $w = kh$, then the state space may be reduced to $\{0, h, 2h, \ldots\}$.

The interest in the Lindley process stems classically from the way it comes up in the $GI/G/1$ queue (see the basic Example 6.1 below), but it is quite common that in a particular queueing model one or more of the processes of interest may be related to a process that is Lindley or at least of a somewhat similar structure. One example has already been given in I.5.7, one more follows below in Example 6.2, and further examples are in the Problems and (in continuous time) Section 7.

Example 6.1 Consider the $GI/G/1$ queue and let as in Section 1d W_n be the waiting time of customer $n = 0, 1, \ldots$. What is the sample path relation between W_n, W_{n+1}? Say that customer n arrives at time t and customer $n+1$ at $t + T_n$. The residual work in the system is W_n just before t, $W_n + U_n$ just after t (recall that U_n is the service time of n) and W_{n+1} just before $t + T_n$. Since the residual work decreases at a unit linear rate in between arrivals so long as it is positive, W_{n+1} will be $W_n + U_n - T_n$ when this quantity is ≥ 0 and 0 when it is ≤ 0 (a graphical illustration of the argument is contained in Fig. 1.4(c) with $n = 2$). Hence (6.1) holds with $X_n = U_n - T_n$ and clearly the X_n are i.i.d. For example in the $M/M/1$ case $\mathbb{P}(U_n > u) = \mathrm{e}^{-\delta u}$, $\mathbb{P}(T_n > t) = \mathrm{e}^{-\beta t}$, it is readily seen that F is the *doubly exponential* or *Laplace distribution* with density

$$f(x) = \begin{cases} \frac{\beta\delta}{\beta+\delta}\mathrm{e}^{-\delta x} & x \geq 0 \\ \frac{\beta\delta}{\beta+\delta}\mathrm{e}^{\beta x} & x \leq 0 \end{cases}. \qquad \square$$

Example 6.2 This is the embedded Markov chain in $GI/M/1$ mentioned in Section 1d, i.e. W_n is the number of customers just before the arrival of customer n. Let A denote the interarrival distibution and δ the service intensity. We may think of service events being given in terms of a Poisson process with intensity δ, such that an event in the Poisson process corresponds to a customer being served if the queue is nonempty and is just dummy otherwise. To describe the relation between W_n and W_{n+1}, let K_n be the number of Poisson events in the interval between arrivals of customers n and $n+1$, and define

$$q_k \;=\; \mathbb{P}(K_n = k) \;=\; \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^k}{k!}\,A(\mathrm{d}t). \tag{6.3}$$

Then clearly $W_n + 1$ customers are present just after the arrival of n and $W_{n+1} = (W_n + 1 - K_n)^+$ just before the arrival of $n+1$. Thus (6.1) holds with $X_n = 1 - K_n$ (clearly, the X_n are i.i.d.). We have $E = \mathbb{N}$, and letting $r_n = q_{n+1} + q_{n+2} + \cdots$, the transition matrix of $\{W_n\}$ is easily seen to be

$$\begin{pmatrix} r_0 & q_0 & 0 & 0 & \ldots \\ r_1 & q_1 & q_0 & 0 & \\ r_2 & q_2 & q_1 & q_0 & \\ \vdots & & & & \ddots \end{pmatrix}. \tag{6.4}$$

□

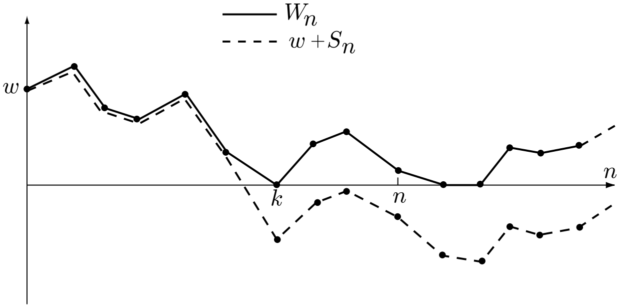

Figure 6.1

Now define $S_0 = 0$, $S_n = X_0 + \cdots + X_{n-1}$. Then (6.1) reflects that the Lindley process $\{W_n\}$ has the same transition mechanism as the random walk $\{S_n\}$ except when the random walk crosses from positive to negative values (the Lindley process then stays at 0). The relation is illustrated in Fig. 6.1. It is actually typical for many queueing processes that they are nonnegative but can be described by modifications at (or near) 0 of a process on the whole line (the *netput process*) with some basic simple structure (such modifications may well be more complex than in the present case).

Exploiting the relation between the paths of $\{W_n\}$ and $\{S_n\}$ even further yields:

Proposition 6.3 $W_n = \max\bigl(W_0 + S_n, S_n - S_1, \ldots, S_n - S_{n-1}, 0\bigr)$.

Proof. By (6.1), the increments of $\{W_n\}$ are at least those of $\{S_n\}$ so that

$$W_n - W_{n-k} \ \geq\ S_n - S_{n-k}, \quad k = 0, \ldots, n. \tag{6.5}$$

Letting $k = n$ yields $W_n \geq W_0 + S_n$ and using $W_{n-k} \geq 0$ we get $W_n \geq S_n - S_{n-k}$, proving $W_n \geq \max(\ldots)$. For the converse, we shall show that either $W_n = W_0 + S_n$ or $W_n = S_n - S_{n-k}$ for some k. The first case occurs apparently if $W_0 + S_k \geq 0$ for all $k \leq n$. Otherwise, $W_\ell = 0$ for some $\ell \leq n$, and letting k be the last such ℓ, (6.1) yields $W_n = S_n - S_{n-k}$; see Fig. 6.1. □

Now define $M_n = \max_{0 \leq k \leq n} S_k$, $M = \max_{0 \leq k < \infty} S_k$. Since the distribution of $\bigl(S_n, S_n - S_1, \ldots, S_n - S_{n-1}, 0\bigr)$ is the same as the distribution of $\bigl(S_n, S_{n-1}, \ldots, S_1, S_0 = 0\bigr)$, we get:

Corollary 6.4 $W_n \overset{\mathscr{D}}{=} \max(W_0 + S_n, M_{n-1})$. *In particular, if* $W_0 = 0$, *then* $W_n \overset{\mathscr{D}}{=} M_n$.

It should be noted that this holds in the sense of one–dimensional distributions only and not processes. For example in the case $W_0 = 0$, the paths of $\{M_n\}$ are nondecreasing, those of $\{W_n\}$ not.

Suppose now $\mathbb{E}|X_n| < \infty$ and write $\mu = \mathbb{E}X_n$.

Corollary 6.5 *If* $\mu < 0$, *then* $M < \infty$ *a.s. and* $W_n \xrightarrow{\mathscr{D}} M$ (and in t.v.).

Proof. From $S_n/n \overset{\text{a.s.}}{\to} \mu$ we have $S_n \overset{\text{a.s.}}{\to} -\infty$. This implies in particular $M < \infty$. Also $W_0 + S_n \overset{\text{a.s.}}{\to} -\infty$ and $M_n \uparrow M$ a.s. and in distribution. Thus $\max(W_0 + S_n, M_{n-1}) = M_{n-1}$ eventually and $W_n \xrightarrow{\mathscr{D}} M$ follows. □

For the limiting behaviour for $\mu \geq 0$, see Problem 6.6.

Corollary 6.6 *If* $\mu < 0$, *then* $M \overset{\mathscr{D}}{=} (M + X)^+$, *where* X *is independent of* M *with distribution* F. *Furthermore* $H(m) = \mathbb{P}(M \leq m)$ *is the unique distribution function on* $[0, \infty)$ *which solves Lindley's integral equation*

$$H(m) \ =\ \int_{-\infty}^{m} H(m - x)\, F(\mathrm{d}x), \quad m \geq 0. \tag{6.6}$$

Proof. The first statement can be proved by a limiting argument or from

$$\begin{aligned}(M + X)^+ \ &=\ \max(0, M + X) \ =\ \max\bigl(0, X, X + X_0, X + X_0 + X_1, \ldots\bigr) \\ &\overset{\mathscr{D}}{=}\ \max\bigl(0, X_0, X_0 + X_1, X_0 + X_1 + X_2, \ldots\bigr) \ =\ M.\end{aligned}$$

Furthermore the r.h.s. of (6.6) is just $\mathbb{P}(M + X \leq m) = \mathbb{P}\bigl((M + X)^+ \leq m\bigr)$ evaluated by conditioning upon $X = x$. Thus (6.6) is equivalent to $M \overset{\mathscr{D}}{=} (M + X)^+$, i.e. H being stationary for $\{W_n\}$. Thus if H_1, H_2 both solve (6.6), we may consider the two stationary chains with initial distributions

H_1, resp. H_2. From the fact that they both converge in distribution to M we conclude that $H_1 = H_2$. □

It is nontrivial even in such simple models as Example 6.2 or the doubly exponential $M/M/1$ case in Example 6.1 to derive the distribution H of M, and this is in fact the subject of the detailed investigations in VIII.5 (of course, given a trial solution H, one can just check whether H solves Lindley's integral equation).

Problems

6.1 Let $\{A_n\}_0^\infty$, $\{B_n\}_0^\infty$ be independent sequences of i.i.d. r.v.'s and define $W_0 = B_0$, $W_{n+1} = (W_n - A_n)^+ + B_{n+1}$. Show that $\{W_n - B_n\}$ is a Lindley process corresponding to $X_n = B_n - A_n$. Show that if $\mathbb{E}B_n < \mathbb{E}A_n$, then in the limit W_n is distributed as $M + B$ where $M = \sup S_n$ and B is an independent r.v. distributed as B_n.

6.2 Consider the *fixed–cycle traffic light*, with the cycles divided into the *green period* where customers (say cars or pedestrians) can pass and the *red period* where they cannot. Let W_n^G be the number of customers just after the start of the nth green period and B_n^G the number of customers arriving during the nth green period (similar conventions define W_n^R, B_n^G). Assuming that the maximal number of customers which can pass during a green period is some fixed number p, show that

$$W_{n+1}^R = \Bigl(W_n^R + B_{n+1}^G + B_n^R - p\Bigr)^+, \quad W_{n+1}^G = \Bigl(W_n^G + B_n^G - p\Bigr)^+ + B_n^R.$$

6.3 Consider the $M/G/1$ queue and let W_n be the queue length just after the nth departure, B_n the number of customers arriving during the nth service period. Show that $W_{n+1} = (W_n - 1)^+ + B_{n+1}$, and find the distribution of B_n in terms of a formula similar to (6.3). Show also in the $M/M/1$ case with $\rho < 1$ that the stationary distribution of $\{W_n\}$ is given by $\pi_n = (1-\rho)\rho^n$.

6.4 Assume that F has negative mean and density $p\delta e^{-\delta x}$ on $(0,\infty)$ (clearly, $p = \overline{F}(0)$). Let γ satisfy $\int_{-\infty}^{\infty} e^{\gamma x} F(dx) = 1$. Show by direct calculation that $H(m) = 1 - (1-\gamma/\delta)e^{-\gamma m}$ is the unique solution to Lindley's integral equation. Show hereby that the steady–state $GI/M/1$ waiting time distribution is of this form.

6.5 Compute the matrix (6.4) for the $M/M/1$ case and show that $\pi_n = (1-\rho)\rho^n$ is stationary when $\rho < 1$.

6.6 Consider a Lindley process with $\mu > 0$. Show that $W_n/n \overset{\text{a.s.}}{\to} \mu$. Show also that the process is a null recurrent Markov chain in the lattice case with $\mu = 0$. [*Hint:* Let $\tau = \inf\{n : S_n \le 0\}$ and show that $\mathbb{E}\tau < \infty$ contradicts Wald's identity.]

Notes Among textbooks with systematic discussion of Lindley processes, we mention in particular Feller (1971,VI.9) and Borovkov (1976). A Lindley process is most naturally viewed as a reflected random walk, and we return to a more systematic study of reflection allowing also for dependence in IX.2 (a preliminary discussion of the continuous time case is in the next section).

Lindley's integral equation is of *Wiener–Hopf type*,; see further the Notes to VIII.3.

The Laplace distribution in Example 6.1 has received considerable attention outside of queueing theory; see Kotz *et al.* (2001).

7 A First Look at Reflected Lévy Processes

A natural way to define a random walk in continuous time is as a process $\{S_t\}_{t\geq 0}$ with $S_0 = 0$ and with stationary independent increments. We will return to a study of such processes, usually referred to as *Lévy processes*, in IX.1. In this chapter, it will suffice to note that basic examples are a linear deterministic drift ($S_t = \theta t$), standard Brownian motion and a compound Poisson process, and independent sums of such processes. If the mean is well defined, it must be linear, $\mathbb{E}S_t = \mu t$.

How to define a reflected version $\{V_t\}_{t\geq 0}$ of $\{S_t\}$ in continuous time is less obvious than in discrete time. Also this topic is studied more systematically later; see IX.2. The definition used there is the continuous–time analogue of Proposition 6.3,

$$V_t \;=\; (V_0 + S_t) \vee \max_{0\leq s\leq t}(S_t - S_s)\,; \tag{7.1}$$

when $x = V_0$ is of importance, we will write $V_t = V_t(x)$. The following results are obtained as special cases of results to be shown in IX.2 (for the strong Markov property in Proposition 7.1, combine with I.8.3). Define $M_T = \sup_{0\leq t\leq T} S_t$, $M = \sup_{0\leq t<\infty} S_t$.

Proposition 7.1 *$\{V_t\}$ is a strong Markov process.*

Corollary 7.2 *$V_T \overset{\mathscr{D}}{=} (V_0+S_T)\wedge M_T$. If $\mu < 0$, then $M < \infty$ and $V_T \to M$ in total variation.*

Proposition 7.3 *Define $\omega = \inf\{t > 0 : V_0 + S_t \leq 0\}$. Then also $\omega = \inf\{t > 0 : V_t = 0\}$, and $V_t = V_0 + S_t$ for $t < \omega$.*

The content of this last result is that $\{V_t\}$ evolves as $\{S_t\}$ until the first hitting time of 0, so that in examples the crux is to describe the behaviour starting from $V_0 = 0$ or, equivalently, in Markov state $x = 0$.

Example 7.4 Consider a compound Poisson process of the form $S_t = N_t^{(\beta)} - N_t^{(\delta)}$ where $\{N_t^{(\beta)}\}$, $\{N_t^{(\delta)}\}$ are independent Poisson processes with intensities β, resp. δ. The reflection then means that jumps of $\{N_t^{(\delta)}\}$ are ignored when $V_t = 0$. Thus, $\{V_t\}$ is a Markov process on $\mathbb{N}$ where the only nonzero off–diagonal intensities are $\lambda(i, i-1) = \delta$, $i = 1, 2, \ldots$, $\lambda(i, i+1) = \beta$, $i = 0, 1, \ldots$, and we recognize $\{V_t\}$ as the $M/M/1$ queue length process in Section 3a. □

Example 7.5 Consider a compound Poisson process with only positive jumps and a negative drift, $S_t = \sum_1^{N_t} U_i - t$ where $\{N_t\}$ is a Poisson process with intensity β, and $U_1, U_2, \ldots$ are i.i.d. with common distribution B (concentrated on $(0, \infty)$) and independent of $\{N_t\}$. The reflection then means that the downward drift at unit rate is cut off when $V_t = 0$. Thus, $\{V_t\}$ has the same upward jumps as $\{S_t\}$ and a downward drift at unit rate in states $x > 0$ so that we recognize $\{V_t\}$ as the $M/G/1$ workload process. □

In the next example as well as at other places in the book, we shall need:

Proposition 7.6 *If $\{S_t\}$ is standard Brownian motion, then the joint distribution of S_t and M_t is given by*

$$\mathbb{P}(S_t \le x - y,\, M_t \ge x) \;=\; \mathbb{P}(S_t \ge x + y) \tag{7.2}$$

for $x, y \ge 0$. Also $M_t \overset{\mathscr{D}}{=} |S_t|$.

Proof. Define $\tau(x) = \inf\{t > 0 : S_t \ge x\}$. Since $M_t \ge x$ is equivalent to $\tau(x) \le t$ and is automatic if $S_t \ge x + y$, we may rewrite (7.2) as

$$\mathbb{P}(S_t \le x - y,\, \tau(x) \le t) \;=\; \mathbb{P}(S_t \ge x + y,\, \tau(x) \le t).$$

The truth of this follows by the *reflection principle*, which states that Brownian motion, being symmetric ($S_t \overset{\mathscr{D}}{=} -S_t$), is equally likely to proceed from $S_{\tau(x)} = x$ to levels $\ge x + y$ or $\le x - y$ within $s = t - \tau(x)$ time units (here we have also used the strong Markov property). We then get

$$\begin{aligned}\mathbb{P}(M_t \ge x) \;&=\; \mathbb{P}(S_t \le x \le M_t) + \mathbb{P}(S_t \ge x) \;=\; \mathbb{P}(S_t \ge x) + \mathbb{P}(S_t > x)\\ &=\; 2\mathbb{P}(S_t \ge x) \;=\; \mathbb{P}(|S_t| \ge x),\end{aligned}$$

using (7.2) in the second step. □

Example 7.7 Assume that $\{S_t\}$ is Brownian motion with zero drift and unit variance. Then $\{V_t(x)\}$ is reflected standard Brownian motion starting from x and has the same distribution as $\{|x + S_t|\}$. In fact, since both processes are strong Markov and evolve as $\{x + S_t\}$ until the first hitting time of 0, it suffices to show that the transition functions starting from $x = 0$ are the same, i.e. that $V_T(0) \overset{\mathscr{D}}{=} |S_T|$. However, $V_T(0) \overset{\mathscr{D}}{=} M_T$ by Corollary 7.2 and $M_T \overset{\mathscr{D}}{=} |S_T|$ by Proposition 7.6. □

Problems

7.1 Show that the number of stages in $M/E_k/1$ is a reflected Lévy process.

Notes We return to reflected Lévy processes in IX.2. Aspects of the topic are closely related to dams and storage processes; see Chapter XIV, Prabhu (1980) and references therein.

8 Time–Dependent Properties of $M/M/1$

8a. The Doubly Infinite Queue and Its Maximum
8b. The Transition Probabilities
8c. The Busy Period Distribution
8d. Transform Methods
8e. The Relaxation Time

8a The Doubly Infinite Queue and Its Maximum

The key to our more refined study of the $M/M/1$ queue length process $\{X_t\}_{t\geq 0}$ is the Lindley process representation in Example 7.4. Combining with Corollary 7.2, we have:

Proposition 8.1 *The distribution of the $M/M/1$ queue length X_t at time t given $X_0 = i$ is that of* $\max(i+S_t, M_t)$ *where $S_t = B_t - D_t$ is the difference between two independent Poisson processes with intensities β and δ, and $M_t = \sup_{0\leq v\leq t} S_v$.*

The process $\{S_t\}_{t\geq 0}$ is frequently denoted as the *doubly infinite queue.* It models, for example, a queueing situation with taxis and passengers in front of a railway station, with B_t, D_t denoting the number of passengers, resp. taxis arriving before t. Thus if $S_t > 0$ at time t, there is a queue of length S_t of passengers, whereas if $S_t < 0$ there is a queue of length $-S_t$ of taxis.

Letting $M = \sup_{0\leq t<\infty} S_t$, the following simple observation will be useful in the following:

Proposition 8.2 *Let $\rho = \beta/\delta$. Then a.s.:* (i) *$S_t \to -\infty$, $M < \infty$ when $\rho < 1$;* (ii) *$S_t \to +\infty$, $M = \infty$ when $\rho > 1$;* (iii) *$\overline{\lim}_{t\to\infty} S_t = +\infty$, $\underline{\lim}_{t\to\infty} S_t = -\infty$ when $\rho = 1$.*

Proof. Let T_n be the value of S_t just after the nth jump, $T_0 = 0$. Then $\{T_n\}$ is a Bernoulli random walk with

$$p = \frac{\beta}{\beta+\delta} = \frac{\rho}{1+\rho}, \qquad q = \frac{\delta}{\beta+\delta} = \frac{1}{1+\rho}.$$

Hence if $\rho < 1$, $\mathbb{E}T_1 = p - q < 0$ and by the LLN $T_n/n \overset{\text{a.s.}}{\to} p - q$, implying $T_n \overset{\text{a.s.}}{\to} -\infty$, $S_t \overset{\text{a.s.}}{\to} -\infty$ and hence $M < \infty$. The case $\rho > 1$ is treated similarly. The case $\rho = 1$ is slightly more intricate and can be treated either by appealing to a general random walk result given in VIII.2.4 or by a direct argument (see Problem 8.1). □

Despite the simple relation to the Poisson distribution, the explicit form of the point probabilities of S_n is not elementary. Define the *modified Bessel* *function of integer order* $n \in \mathbb{Z}$ by

$$I_n(x) \;=\; \sum_{k=0}^{\infty} \frac{(x/2)^{n+2k}}{k!(n+k)!}, \quad I_{-n}(x) \;=\; I_n(x), \quad n \in \mathbb{N},$$

and let $\mu = \sqrt{\beta\delta}$, $\rho = \beta/\delta$. The argument of I_n will be $x = 2\mu t$ throughout, so unless otherwise stated I_n just denotes $I_n(2\mu t)$, and we shall let $\iota_n = \mathrm{e}^{-(\beta+\delta)t}\rho^{n/2}I_n$, $n \in \mathbb{Z}$, so that

$$\iota_{-n} \;=\; \rho^{-n}\iota_n, \quad n \in \mathbb{Z}. \tag{8.1}$$

As a technical tool (a particular case of the change of measure technique studied in Chapter XIII), we shall use a process with β, δ both replaced by μ. This is denoted by $\mathbb{P}_0$ and has traffic intensity 1 and the same value of μ, and we have

$$\begin{aligned}\mathbb{P}(B_t = \ell, D_t = k) \;&=\; \mathrm{e}^{-(\beta+\delta)t}\frac{(\beta t)^\ell}{\ell!}\frac{(\delta t)^k}{k!} \\ &=\; \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{(\ell-k)/2}\mathbb{P}_0(B_t = \ell, D_t = k). \quad (8.2)\end{aligned}$$

Proposition 8.3 $\mathbb{P}(S_t = n) \;=\; \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{n/2}\mathbb{P}_0(S_t = n) \;=\; \iota_n$*,* $n \in \mathbb{Z}$*.*

Proof. For $n \geq 0$ we get

$$\begin{aligned}\mathbb{P}(S_t = n) \;&=\; \sum_{k=0}^{\infty}\mathbb{P}(B_t = n+k, D_t = k) \;=\; \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{n/2}\mathbb{P}_0(S_t = n) \\ =\; \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{n/2}&\sum_{k=0}^{\infty}\mathrm{e}^{-\mu t}\frac{(\mu t)^{n+k}}{(n+k)!}\mathrm{e}^{-\mu t}\frac{(\mu t)^k}{k!} \;=\; \mathrm{e}^{-(\beta+\delta)t}\rho^{n/2}I_n \;=\iota_n.\end{aligned}$$

The case $n \leq 0$ is treated similarly or by a symmetry argument (when passing from S_t to $-S_t$, the arrival and service intensities are interchanged and ρ changed to ρ^{-1}). □

As a corollary, which will be used in formula manipulations in the following, we note the identity

$$1 \;=\; \sum_{n=-\infty}^{\infty}\iota_n\,. \tag{8.3}$$

In order to apply Proposition 8.1, we have more generally to find the joint distribution of S_t and M_t:

Lemma 8.4 *For* $n+m \geq 0$*,* $m \geq 0$

$$\mathbb{P}(S_t = n, M_t \geq n+m) \;=\; \rho^{-m}\iota_{n+2m}, \tag{8.4}$$

$$\mathbb{P}(S_t = n, M_t = n+m) \;=\; \rho^{-m}(\iota_{n+2m} - \rho^{-1}\iota_{n+2m+2}). \tag{8.5}$$

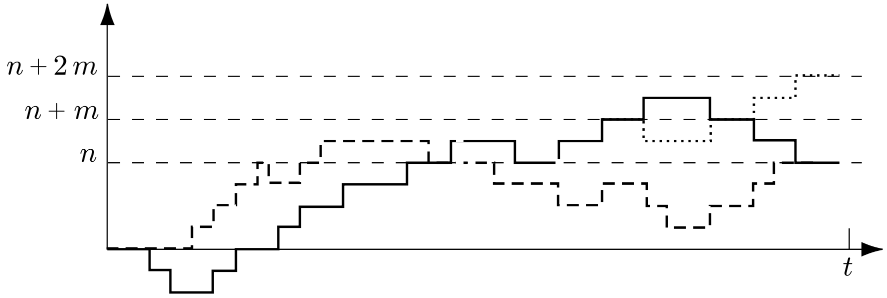

Figure 8.1 The path - - - has the same B_t and D_t as —— but a different M_t. The path · · · is obtained by reflection after $n+m$ has been reached.

Proof. Let $F = \{S_t = n, M_t \geq n+m\}$, $G_k = \{B_t = n+k, D_t = k\}$. Then conditionally upon G_k it depends solely on the order of the $n+k$ increments of S_t and the k decrements whether or not F occurs; cf. Fig. 8.1. But by well–known properties of the Poisson process, this ordering is determined by two independent samples of sizes $n+k$, k from the uniform distribution on $[0,t]$. Hence $\mathbb{P}(F|G_k)$ is independent of the intensities and in particular, $\mathbb{P}(F|G_k) = \mathbb{P}_0(F|G_k)$ so that using (8.2) we get

$$\begin{aligned}\mathbb{P}F &= \sum_{k\geq 0:\, n+k\geq 0} \mathbb{P}(FG_k) = \sum_{k\geq 0:\, n+k\geq 0} \frac{\mathbb{P}G_k}{\mathbb{P}_0 G_k}\mathbb{P}_0(FG_k) \\ &= \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{n/2}\mathbb{P}_0 F.\end{aligned}$$

Now since the $\mathbb{P}_0$–process is symmetric, it follows by the reflection principle just as for Brownian motion in Proposition 7.6 (cf. also Fig. 8.1) that

$$\mathbb{P}_0 F = \mathbb{P}_0(S_t = n+2m) = \mathrm{e}^{-2\mu t} I_{n+2m}.$$

Hence

$$\mathbb{P}F = \mathrm{e}^{(2\mu-\beta-\delta)t}\rho^{n/2}\mathrm{e}^{-2\mu t} I_{n+2m} = \rho^{-m}\iota_{n+2m}$$

and (8.4) follows. Finally (8.5) is a consequence of (8.4). □

8b The Transition Probabilities

Proposition 8.1 and Lemma 8.4 solve in principle the problem of evaluating

$$p_{ij}^t = \mathbb{P}(X_t = j \,|\, X_0 = i) = \mathbb{P}F \quad \text{where } F = \{\max(i+S_t, M_t) = j\},$$

and it only remains to collect terms. Now F is the disjoint union of

$$F_1 = \{S_t = j-i \leq M_t \leq j\} \quad \text{and} \quad F_2 = \{S_t < j-i, M_t = j\}$$

But

$$\mathbb{P}F_1 = \sum_{m=0}^{i} \mathbb{P}(S_t = j-i, M_t = j-i+m)$$

$$\begin{aligned}
&= \sum_{m=0}^{i} \rho^{-m}\bigl(\iota_{j-i+2m} - \rho^{-1}\iota_{j-i+2m+2}\bigr) \\
&= \iota_{j-i} - \rho^{-i-1}\iota_{j+i+2} \;=\; \iota_{j-i} - \rho^{j+1}\iota_{-j-i-2}, \\
\mathbb{P}F_2 &= \sum_{n=-\infty}^{j-i-1} \rho^{n-j}\bigl(\iota_{n+2(j-n)} - \rho^{-1}\iota_{n+2(j-n)+2}\bigr) \\
&= \rho^{j}\sum_{n=-\infty}^{-j-i-1} \iota_n \;-\; \rho^{j+1}\sum_{n=-\infty}^{-j-i-1} \iota_{n-2} \\
&= \rho^{-i-1}\iota_{i+j+1} \;+\; \rho^{j}\sum_{n=-\infty}^{-j-i-2} \iota_n \;-\; \rho^{j+1}\sum_{n=-\infty}^{-j-i-3} \iota_n
\end{aligned}$$

where we have used (8.1) repeatedly. Adding these two expressions, we obtain:

Theorem 8.5 *In the $M/M/1$ queue with $0 < \rho < \infty$, $p^t_{ij} = \mathbb{P}(X_t = j \,|\, X_0 = i)$ is given by*

$$\iota_{j-i} \;+\; \rho^{-i-1}\iota_{i+j+1} \;+\; (1-\rho)\rho^{j}\sum_{n=-\infty}^{-j-i-2} \iota_n. \tag{8.6}$$

Analytical manipulations or alternative derivations provide a number of alternative expressions for p^t_{ij}, for example

$$\begin{aligned}
p^t_{ij} &= (1-\rho)\rho^{j} \;+\; \iota_{j-i} \;-\; \rho^{(j-i)/2}J_1 & (8.7)\\
&= (1-\rho)\rho^{j}I(\rho<1) \;+\; J_2 & (8.8)
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \int_t^{\infty} \mathrm{e}^{-(\beta+\delta)s}\bigl[I_{i+j}(2\mu s) - 2\rho^{1/2}I_{i+j+1}(2\mu s) + \beta I_{i+j+2}(2\mu s)\bigr]\,\mathrm{d}s, \\
J_2 &= \frac{2\mathrm{e}^{-(\beta+\delta)t}\rho^{(j-i)/2}}{\pi}\int_0^{\pi} \frac{\mathrm{e}^{2\mu t\cos\theta}}{1-2\rho^{1/2}\cos\theta+\rho}g_i(\theta)g_j(\theta)\,\mathrm{d}\theta
\end{aligned}$$

with $g_i(\theta) = \sin i\theta - \rho^{1/2}\sin(i+1)\theta$ (see Cohen, 1982, p. 178, for (8.7) and Takács, 1962, p. 23, for (8.8)). That trigonometric functions are involved may be understood from the standard analytical identity

$$I_n(t) \;=\; \frac{1}{\pi}\int_0^{\pi} \mathrm{e}^{t\cos\theta}\cos n\theta\,\mathrm{d}\theta.$$

The message of these formulas, as well as those for the busy–period distribution to be derived shortly, is perhaps not so much their particular form, but rather that they are extremely complicated. The $M/M/1$ queue being the very simplest queueing system, this probably suggests that time–dependent explicit solutions are in general not possible and indeed this is the case. Even the numerical evaluation on a computer requires some care and the most feasible approach may be to apply numerical integration of formulas such as (8.8), thereby avoiding manipulation of infinite sums or integrals of Bessel functions of high order.

8c The Busy Period Distribution

We first prove:

Proposition 8.6 *The first passage time* $\tau = \inf\{t > 0 : S_t = 1\}$ *of the doubly infinite queue from* 0 *to* 1 *has density*

$$f(t) \;=\; \beta \mathrm{e}^{-(\beta+\delta)t}[I_0(2\mu t) - I_2(2\mu t)] \;=\; \frac{\rho^{1/2}}{t}\mathrm{e}^{-(\beta+\delta)t} I_1(2\mu t). \tag{8.9}$$

Note that if $\rho < 1$, then by Proposition 8.2 S_t drifts to $-\infty$ and hence (8.9) is defective, $\int f = \mathbb{P}(\tau < \infty) < 1$. More precisely, from $\mathbb{P}(\tau < \infty) = \mathbb{P}(M \geq 1) = \mathbb{P}_e(X_t \geq 1)$ it follows that $\int f = \rho$.

Proof. Using (8.5) we get

$$\begin{aligned} \mathbb{P}(\tau > t) \;&=\; \mathbb{P}(M_t = 0) \;=\; \sum_{n=-\infty}^{0} \mathbb{P}(S_t = n, M_t = 0) \\ &=\; \sum_{n=-\infty}^{0} \rho^n(\iota_{-n} - \rho^{-1}\iota_{2-n}) \;=\; C_0(t) - \rho C_2(t) \end{aligned}$$

where

$$C_N(t) \;=\; \sum_{n=N}^{\infty} \mathrm{e}^{-(\beta+\delta)t}\rho^{-n/2} I_n(2\mu t).$$

Hence $f(t) = \rho C_2'(t) - C_0'(t)$ and to evaluate the derivatives, we need the formulas

$$I_0'(t) = I_1(t), \quad I_n'(t) = \frac{1}{2}[I_{n-1}(t) + I_{n+1}(t)], \; n = 1, 2, \ldots, \tag{8.10}$$

$$I_{n-1}(t) - I_{n+1}(t) = \frac{2n}{t} I_n(t), \; n = 1, 2, \ldots, \tag{8.11}$$

which may easily be seen from the power series definition of I_n. Using (8.10), we get for $N \geq 1$ that

$$\begin{aligned} C_N'(t) \;&=\; -(\beta+\delta)C_N(t) \;+\; \sum_{n=N}^{\infty} \mathrm{e}^{-(\beta+\delta)t}\rho^{-n/2}\mu(I_{n-1} + I_{n+1}) \\ &=\; -(\beta+\delta)C_N(t) + \delta C_{N-1}(t) + \beta C_{N+1}(t) \\ &=\; \mathrm{e}^{-(\beta+\delta)t}[\delta\rho^{-(N-1)/2} I_{N-1} - \beta\rho^{-N/2} I_N]. \end{aligned}$$

Hence

$$f(t) \;=\; \rho C_2'(t) - C_1'(t) - \frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{-(\beta+\delta)t} I_0(2\mu t)\right]$$

$$
\begin{aligned}
&= \mathrm{e}^{-(\beta+\delta)t}\big[\delta\rho^{1/2}I_1 - \beta I_2 - \delta I_0 + \beta\rho^{-1/2}I_1 + (\beta+\delta)I_0 - 2\mu I_1\big] \\
&= \mathrm{e}^{-(\beta+\delta)t}\beta[I_0 - I_2],
\end{aligned}
$$

and combining with (8.11), the proof is complete. □

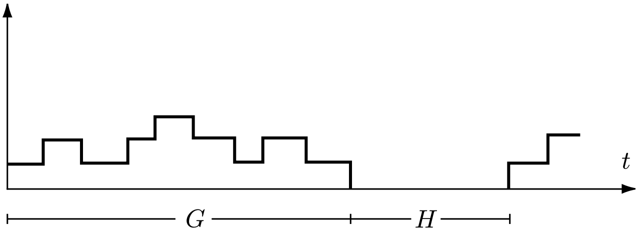

Figure 8.2

We understand the *busy period* G of the queue to be the time from when a customer enters an empty system until the system is empty again. The busy period is followed by an interval of length H where the system is empty, the *idle period*, and $G + H$ constitute the *busy cycle*; see Fig. 8.2 (the notation G, H is used only at this place). In the $M/M/1$ case, it is clear that G and H are independent and H exponentially distributed with intensity β. Furthermore, we may identify G with the time of passage of the doubly infinite queue from 0 to -1. The distribution of this follows by a symmetry argument since we just have to interchange β and δ in Proposition 8.6, and thus:

Corollary 8.7 *The busy–period distribution of the $M/M/1$ queue is given by the density*

$$
g(t) \;=\; \delta\mathrm{e}^{-(\beta+\delta)t}[I_0(2\mu t) - I_2(2\mu t)] \;=\; \frac{\rho^{-1/2}}{t}\mathrm{e}^{-(\beta+\delta)t}I_1(2\mu t). \quad (8.12)
$$

As above, g is defective for $\rho > 1$. Moments will be derived shortly.

8d Transform Methods

In some cases, where quantities such as the p_{ij}^t cannot be derived in closed analytical form, it may still be possible to find explicit expressions for transforms or double (bivariate) transforms. As an example, we quote for the present $M/M/1$ case the formula (Problem IX.3.4 or Prabhu, 1965, Chapter 1.2a)

$$
\int_0^\infty \mathrm{e}^{-\theta t}\sum_{j=0}^\infty s^j p_{ij}^t \,\mathrm{d}t \;=\; \frac{s^{i+1} - (1-s)\xi^{i+1}/(1-\xi)}{\beta(s-\xi)(\eta - s)}, \quad (8.13)
$$

where $\xi = \xi(\theta)$, $\eta = \eta(\theta)$ are the two roots of

$$
\beta z^2 - (\beta+\delta+\theta)z + \delta \;=\; 0, \quad (8.14)
$$

with $\xi \le \eta$, i.e.

$$\xi = \frac{\beta+\delta+\theta-R(\theta)}{2\beta}, \quad \eta = \frac{\beta+\delta+\theta+R(\theta)}{2\beta}, \tag{8.15}$$

where $R(\theta) = \sqrt{(\beta+\delta+\theta)^2 - 4\beta\delta}$. Classically this is proved, for example, by careful manipulation of the differential equations for the p_{ij}^t (but see IX.3 for a more elegant approach). Though more generally applicable, this method has serious drawbacks, however. In more complex cases, the expressions are even less transparent than (8.13) and their derivation may require much ingenuity. Also, even in the present case, the inversion of (8.13) is not easy no matter whether the purpose is to derive formulas like (8.6)–(8.8) or numerical computation.

We shall consider only one example where the transform calculations work out in a quite elegant fashion, namely the distribution of the busy period G and its moments. We first note that since $\{B_t\}$, $\{D_t\}$ are (independent) Lévy processes with the so–called *Lévy exponents* (see further IX.1) $\log \mathbb{E}e^{\alpha B_1}$, $\log \mathbb{E}e^{\alpha D_1}$ given by $\beta(e^\alpha - 1)$, resp. $\delta(e^\alpha - 1)$, then also $\{S_t\} = \{B_t - D_t\}$ is a Lévy process with Lévy exponent

$$\kappa(\alpha) = \log \mathbb{E}e^{\alpha S_1} = \log\bigl[\mathbb{E}e^{\alpha B_1}\cdot \mathbb{E}e^{-\alpha D_1}\bigr] = \beta(e^\alpha-1)+\delta(e^{-\alpha}-1), \tag{8.16}$$

and we have:

Lemma 8.8 *For any α, $\{Y_t\} = \bigl\{e^{\alpha S_t - t\kappa(\alpha)}\bigr\}$ is a continuous–time martingale, and we have for $\rho \le 1$ and $\kappa'(\alpha) \le 0$* (i.e. $\alpha \le -\log\rho/2$) *that*

$$1 = \mathbb{E}Y_0 = \mathbb{E}Y_G = e^{-\alpha}\mathbb{E}e^{-G\kappa(\alpha)} \tag{8.17}$$

Proof. The martingale property follows from

$$\begin{aligned}
&\mathbb{E}\bigl[\exp\{\alpha S_{t+v} - (t+v)\kappa(\alpha)\}\bigm| (S_u)_{0\le u\le t}\bigr] \\
&\quad= \exp\{\alpha S_t - t\kappa(\alpha)\}\,\mathbb{E}\bigl[\exp\{\alpha(S_{t+v}-S_t) - v\kappa(\alpha)\}\bigm| (S_u)_{0\le u\le t}\bigr] \\
&\quad= \exp\{\alpha S_t - t\kappa(\alpha)\}\,\mathbb{E}\exp\{\alpha S_v - v\kappa(\alpha)\} = \exp\{\alpha S_t - t\kappa(\alpha)\}.
\end{aligned}$$

Now G, being the first passage time from 0 to -1 of $\{S_t\}$, is clearly a stopping time and the assumption $\rho \le 1$ ensures $G < \infty$ a.s. Since clearly $Y_0 = 1$ and $S_G = -1$, the formula (8.17) thus follows by verifying the conditions of a suitable version of the optional stopping theorem given in XIII.4.2(b) (the discussion there also explains why the condition $\kappa'(\alpha) \le 0$ is involved). □

Remark 8.9 It is clear that the proof of Lemma 8.8 shows that $\{Y_t\}$ is a martingale for any Lévy process. This martingale is known as the *Wald martingale* and will be used at a number of later occasions. Its analogue for a discrete time random walk $\{S_n\}$ is $Y_n = e^{\alpha S_n}/\widehat{F}[\alpha]^n$ where $\widehat{F}[\alpha] = \mathbb{E}e^{\alpha S_1}$. □

Proposition 8.10 *For $\rho < 1$, the Laplace transform of the $M/M/1$ busy period G is given by*

$$\mathbb{E}e^{-\theta G} \;=\; \xi(\theta) \;=\; \frac{1}{2\beta}\Bigl(\beta+\delta+\theta-\sqrt{(\beta+\delta+\theta)^2-4\beta\delta}\Bigr) \qquad (8.18)$$

for $\theta \geq 0$, cf. (8.15). *In particular, the mean and variance are given by*

$$\mathbb{E}G \;=\; -\xi'(0) \;=\; \frac{1}{\delta(1-\rho)},\ \mathbb{V}ar\,G \;=\; \xi''(0)-\xi'(0)^2 \;=\; \frac{1+\rho}{\delta^2(1-\rho)^3}. \quad (8.19)$$

Proof. Let $\widetilde{\xi}(\theta) = \mathbb{E}e^{-\theta G}$. It follows readily from (8.16) that $\kappa(\alpha)$ decreases monotonically from ∞ to 0 on $(-\infty, 0]$. Hence for $\theta \geq 0$ we may find $\alpha \leq 0$ such that $\kappa(\alpha) = \theta$, and (8.17) then yields $\widetilde{\xi}(\theta) = e^{\alpha}$. Now letting $\widetilde{\xi} = e^{\alpha}$ in (8.16) we get $\theta = \beta(\widetilde{\xi}-1)+\delta(\widetilde{\xi}^{-1}-1)$ which after some algebra shows that $\widetilde{\xi}$ solves the quadratic (8.14). But in (8.15) we have $\eta(\theta) > 1$, $\theta > 0$, hence $\widetilde{\xi} = \xi$ since $\widetilde{\xi} < 1$. Finally (8.19) follows by some more algebra from $R(0) = \delta-\beta$, $R'(\theta) = (\beta+\delta+\theta)/R(\theta)$ and

$$\xi'(\theta) \;=\; \frac{1}{2\beta}\left(1-\frac{\beta+\delta+\theta}{R(\theta)}\right),\quad \xi''(\theta) \;=\; \frac{(\beta+\delta+\theta)^2-R(\theta)^2}{2\beta R(\theta)^3}. \qquad \square$$

Distributional properties of the busy cycle follow easily as a corollary, since we just have to convolve the busy–period distribution with the distribution of the idle period, which is exponential with intensity β. For example, the mean busy cycle is

$$\mathbb{E}G+\mathbb{E}H \;=\; \frac{1}{\delta(1-\rho)}+\frac{1}{\beta} \;=\; \frac{1}{\beta(1-\rho)}\,.$$

8e The Relaxation Time

Suppose $\rho < 1$. We shall consider the question on the rate of convergence of p_{ij}^t, as given by the formulas of Section 8b, to its limiting value $(1-\rho)\rho^j$. To this end, we need the asymptotic properties of the Bessel functions:

Lemma 8.11 *As $t \to \infty$,*

$$I_n(t) \;=\; \frac{e^t}{\sqrt{2\pi}}\left\{t^{-1/2}-t^{-3/2}\frac{4n^2-1}{8}\right\} \;+\; n^4t^{-3/2}e^t o(1)\,,$$

where the $o(1)$ *terms here and in the proof are uniform in n.*

Proof. Letting $\beta = \delta = \mu = 1/2$ for the moment, we shall appeal to the interpretation $I_n(t) = e^t\mathbb{P}(S_t = n)$, cf. Proposition 8.3, and apply the higher–order expansion in the local CLT for lattice distributions (see Bhattacharya and Rao, 1976, p. 231, or Gnedenko and Kolmogorov, 1954, p. 241; the result is stated there for discrete time random walks but is also valid in continuous time as may be seen by the same proof or by the method of discrete skeletons; cf. A11). To this end we need the cumulants κ_r of S_1, i.e. according to (8.16)

$$\kappa_r = \kappa^{(r)}(0) = \frac{\mathrm{d}^r}{\mathrm{d}\alpha^r}\frac{1}{2}(\mathrm{e}^{\alpha}+\mathrm{e}^{-\alpha})\Big|_{\alpha=0} = \left\{\begin{array}{ll} 0 & r \text{ uneven} \\ 1 & r \text{ even} \end{array}\right. .$$

Hence if φ_r denotes the rth derivative of the standard normal density evaluated at $y_n = n/\sqrt{t}$, we have

$$\begin{aligned}
&\mathbb{P}(S_t = n) \\
&= \frac{1}{\sqrt{t}}\left\{\varphi_0 - \frac{1}{6\sqrt{t}}\frac{\kappa_3}{\kappa_2^{3/2}}\varphi_3 + \frac{1}{24t}\left[\frac{\kappa_4}{\kappa_2^2}\varphi_4 + \frac{\kappa_3^2}{3\kappa_2^3}\varphi_6\right]\right\} + t^{-3/2}\mathrm{o}(1) \\
&= \frac{1}{\sqrt{t}}\left\{\varphi_0 + \frac{\varphi_4}{24t}\right\} + t^{-3/2}\mathrm{o}(1) \\
&= \frac{\varphi(y_n)}{\sqrt{t}}\left\{1 + \frac{y_n^4 - 6y_n^2 + 3}{24t}\right\} + t^{-3/2}\mathrm{o}(1) \\
&= \frac{1 - n^2/2t + n^4t^{-2}\mathrm{o}(1)}{\sqrt{2\pi t}}\left\{1 + \frac{1}{8t} + n^4t^{-2}\mathrm{o}(1)\right\} + t^{-3/2}\mathrm{o}(1) \\
&= \frac{1}{\sqrt{2\pi t}}\left(1 - \frac{4n^2-1}{8t}\right) + n^4t^{-3/2}\mathrm{o}(1).
\end{aligned}$$

□

We can now evaluate the desired rate of convergence. Using (8.6), (8.3) and Lemma 8.11 we get

$$\begin{aligned}
(1-\rho)\rho^j - p_{ij}^t &= (1-\rho)\rho^j \sum_{n=-j-i-1}^{\infty} \iota_n - \iota_{j-i} - \rho^{-i-1}\iota_{i+j+1} \\
&= \mathrm{e}^{-(\beta+\delta)t}\Bigg\{(1-\rho)\rho^j \sum_{n=-j-i-1}^{\infty} \rho^{n/2} I_n \\
&\qquad - \rho^{(j-i)/2} I_{j-i} - \rho^{(j-i+1)/2} I_{i+j+1}\Bigg\} \\
&= \frac{\mathrm{e}^{(2\mu-\beta-\delta)t}}{\sqrt{2\pi}}\left\{\frac{C_1(i,j) - \rho^{(j-i)/2} - \rho^{(j-i+1)/2}}{(2\mu t)^{1/2}} + \frac{C_2(i,j)}{(2\mu t)^{3/2}}\right\} + \mathrm{o}(t^{-3/2}),
\end{aligned}$$

where

$$\begin{aligned}
C_1(i,j) &= (1-\rho)\rho^j \sum_{n=-j-i-1}^{\infty} \rho^{n/2} = (1-\rho)\rho^j \frac{\rho^{-(j+i+1)/2}}{1-\rho^{1/2}} \\
&= \rho^{(j-i)/2} + \rho^{(j-i+1)/2}, \\
C_2(i,j) &= (1-\rho)\rho^j \sum_{n=-j-i-1}^{\infty} \rho^{n/2}\frac{4n^2-1}{8} - \rho^{(j-i)/2}\frac{4(j-i)^2-1}{8} \\
&\qquad - \rho^{(j-i-1)/2}\frac{4(i+j+1)^2-1}{8}
\end{aligned}$$

(C_2 can be reduced, but we shall not carry out the tedious algebra). Hence the $t^{-1/2}$ term vanishes, and we have proved:

Theorem 8.12 *If $\rho < 1$, then*

$$p_{ij}^t \;=\; (1-\rho)\rho^j \;+\; \frac{\mathrm{e}^{-rt}}{4\sqrt{\pi}(\beta\delta)^{3/2}} t^{-3/2} C_2(i,j) \;+\; \mathrm{o}\Big(\frac{\mathrm{e}^{-rt}}{t^{3/2}}\Big)$$

as $t \to \infty$, where $r = \beta + \delta - 2\mu = (\sqrt{\delta} - \sqrt{\beta})^2$.

It is seen that the remainder term decreases essentially exponentially at rate r no matter i, j, and for this reason r^{-1} (or some multiple) is frequently denoted as the *relaxation time* of the system, measuring in some appropriate sense the time needed for the initial condition $X_0 = i$ to become unimportant and the system to relax in the steady state.

Problems

8.1 Show that if $\rho = 1$ in Proposition 8.2, then $\{T_n\}$ is a recurrent Markov chain by (a) explicit calculations of $\sum_0^\infty p_{00}^{(2n)}$, (b) test function techniques and $\mathbb{E}_i|T_1| = |i|$, $i \neq 0$.
8.2 Find asymptotic expressions for the tails $\mathbb{P}(G > t)$ and $\mathbb{P}(G + H > t)$ of the busy period, resp. the busy cycle, in $M/M/1$.
8.3 Evaluate the Laplace transform (or generating function) of the number N of customers served in a busy period. Check the formula by $\mathbb{E}N = \delta\mathbb{E}G$ (and explain why this is true!).
8.4 Consider $M/M/1$ with $\rho \geq 1$ and let $S_t = \sum_1^{N_t} U_n - t$ with $\{N_t\}$ the arrival process and $U_1, U_2, \ldots$ the service times, and define $\tau(u) = \inf\{t > 0 : S_t > u\}$, $B(u) = S_{\tau(u)} - u$. Evaluate the Lévy exponent $\kappa(\alpha) = \log \mathbb{E}\mathrm{e}^{\alpha S_1}$. Explain heuristically that $B(u)$ is independent of $\tau(u)$ with $\mathbb{P}(B(u) > x) = \mathrm{e}^{-\delta x}$ and that $\mathbb{E}\exp\big\{\alpha S_{\tau(u)} - \tau(u)\kappa(\alpha)\big\} = 1$. Evaluate thereby the Laplace transform of $\tau(u)$. How is $\tau(u)$ related to time–dependent properties of the workload?
8.5 Let $\tau = \inf\{t > 0 : |S_t| = 2\}$ with $\{S_t\}$ the doubly infinite queue. Evaluate in the symmetric case $\beta = \delta$ the Laplace transform of τ by a similar method as used for the busy period, and check with Proposition 4.1.

Notes What we have called time–dependent properties are often referred to in the literature as *transient* properties (not a good terminology!). Textbook treatments of the topic for queues in general and for $M/M/1$ in particular can be found in (among many) Takács (1962), Cox and Smith (1961), Prabhu (1965) and Cohen (1982), whereas a detailed study of aspects of $M/M/1$ is in Abate and Whitt (1988). Cohen's book is also a monumental treatise of the general area of transform methods in queueing theory. For Bessel functions, see for example Abramowitz and Stegun (1972).

9 Waiting Times and Queue Disciplines in $M/M/1$

1a. Waiting Times and Workload in the FIFO Case
1b. The LIFO Case
1c. The SIRO Case
1d. The PS Sojourn Time

So far queue length processes have received considerably more attention than the waiting times. The main reason has simply been that as discrete state space processes they are easier to handle by Markovian methods than the continuous waiting times, but we have now actually collected enough results to be able to say something about waiting times. We shall concentrate on $M/M/1$ in the steady state and the effects of changing the queue discipline.

9a Waiting Times and Workload in the FIFO Case

Theorem 9.1 *Consider the $M/M/1$ queue in the steady state. Then the waiting time and the workload have a common distribution that is a mixture with weights $1-\rho$, ρ of an atom at 0 and an exponential distribution with intensity $\gamma = \delta - \beta$,*

$$\mathbb{P}_e(W_n \le y) \;=\; \mathbb{P}_e(V_t \le y) \;=\; 1-\rho+\rho(1-\mathrm{e}^{-\gamma y}) \;=\; 1-\rho\mathrm{e}^{-\gamma y}. \quad (9.1)$$

Before embarking into the proof, we stress that when talking about "the $M/M/1$ queue in the steady state" we must distinguish between time and customer stationarity. More precisely, a time–stationary version $\{V_t^*\}_{t\ge 0}$ of the workload process is not customer–stationary since, for example, the first customer to arrive has a waiting time with distribution different from (9.1). Indeed, his waiting time is $V_{T_1-}^*$ where T_1 is the first arrival time, and obviously $V_{T_1-}^* = (V_0^* - T_1)^+$ is effectively smaller than the representative V_0^* for (9.1). In different terms, the particular customer is not "sampled at random." We return to such phenomena in V.3 and VII.6.

Proof. The workload at time t is 0 if the system is empty which occurs w.p. $1-\rho$ in the steady state. If $X_t = n > 0$ customers are being served, the workload is the residual service time Y_1 of the customer in service plus the service times $Y_2, \ldots, Y_n$ of the ones waiting in line. But by the memoryless property of the exponential distribution, Y_1 is exponential with intensity δ and independent of $Y_2, \ldots, Y_n$. Hence $Y_1, \ldots, Y_n$ are i.i.d. exponentials with intensity δ given $X_t = n$ and thus

$$\mathbb{P}_e(V_t \le y) \;=\; 1-\rho+\sum_{n=1}^{\infty}(1-\rho)\rho^n\mathbb{P}(Y_1+\cdots+Y_n \le y)$$

which reduces to the r.h.s. of (9.1); cf. II.3.2.

We shall give three different proofs that also $\mathbb{P}_e(W_n \le y)$ is as in (9.1) (some routine calculations are omitted in (a), (b)):

(a) Apply the uniqueness of the solution to Lindley's integral equation (6.6) and show directly that if F is doubly exponential as in Example 6.1, then the r.h.s. of (9.1) is a solution (cf. also Problem 6.4).

(b) Letting $M(k) = T_1 + \cdots + T_{k-1}$ be the arrival time of customer k, we have $W_k = V_{M(k)-}$. Hence it suffices to show that the limiting stationary probabilities for the Markov chain $\{X_{M(k)-}\}$ are the same as the ones $\pi_n = (1-\rho)\rho^n$ for $\{X_t\}$ since we may then condition upon $X_{M(k)-}$ and proceed as above. But $\{X_{M(k)-}\}$ is the Markov chain studied in Example 6.2. Inserting $A(\mathrm{d}t) = \beta \mathrm{e}^{-\beta t}\,\mathrm{d}t$ yields $q_k = \rho/(1+\rho)^{k+1}$, and it is then immediately checked that π is stationary for the transition matrix (6.4) (cf. Problem 6.5).

(c) We use the maximum representations of r.v.'s W, V having the steady–state distributions, i.e.

$$W \overset{\mathscr{D}}{=} \max\{0, U_0 - T_0, U_0 + U_1 - T_0 - T_1, \ldots\}, \quad V \overset{\mathscr{D}}{=} \max_{0 \le t < \infty} \{S_t^{\uparrow} - t\}$$

where $S_t^{\uparrow} = \sum_1^{N_t} U_k$, cf. Corollaries 6.5 and 7.2 and Examples 6.1 and 7.5. Then $\{S_t^{\uparrow} - t\}$ increases only at times $M(1), M(2), \ldots$ so that the maximum is attained either at time 0 or at some $M(k)$. Now just note that by sample path insection, $S_{M(k)}^{\uparrow} - M(k) = \sum_0^{k-1}(U_n - T_n)$ so that the two maxima above are equal. □

In connection with (b), one may feel on intuitive grounds that it can be inferred from the Poisson arrivals alone and without calculations that the steady–state distribution of $\{X_{M(k)-}\}$ is the same as that of $\{X_t\}$, the reason being heuristically that the state of $\{X_t\}$ seen by the arriving customers is "chosen at random." Such reasoning is obviously important for intuition but requires some tightening which will be done in VII.6 in the framework of PASTA (Poisson Arrivals See Time Averages).

Note also that the proof of (c) immediately yields:

Corollary 9.2 *The steady–state workload and the steady–state waiting time in the FIFO $M/G/1$ queue have the same distribution.*

9b The LIFO Case

The basic observation is that passing from FIFO to (nonpreemptive) LIFO discipline neither changes the distribution of the queue length at a fixed time t nor prior to an arrival. Hence exactly as above, we may evaluate $\mathbb{P}_e(W_k \le y)$ by conditioning upon the events $\{X_{M(k)-} = n\}$ having probabilities $(1-\rho)\rho^n$. Now clearly $W_k = 0$ if $n = 0$. If $n > 0$, customer k must wait for the server to finish the customer presently in front of him and to clear customers arriving later than customer k. Thus his service can start at time

$$M(k)+W(k) \;=\; \inf\bigl\{t \ge M(k) :\, X_t = n-1\bigr\}.$$

But this shows that independently of $n \ge 1$ W_k is distributed as the time of first passage of the doubly infinite queue from 0 to -1, or equivalently as the busy period; cf. Section 8c. Hence by Corollary 8.7:

Corollary 9.3 *Consider the nonpreemptive LIFO $M/M/1$ queue in the steady state. Then*

$$\mathbb{P}_e(W_n \le y) \;=\; 1-\rho \,+\, \rho^{1/2}\int_0^y \frac{1}{t}\mathrm{e}^{-(\beta+\delta)t} I_1(2\mu t)\,\mathrm{d}t. \tag{9.2}$$

9c The SIRO Case

We say that a customer is of type n if he meets n other customers in the system upon arrival (this occurs in the steady state w.p. $\pi_n = (1-\rho)\rho^n$, cf. proof (b) of Theorem 9.1; obviously the queue length has the same distribution for both the FIFO and SIRO cases).

Considering the steady–state SIRO case, let $H_n(y)$ denote the probability that the waiting time of a customer of type n strictly exceeeds $y \ge 0$. Then

$$\mathbb{P}_e(W_n > y) \;=\; \sum_{n=0}^{\infty} \pi_n H_n(y) \;=\; (1-\rho)\sum_{n=0}^{\infty}\rho^n H_n(y) \tag{9.3}$$

(here obviously $H_0(y)=0$, $y \ge 0$).

Consider a n–type customer ($n \ge 1$) arriving at time 0 and let u be the time of first exit from state n. If u is a departure time, then the customer is selected for service w.p. $1/n$. That is, he continues to wait w.p. $(n-1)/n$ and behaves then as a type $n-1$ customer who has already waited u time units. It follows that

$$\begin{aligned} H_n(y+t) \;&=\; \mathrm{e}^{-(\beta+\delta)t}H_n(y) \\ &\quad + \int_0^t \Bigl\{\beta H_{n+1}(y+t-u) + \delta H_{n-1}(y+t-u)\frac{n-1}{n}\Bigr\}\,\mathrm{e}^{-(\beta+\delta)u}\,\mathrm{d}u. \end{aligned}$$

Hence up to $\mathrm{o}(h)$ terms

$$H_n(y)+H_n'(y)h \;=\; \bigl(1-(\beta+\delta)h\bigr)H_n(y)+\beta H_{n+1}(y)h+\delta H_{n-1}(y)\frac{n-1}{n}h,$$

$$H_n'(y) \;=\; \delta H_{n-1}(y)\frac{n-1}{n} - (\beta+\delta)H_n(y)+\beta H_{n+1}(y) \tag{9.4}$$

It follows by induction from (9.4) that H_n is C^∞ on $[0,\infty)$ with $\bigl|H_n^{(k)}(y)\bigr| \le (\beta+\delta)^k$. This is sufficient to ensure the validity of the series expansions

$$H_n(y) \;=\; \sum_{k=0}^{\infty} h_n^{(k)}\frac{y^k}{k!}, \qquad \mathbb{P}_e(W_n > y) \;=\; \sum_{k=0}^{\infty} h^{(k)}\frac{y^k}{k!}, \tag{9.5}$$

where $h^{(k)} = \sum_0^\infty \pi_n h_n^{(k)}$ and

$$h_n^{(k)} \;=\; H_n^{(k)}(0) \;=\; \delta h_{n-1}^{(k-1)}\frac{n-1}{n} - (\beta+\delta)h_n^{(k-1)} + \beta h_{n+1}^{(k-1)}, \qquad (9.6)$$

$$h_0^{(k)} \;=\; 0, \quad h_n^{(0)} \;=\; 1 \qquad (9.7)$$

(note for (9.7) that $H_0(y) = 0$ and $H_n(0) = 1$ for $n \geq 1$). In summary:

Theorem 9.4 *The steady–state $M/M/1$ SIRO waiting time distribution is given by* (9.5) *with the $h_n^{(k)}$ recursively determined by* (9.6) *and* (9.7).

Notes Flatto (1997) gave the complete form of the SIRO waiting time distribution:

$$\mathbb{P}_e(W_n > y) \;=\; \frac{2(1-\rho)}{\rho}\int_0^\pi \frac{\mathrm{e}^{(2\xi(\theta)-\theta)\cot\theta - \beta\eta(\theta)y}}{(\mathrm{e}^{\pi\cot\theta}+1)\eta(\theta)^2}\sin\theta\,\mathrm{d}\theta \qquad (9.8)$$

where $\xi(\theta) = \arctan\Bigl(\sin\theta/[\cos\theta - 1/\sqrt{\rho}]\Bigr)$, $\eta(\theta) = 1 - 2\cos\theta/\sqrt{\rho} + 1/\rho$. In particular, this implies the intriguing asymptotics

$$\mathbb{P}_e(W_n > y) \;\sim\; c_1(\beta y)^{5/6}\mathrm{e}^{-c_2\beta y - c_3(\beta y)^{1/3}}, \quad y \to \infty, \qquad (9.9)$$

where $c_2 = (1/\sqrt{\rho}-1)^2$, $c_2 = 3(\pi/2)^{2/3}\rho^{-1/6}$,

$$c_1 \;=\; 2^{2/3}3^{-1/2}\pi^{5/6}\rho^{17/12}\frac{1+\sqrt{\rho}}{(1-\sqrt{\rho})^3}\mathrm{e}^{(1+\sqrt{\rho})/(1-\sqrt{\rho})}.$$

9d The PS Sojourn Time

In PS, the waiting time cannot be given a similar sense as for FIFO, LIFO and SIRO (or at least it is then always 0) since a customer k starts service as soon as he arrives (though in general at a reduced rate). Instead we shall be interested in his sojourn time W_k^*. We always have $W_k^* \geq U_k$ where U_k is the service time, and so $W_k = W_k^* - U_k$ may be interpreted as the delay caused by the possible presence of other customers.

A type n customer is defined as for SIRO and we let $K_n(y)$ denote the steady–state probability that his sojourn time strictly exceeds $y \geq 0$. A service event within h time units after arrival of a type n customer will terminate the sojourn of any particular customer w.p. $1/(n+1) + \mathrm{o}(h)$. Letting $K_{-1}(y) = 0$, it follows that up to $\mathrm{o}(h)$ terms

$$K_n(y+h) \;=\; (1-(\beta+\delta)h)K_n(y) + \beta K_{n+1}(y)h + \delta K_{n-1}(y)\frac{n}{n+1}h,$$

$$K_n'(y) \;=\; \delta K_{n-1}(y)\frac{n}{n+1} - (\beta+\delta)K_n(y) + \beta K_{n+1}(y). \qquad (9.10)$$

This means that $\widetilde{H}_0 = K_{-1}$, $\widetilde{H}_1 = K_0$, $\widetilde{H}_2 = K_1$ satisfy exactly the same set of equations (9.4), (9.6), (9.7) as $H_0, H_1, H_2, \ldots$ in the SIRO case. The solution being unique, we get:

Theorem 9.5 *The distribution of the steady–state $M/M/1$ PS sojourn time W_k^* is given by*

$$\begin{aligned}\mathbb{P}_e(W_k^* > y) &= \sum_{n=0}^{\infty} \pi_n K_n(y) \;=\; \sum_{n=0}^{\infty} \pi_n H_{n+1}(y) \\ &= (1-\rho)\sum_{k=0}^{\infty}\frac{y^k}{k!}\sum_{n=0}^{\infty}\rho^n h_{n+1}^{(k)}.\end{aligned}$$

In particular, $\rho\mathbb{P}_e(W_k^* > y) = \mathbb{P}_e(W_k^{\text{SIRO}} > y)$ where W_k^{SIRO} is the SIRO waiting time from Section 9c.

It is also of interest to ask for the conditional distribution of the sojourn time W_k^* given the service time U_k. For example in a time–sharing computer, U_k represents the size of the job (ideal execution time) and W_k^* the actual execution time. At the intutitive level, the following result states that PS (and therefore presumably also RR with a small quantum) is fair in the sense that the average sojourn time is proportionally dependent on U_k:

Theorem 9.6 $\mathbb{E}_e[W_k^* \,|\, U_k] \;=\; U_k/(1-\rho)$.

Proof. We let $m_n(u)$ denote the conditional expectation of the sojourn time of a type n customer with service time u and write $m(u) = \sum_0^\infty \pi_n m_n(u)$ so that we have to show $m(u) = u/(1-\rho)$. Now if no services or arrivals occur within t time units after the arrival of a type n customer with service time u, he will have attained $t/(n+1)$ units of service and hence behave like a type n customer with service time $u - t/(n+1)$. Also the sojourn time has increased by t unless the customer has been served. For small t, this service event occurs with intensity $\delta/(n+1)$ and other service events with intensity $n\delta/(n+1)$. Letting $t = h$, we get up to o(h) terms that

$$\begin{aligned}m_n(u) &= h\Bigl(1-\delta\frac{1}{n+1}h\Bigr) + m_n\Bigl(u-\frac{h}{n+1}\Bigr)\Bigl\{1-\beta h-\delta\frac{n}{n+1}h\Bigr\} \\ &\quad + \beta m_{n+1}(u)h + \delta m_{n-1}(u)\frac{n}{n+1}h \\ &= h + m_n(u) - m_n'(u)\frac{h}{n+1} - \beta m_n(u)h - \delta m_n(u)\frac{n}{n+1}h \\ &\quad + \beta m_{n+1}(u)h + \delta m_{n-1}(u)\frac{n}{n+1}h, \\ m_n' &= n+1-\beta(n+1)m_n - n\delta m_n + \beta(n+1)m_{n+1} + \delta n m_{n-1}, \qquad (9.11) \\ m' &= \sum_{n=0}^{\infty}\pi_n m_n' \\ &= \sum_{n=0}^{\infty}(n+1)\pi_n - \beta\sum_{n=0}^{\infty}(n+1)\pi_n m_n - \delta\sum_{n=0}^{\infty}n\pi_n m_n\end{aligned}$$

$$+\beta\sum_{n=1}^{\infty} n\pi_{n-1}m_n \;+\; \delta\sum_{n=0}^{\infty}(n+1)\pi_{n+1}m_n. \qquad (9.12)$$

Using $\pi_{n+1} = \rho\pi_n$, $\sum n\pi_n = \rho/(1-\rho)$, this reduces to

$$\begin{aligned} & \frac{\rho}{1-\rho} \;+\; 1 \;-\; \beta\sum_{n=0}^{\infty} n\pi_n m_n \;-\; \beta m \;-\; \delta\sum_{n=0}^{\infty} n\pi_n m_n \\ & +\frac{\beta}{\rho}\sum_{n=0}^{\infty} n\pi_n m_n \;+\; \rho\delta\sum_{n=0}^{\infty} n\pi_n m_n \;+\; \rho\delta m \\ = \;& \frac{1}{1-\rho} \;+\; \sum_{n=0}^{\infty} n\pi_n m_n\Big\{-\beta-\delta+\frac{\beta}{\rho}+\rho\delta\Big\} \;+\; m\left\{-\beta+\rho\delta\right\}\,. \end{aligned}$$

But since $\rho = \beta/\delta$, the two $\{\cdot\}$ are 0. Thus $m'(u) = 1/(1-\rho)$ and since clearly $m(0) = 0$, we have $m(u) = u/(1-\rho)$ as desired. □

It should be noted that some technical details (not coming up in the proof of Theorem 9.5) have been omitted. For example, to differentiate under the sum sign in (9.12) we need some bound on the m_n' (or equivalently the m_n; cf. (9.11)).

Notes The proof of Theorem 9.6 carries over also to find the Laplace transform of W_n^* given $U_n = u$. For references to this and other early work on PS, see the surveys by Yashkov (1987, 1992) (where also a relation to branching processes is discussed). For more on the connection to SIRO, see Borst *et al.* (2003). Other recent contributions include Zwart and Boxma (2000) and Jelenkovic and Momcilovic (2003) who derive heavy–tailed asymptotics.

Generalized processor sharing is a queueing discipline for $r > 1$ customer classes where class i receives service at rate ϕ_i ($\phi_1+\cdots+\phi_r = 1$) and the service discipline within each class is FIFO. See, for example, Dupuis and Ramanan (1998) and Jelenkovic and Momcilovic (2002).

IV
Queueing Networks and Insensitivity

1 Poisson Departure Processes and Series of Queues

We start by noting:

Proposition 1.1 *Any ergodic birth–death process is time reversible.*

Proof. We must check the conditions $\pi_i \lambda(i,j) = \pi_j \lambda(j,i)$ of detailed balance in II.5.3. If $|i-j| > 1$, then by the skip–free property both sides are zero so we can suppose $|i-j| = 1$, say $i = n$, $j = n+1$, where the condition reduces to $\pi_n \beta_n = \pi_{n+1}\delta_{n+1}$ which is clear from III.(2.4). It is instructive to indicate how alternatively the proof can be carried out without first computing $\boldsymbol{\pi}$: in equilibrium, the flow from $\{0,\ldots,n\}$ to $\{n, n+1, \ldots\}$ must balance the flow the other way. But the only possible transition the first way is from n to $n+1$ so the flow is $\pi_n \beta_n$. Similarly, the flow the other way is $\pi_{n+1}\delta_{n+1}$. □

Consider now a doubly infinite stationary version $\{X_t\}_{-\infty<t<\infty}$ and define $\widetilde{X}_t = X_{-t-} = \lim_{s\uparrow -t} X_s$. Then a departure for $\{X_t\}$ at time s corresponds to an arrival for $\left\{\widetilde{X}_t\right\}$ at time $-s$. Considering the case of Poisson arrivals, $\beta_n = \beta$, the collection of such instants $-s$ form a Poisson process with intensity β by reversibility. Hence the instants s do so too, and we have proved:

Corollary 1.2 *The departure process of an ergodic birth–death queue with Poisson arrivals with intensity β is itself a Poisson process with intensity β.*

This is the first in a series of results in this and the next section that may be argued to be contrary to intuition or at least surprising. A discussion of this is deferred to Section 2d.

A main application of the Poisson departure property is to series of K queues, where customers enter queue 1 according to a Poisson process with intensity β, after being served proceed to queue 2, from there to queue 3 and so on (the system with $K = 2$ is called a *tandem queue*). Suppose that queue k has exponential services with intensity $\delta_n^{(k)}$ at queue length n. Then queue 1 is a birth–death queue with Poisson arrivals, hence has a Poisson departure process in equilibrium. But this process is just the arrival process to queue 2 so that this is a birth–death queue with Poisson arrivals which hence delivers Poisson input to queue 3 and so on. It follows that in equilibrium the number of customers at queue k and their waiting times have the same characteristics as if the queue was considered in isolation and subject to Poisson arrivals. We generalize below this reasoning to the simultaneous behaviour of the queues, but first we shall give one more of the classical examples of a Poisson departure process.

Theorem 1.3 *The stationary $M/G/\infty$ queue $\{X_t\}_{-\infty<t<\infty}$ with doubly infinite time has a Poisson departure process.*

Proof. The departure process M has epochs $\{\sigma(n)+U_n\}_{n\in\mathbb{Z}}$ where $\{\sigma(n)\}_{n\in\mathbb{Z}}$ are the epochs of of the arrival process N and $\{U_0, U_{\pm 1},, U_{\pm 2}, \ldots\}$ the sequence of service times. It is a general fact about the Poisson process that such an i.i.d. translation is Poisson with the same intensity (say β) as N but here is a self–contained proof. The idea is to observe that this is trivial if the U_n are discrete, and next to apply a discrete approximation: Suppose first that the U_n can assume only the values $0, \pm\delta, \pm 2\delta, \ldots$ and let $p_k = \mathbb{P}(U_n = k\delta)$, $k \in \mathbb{Z}$. Then, by standard properties of Poisson thinning, the $\sigma(n)$ with $U_n = k\delta$ form a Poisson process $N^{(k)}$ with intensity βp_k, and the $N^{(k)}$ are independent. Letting $M^{(k)}$ denote $N^{(k)}$ translated by $k\delta$, $M^{(k)}$ is Poisson with intensity βp_k and the $M^{(k)}$ are independent. Hence $M = \sum_{k\in\mathbb{Z}} M^{(k)}$ is Poisson with intensity β as asserted.

To deal with the general case, let $U_n^{(\delta)} = k\delta$, $k\delta \le U_n < (k+1)\delta$, and let M_δ have epochs $\bigl\{\sigma(n) + U_n^{(\delta)}\bigr\}_{n\in\mathbb{Z}}$. If $I_1, \ldots, I_R$ are disjoint intervals, then $M_\delta(I_r) \overset{\text{a.s.}}{\to} M(I_r)$ for all r as $\delta \downarrow 0$ since $U_n^{(\delta)} \overset{\text{a.s.}}{\to} U_n$ and the probability of an epoch of M at a boundary point of I_r is zero. Hence $\{M_\delta(I_r)\}_1^R \overset{\mathscr{D}}{\to} \{M(I_r)\}_1^R$. But by what has just been proved, M_δ is Poisson with intensity β. Hence the joint distribution of the $M(I_r)$ is the common joint distribution of the $M_\delta(I_r)$ so that $M \overset{\mathscr{D}}{=} M_\delta$. □

Preparing for a more thorough study of series of queues, we start by noting:

Proposition 1.4 *Consider the equilibrium version $\{X_t\}_{-\infty<t<\infty}$ of an ergodic birth–death process with Poisson arrivals at rate β. Then the departure process prior to t is Poisson at rate β and independent of X_t.*

Proof. It only remains to check the independence. The argument is a slight variant of the proof of Corollary 1.2: the departure process prior to t is the arrival process after $-t$ of the time–reversed process and hence independent of its queue length $\widetilde{X}_{-t}$ which a.s. coincides with X_t. □

Corollary 1.5 *Consider a series of K queues in series where the arrivals to the first are Poisson at rate β and the K servers work independently, with rate $\delta_n^{(k)}$ for server k at queue length n. Suppose that for each k the birth–death queue with $\beta_n = \beta$, $\delta_n = \delta_n^{(k)}$ is ergodic, and let $\boldsymbol{\pi}^{(k)}$ denote its stationary distribution, $\pi_n^{(k)} = S_k^{-1}\beta^n/\bigl(\delta_1^{(k)}\cdots\delta_n^{(k)}\bigr)$. Then the series system is also ergodic and the steady state is described by the queue lengths $X_t^{(1)},\ldots,X_t^{(K)}$ being independent with $X_t^{(k)}$ governed by $\boldsymbol{\pi}^{(k)}$,*

$$\mathbb{P}_e\bigl(X_t^{(1)} = n(1),\ldots,X_t^{(K)} = n(K)\bigr) \;=\; \pi^{(1)}_{n(1)}\cdots\pi^{(K)}_{n(K)}. \tag{1.1}$$

Proof. Letting $\boldsymbol{\pi} = \boldsymbol{\pi}^{(1)}\otimes\cdots\otimes\boldsymbol{\pi}^{(K)}$ be the distribution (1.1), we proceed by showing $\boldsymbol{\pi}\boldsymbol{P}^t = \boldsymbol{\pi}$ for any fixed t (an alternative proof involving $\boldsymbol{\pi}\boldsymbol{\Lambda} = \boldsymbol{0}$ is in Problem 1.2). Suppose thus that the $X_0^{(k)}$ has been assigned initial joint distribution $\boldsymbol{\pi}$ and let $N^{(k)}$ be the departure process from queue k in $[0,t]$. The conclusion will follow if we can show that for each k, $X_t^{(1)},\ldots,X_t^{(k)},N^{(k)}$ are independent, governed by $\boldsymbol{\pi}^{(1)},\ldots,\boldsymbol{\pi}^{(K)}$ and the distribution of the Poisson process respectively. The case $k=1$ is just Proposition 1.4. Suppose the assertion holds for k. Both $X_t^{(k+1)}$ and $N^{(k+1)}$ depend on $N^{(k)}, X_0^{(k+1)}$ and the action of server $k+1$ only, hence the set $\bigl(X_t^{(k+1)},N^{(k+1)}\bigr)$ is independent of $X_t^{(1)},\ldots,X_t^{(k)},N^{(k)}$. But since $N^{(k)}$ is Poisson and $X_0^{(k)}$ governed by $\boldsymbol{\pi}^{(k)}$, it follows by applying Proposition 1.4 once more that the joint distribution of $X_t^{(k+1)},N^{(k+1)}$ is as asserted. □

The simplest case is of course that of $M/M/1$ queues in series, $\delta_n^{(k)} = \delta^{(k)}$. Then $\rho_k = \beta/\delta^{(k)}$ is the traffic intensity at queue k and

$$\pi_{n(1)\ldots n(K)} \;=\; \prod_{k=1}^{K}\pi^{(k)}_{n(k)} \;=\; \prod_{k=1}^{K}(1-\rho_k)\rho_k^{n(k)}.$$

Problems

1.1 Consider the case $K=3$, but assume that customers leaving queue 1 do not necessarily go to 2 but choose between 2 and 3 with probabilities p, resp. $q = 1-p$. Show that if queue 1 has Poisson input and is ergodic, then in equilibrium the input to 2, 3 are independent Poisson processes. Formulate the criterion for ergodicity of the whole system and show that the stationary distribution is of product form as in (1.1).

1.2 Check that (1.1) satisfies $\boldsymbol{\pi\Lambda} = \mathbf{0}$ and that $\{(X_t^{(1)}, \dots, X_t^{(k)})\}$ is nonexplosive.

Notes References and discussion covering the whole of this chapter IV (except Section 3) are given at the end of the chapter.

2 Jackson Networks

2a. Model and Examples
2b. Ergodic Theory for Open Networks
2c. Ergodic Theory for Closed Networks
2d. Pitfalls for Intuition

2a Model and Examples

One of the simplest examples of a queueing network, queues in series, has already been encountered in the preceding section. It imposes, however, together with simple generalizations as in Problem 1.1, the restriction that customers can only move along a feed–forward path.

We consider now more generally a queueing network where there is a finite number K of individual queues, the *nodes* of the network, at which customers arrive from external sources according to independent Poisson processes with intensities $\alpha_1, \dots, \alpha_K$. A customer having completed service at node k goes to node ℓ w.p. $\gamma_{k\ell}$ and leaves the system with w.p. $\gamma_{k0} = 1 - \sum_1^K p_{k\ell}$ (external drain). A graphical illustration is given in Fig. 2.1(a), an arrow from k to ℓ denoting $\gamma_{k\ell} > 0$ and arrows to and from the external world (node 0) denoting $\gamma_{k0} > 0$ and $\alpha_k > 0$, respectively.

In this section, we consider the case of a single exponential server at each node (one then talks frequently about *Jackson networks*), and we denote the corresponding service rates by $\delta_1, \dots, \delta_k$. Two main types are considered, *open* networks where external input is received and external output is delivered by customers entering and leaving the system, and *closed* networks where the customers can only move internally in the system. Thus a closed network has all $\alpha_k = 0$ and all $\gamma_{k0} = 0$, and the total number of customers in the system does not vary with time. In an open network some $\alpha_k > 0$ and some $\gamma_{k0} > 0$, and the number in system is a nondegenerate stochastic process.

Networks of this type come up in a great variety of problems, the most important of which are of rather recent date and associated with data communication systems and the internal organization of time–sharing computers. Also, colonies of biological individuals with migration between colonies have been modelled in this way. We shall give two examples, one of an open and one of a closed network, with the reservation that it should be stressed that the above model does not pretend to be anything but a crude first approximation. For example, there would frequently be a strong positive correlation between the service times of a customer at the different nodes. In other cases the choice of the customer of the path along the nodes could depend on the length of the waiting lines.

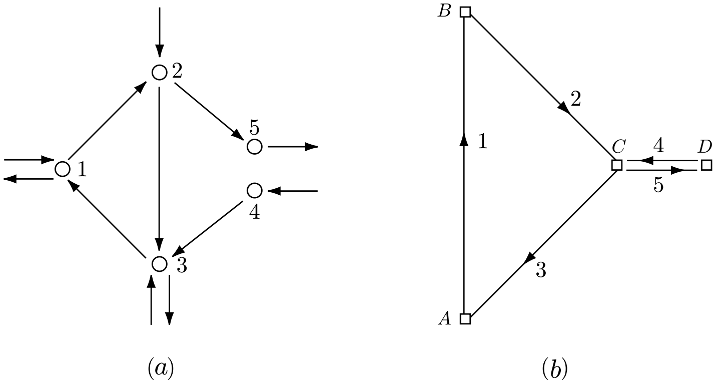

Figure 2.1

Fig. 2.1(b) depicts a communication network, say connecting three branches A, B, D, of a bank, or three computers via a transmission station C. Messages are sent by the directed channels 1, 2, 3, 4, 5 and if a channel is busy, a queue may be formed. Thus if we reinterpret channels as nodes and messages as customers, we arrive at the network in Fig. 2.1(a). This is open since new messages are created currently and a message leaves the system after having reached its destination.

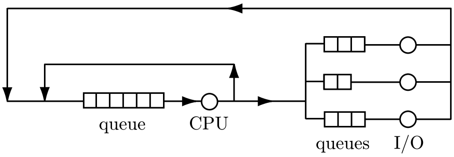

Figure 2.2

Consider next jobs (customers) circulating in a time–sharing computer as in Fig. 2.2. The nodes are the CPU and input/output facilities, the allowance of feedback at the CPU corresponding to some sort of PS or RR. At first glance this looks like an open network. However, the number of steps taken by each job is typically very large and thus within time intervals of moderate length, the number of jobs is fixed and a description by a closed network may be more appropriate.

Now let $X_t^{(k)}$ denote the number of customers at node k at time t and let $\boldsymbol{X}_t = \big(X_t^{(1)}, \ldots, X_t^{(K)}\big)$. The state space is $\mathbb{N}^K$ and the states are denoted $\boldsymbol{n} = n_1 \ldots n_K$. We write $\boldsymbol{n}_k^{(+)}$ for the state obtained by increasing n_k by 1,

$\boldsymbol{n}_k^{(-)}$ for the state (defined only when $n_k > 0$) obtained by decreasing n_k by 1, and $\boldsymbol{n}_{k,\ell}$ for the state (defined only when $n_k > 0$) obtained by decreasing n_k by 1 and increasing n_ℓ by 1. It follows that the possible transitions are

$$\begin{array}{cccc} & \nearrow & \boldsymbol{n}_{k,\ell} & \delta_k\gamma_{k\ell} \\ \boldsymbol{n} & \rightarrow & \boldsymbol{n}_k^{(-)} & \delta_k\gamma_{k0} \\ & \searrow & \boldsymbol{n}_k^{(+)} & \alpha_k \end{array} \tag{2.1}$$

with the right column giving the intensities; the first row corresponds to a customer going to node ℓ after being served at node k, the second to a customer leaving the system after being served at node k and the last to an arrival at node k. Thus only the first type of transition occurs in a closed network.

We first need to make an appropriate definition of the throughput rate β_k at node k, that is, the common rate of the input and output processes (these need not be Poisson but the rate should exist in terms of long–term averages). The input rate is the sum of the rate α_k of external arrivals and the rate of internal arrivals from nodes $\ell \neq k$. But customers leave node ℓ at rate β_ℓ and go then to k with probability $\gamma_{\ell k}$, so that we should have the *traffic equations*

$$\beta_k \;=\; \alpha_k + \sum_{\ell=1}^{K} \beta_\ell \gamma_{\ell k}, \quad k = 1, \ldots, K. \tag{2.2}$$

2b Ergodic Theory for Open Networks

We shall assume that each node k may both receive external input and deliver external output (possibly via other nodes), i.e. for each k (i) either $\alpha_k > 0$ or some $\alpha_{\ell_1}\gamma_{\ell_1\ell_2}\cdots\gamma_{\ell_r k} > 0$, and (ii) either γ_{k0} or some $\gamma_{kk_1}\gamma_{k_1k_2}\cdots\gamma_{k_n0} > 0$. This is easily seen to imply irreducibility of $\{\boldsymbol{X}_t\}_{t\geq 0}$.

Proposition 2.1 *The traffic equations* (2.2) *have a unique nonnegative solution* $(\beta_1, \ldots, \beta_K)$*. It satisfies* $0 < \beta_k < \infty$*.*

Proof. Consider a Markov jump process on $\{0, \ldots, K\}$ with off–diagonal intensities $\lambda(k, \ell) = \gamma_{k\ell}$, $k \neq 0$, $\lambda(0, k) = \alpha_k$. Our assumptions apply irreducibility and hence the existence of a stationary distribution $\boldsymbol{\nu}$, uniquely given by $\boldsymbol{\nu\Lambda} = \boldsymbol{0}$, which amounts to the $K + 1$ linear equations

$$\nu_k \;=\; \nu_0\alpha_k + \sum_{\ell=1}^{K} \nu_\ell\gamma_{\ell k}, \quad k = 1, \ldots, K, \tag{2.3}$$

$$\nu_0 \sum_{k=1}^{K} \alpha_k \;=\; \sum_{\ell=1}^{K} \nu_\ell \gamma_{\ell 0}\,. \tag{2.4}$$

It is therefore clear from (2.3) that $\beta_k = \nu_k/\nu_0$ solves (2.2). Suppose conversely $(\beta_1, \ldots, \beta_K)$ is a solution and define $\beta_0 = 1$, $\nu_k^* = \beta_k / \sum_0^K \beta_\ell$. Then

(2.3) holds for $\boldsymbol{\nu}^*$ and (2.4) is a consequence of (2.3), as is seen by summing over $k = 1, \ldots, K$ and performing some algebra. Thus $\nu_k^* = \nu_k$, i.e. $\sum_0^K \beta_\ell = 1/\nu_0$, $\beta_k = \nu_k/\nu_0$, proving uniqueness. □

Theorem 2.2 *Assume that $\rho_k = \beta_k/\delta_k < 1$ for all $k = 1, \ldots, K$. Then $\{\boldsymbol{X}_t\}$ is ergodic with stationary distribution $\boldsymbol{\pi}$ given by*

$$\pi_{\boldsymbol{n}} \;=\; \pi_{n_1 \ldots n_K} \;=\; \prod_{k=1}^K (1-\rho_k)\rho_k^{n_k}.$$

Proof. The intensities are bounded, cf. (2.1), and hence it suffices to show that $\boldsymbol{\pi}\boldsymbol{\Lambda} = \boldsymbol{0}$, cf. II.4.4, which by (2.1) amounts to

$$\begin{aligned} \pi_{\boldsymbol{n}} \sum_{k=1}^K \bigl\{\alpha_k + \delta_k I(n_k > 0)\bigr\} \;&=\; \sum_{k=1}^K \bigl\{\pi_{\boldsymbol{n}_k^{(-)}} \alpha_k I(n_k > 0) + \pi_{\boldsymbol{n}_k^{(+)}} \delta_k \gamma_{k0}\bigr\} \\ &\quad + \sum_{k,\ell=1}^K \pi_{\boldsymbol{n}_{k,\ell}} \delta_\ell \gamma_{\ell k} I(n_k > 0). \qquad (2.5) \end{aligned}$$

Now $\pi_{\boldsymbol{n}_k^{(+)}} = \rho_k \pi_{\boldsymbol{n}}$, $\pi_{\boldsymbol{n}_{k,\ell}} = \rho_k^{-1}\rho_\ell \pi_{\boldsymbol{n}}$. Hence using (2.2) we get

$$\begin{aligned} \sum_{k=1}^K \pi_{\boldsymbol{n}_k^{(+)}} \delta_k \gamma_{k0} \;&=\; \pi_{\boldsymbol{n}} \sum_{k=1}^K \beta_k \gamma_{k0} \;=\; \pi_{\boldsymbol{n}} \sum_{k=1}^K \beta_k \Bigl(1 - \sum_{\ell=1}^K \gamma_{k\ell}\Bigr) \\ &=\; \pi_{\boldsymbol{n}} \sum_{\ell=1}^K \Bigl(\beta_\ell - \sum_{k=1}^K \beta_k \gamma_{k\ell}\Bigr) \;=\; \pi_{\boldsymbol{n}} \sum_{\ell=1}^K \alpha_\ell, \end{aligned}$$

$$\begin{aligned} \pi_{\boldsymbol{n}_k^{(-)}} \alpha_k + \sum_{\ell=1}^K \pi_{\boldsymbol{n}_{k,\ell}} \delta_\ell \gamma_{\ell k} \;&=\; \pi_{\boldsymbol{n}} \rho_k^{-1} \Bigl\{\alpha_k + \sum_{\ell=1}^K \beta_\ell \gamma_{\ell k}\Bigr\} \\ &=\; \pi_{\boldsymbol{n}} \rho_k^{-1} \beta_k \;=\; \pi_{\boldsymbol{n}} \delta_k, \quad n_k > 0, \end{aligned}$$

and (2.5) follows. □

2c Ergodic Theory for Closed Networks

For a closed network, clearly $E_N = \bigl\{\boldsymbol{n} : \sum_1^K n_k = N\bigr\}$ is a closed set, and to discuss irreducibility and ergodicity, we therefore have to restrict the state space of $\{\boldsymbol{X}_t\}$ to E_N. The traffic equations (2.2) reduce in matrix notation to $\boldsymbol{\beta} = \boldsymbol{\beta}\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} = (\gamma_{k\ell})_{1 \le k,\ell \le K}$, and since $\gamma_{k0} = 0$, $\boldsymbol{\Gamma}$ is a transition matrix. We shall assume that $\boldsymbol{\Gamma}$ is irreducible on $\{1, \ldots, K\}$. This implies the existence of a $\boldsymbol{\beta}$ (unique up to a constant) which satisfies $\boldsymbol{\beta} = \boldsymbol{\beta}\boldsymbol{\Gamma}$, and also, as is readily seen, that $\{\boldsymbol{X}_t\}$ is irreducible on E_N and hence ergodic since E_N is finite.

Theorem 2.3 *Under the above assumptions, $\{\boldsymbol{X}_t\}$ is ergodic on E_N with stationary distribution $\boldsymbol{\pi} = \boldsymbol{\pi}^{(N)}$ given by*

$$\pi_{\boldsymbol{n}} \;=\; \pi_{n_1\ldots n_k} \;=\; C_N \prod_{k=1}^{K} \rho_k^{n_k}.$$

where $\rho_k = \beta_k/\delta_k$ and C_N is a normalization constant ensuring $|\boldsymbol{\pi}| = 1$.

Proof. In the same way as for (2.5), we have to check that

$$\pi_{\boldsymbol{n}} \sum_{k=1}^{K} \delta_k I(n_k > 0) \;=\; \sum_{k,\ell=1}^{K} \pi_{\boldsymbol{n}_{k,\ell}} \delta_\ell \gamma_{\ell k} I(n_k > 0) \tag{2.6}$$

But since $\pi_{\boldsymbol{n}_{k,\ell}} = \rho_k^{-1}\rho_\ell \pi_{\boldsymbol{n}}$, we get for $n_k > 0$ that

$$\sum_{\ell=1}^{K} \pi_{\boldsymbol{n}_{k,\ell}} \delta_\ell \gamma_{\ell k} \;=\; \pi_{\boldsymbol{n}} \rho_k^{-1} \sum_{\ell=1}^{K} \beta_\ell \gamma_{\ell k} \;=\; \pi_{\boldsymbol{n}} \rho_k^{-1} \beta_k \;=\; \pi_{\boldsymbol{n}} \delta_k,$$

which implies the truth of (2.6). □

Theorem 2.3 will now be shown to have as a consequence a *bottleneck* type of system behaviour: if N is large, then with high probability most of the N customers will be in the waiting line at the node with the highest ρ_k. Such a knowledge could be useful say for design purposes, since it would in some situations suggest an allocation of the total service capacity $\delta_1 + \cdots + \delta_K$ such that $\max \rho_k$ is minimized.

To illustrate this effect, we shall assume that one ρ_k, say ρ_1, is effectively largest and we may then choose the scale of $\boldsymbol{\beta}$ such that

$$\rho_1 = 1, \quad \rho_2 < 1, \; \ldots, \; \rho_K < 1. \tag{2.7}$$

Consider the marginal steady–state distributions $\boldsymbol{\eta}^{(N)}$, $\boldsymbol{\theta}^{(N)}$ of $\bigl\{X_t^{(1)}\bigr\}$, resp. $\bigl\{(X_t^{(2)},\ldots,X_t^{(K)})\bigr\}$,

$$\begin{aligned}
\eta_n^{(N)} \;&=\; \mathbb{P}_e\bigl(X_t^{(1)} = n\bigr) \;=\; \sum_{n_2+\cdots+n_K=N-n} \pi_{nn_2\ldots n_K}, \\
\theta_{n_2\ldots n_K}^{(N)} \;&=\; \mathbb{P}_e\bigl(X_t^{(2)} = n_2,\ldots,X_t^{(K)} = n_K\bigr) \;=\; \pi_{n_1n_2\ldots n_K},
\end{aligned}$$

where $n_1 = N - n_2 - \cdots - n_K$. Then in the limit $N \to \infty$, $\boldsymbol{\eta}^{(N)}$ becomes degenerate at ∞, whereas $\boldsymbol{\theta}^{(N)}$ has a proper limit of the same form as the steady–state solution of an open network:

Corollary 2.4 *As $N \to \infty$, $\eta_n^{(N)} \to 0$ for all n and*

$$\theta_{n_2\ldots n_K}^{(N)} \;\to\; \prod_{k=2}^{K} (1-\rho_k)\rho_k^{n_k}.$$

Proof. Since $\rho_1 = 1$, we have

$$\begin{aligned}
C_N^{-1} &= \sum_{n_1+\cdots+n_K=N} \prod_{k=1}^{K} \rho_k^{n_k} = \sum_{n_2+\cdots+n_K\le N} \prod_{k=2}^{K} \rho_k^{n_k} \\
&\to \sum_{n(2)=0}^{\infty} \cdots \sum_{n_k=0}^{\infty} \prod_{k=2}^{K} \rho_k^{n_k} = \prod_{k=2} (1-\rho_k)^{-1} = D \text{ (say)}.
\end{aligned}$$

Here $0 < D < \infty$ and hence if $\rho_k < \delta < 1$, $k = 2, \ldots, K$, we get

$$\begin{aligned}
\eta_n^{(N)} &= C_N \sum_{n_2+\cdots+n_K=N-n} \prod_{k=2}^{K} \rho_k^{n_k} \\
&< C_N \binom{K+N-n-2}{N-n} \delta^{N-n} \to 0, \\
\theta_{n_2\ldots n_K}^{(N)} &= C_N \prod_{k=2}^{K} \rho_k^{n_k} \to \prod_{k=2}^{K} (1-\rho_k)\rho_k^{n_k}.
\end{aligned}$$

□

2d Pitfalls for Intuition

We have now given rigorous mathematical proofs of a number of results on queues delivering Poisson output, and queues in series or networks that behave in a certain sense as if they were totally independent and each subject to Poisson arrivals. Many of these results may be difficult to understand on more intuitive grounds. For example, one may ask:

(i) How can even such a simple queue as $M/M/1$ deliver Poisson output at rate β? The server has idle periods with no output and busy periods where departures are Poisson at rate δ. Also, observing the output alone, we can tell the value of β but not δ.

(ii) How can the departure process $M^{\to t}$ prior to t be independent of the departures $M^{t\to}$ after t? Observation of $M^{\to t}$ should tell us something on $\{X_s\}_{s\le t}$ (e.g. if there are few departures just before t, we expect $X_t = 0$ with greater probability than the average $1-\rho$), and conditionally upon X_t, $M^{\to t}$ is certainly not Poisson.

(iii) In a network rather than a series, the arrivals to node k are not in general Poisson. Why then is this not reflected in the behaviour of $X_t^{(k)}$?

(iv) How can $X_t^{(k)}$ and $X_t^{(\ell)}$ in a network be independent? If $X_t^{(k)}$ is large and $\gamma_{k\ell} > 0$, then node ℓ should have received more input prior to t than on the average, hence also $X_t^{(\ell)}$ should be large.

Of course, the mathematical proofs tell that such reasoning has to have gaps or errors, and it is not difficult either to elaborate further on the intuitive reasoning to deduce that important aspects have been ignored (e.g. the independence in a network refers only to a fixed instant of time and not the time evolution). Our point here is merely to stress that intuitive reasoning, though an indispensable part of applied probability, has its pitfalls and that care has to be taken that it can be followed up with a more rigorous proof.

Problems

2.1 Show that the timereverse of a (open or closed) Jackson network is again a Jackson network, with routing probabilities given by $\widetilde{\gamma}_{\ell k} = \beta_k \gamma_{kl}/\beta_\ell$.

3 Insensitivity in Erlang's Loss System

We consider Erlang's loss system as in III.3e with K lines and intensity β for arrivals of calls, but assume now that the duration of a call follows a phase–type distribution B, say with p phases, initial vector $\boldsymbol{\alpha}$ and phase generator $\boldsymbol{T}$ (the exit rate vector is $\boldsymbol{t} = -\boldsymbol{T}\mathbf{1}$). The system can then be modelled as Markov process $\{X_t\}$ whose state space E consists of all $i = n_1 \ldots n_p$ with $n_k = 0, \ldots, K$, $|i| = n_1 + \cdots + n_p \le K$, such that n_k gives the number of lines where the call is currently being handled in phase k. We write $i_r^{(+)}$ for the state (defined only when $|i| < K$) obtained by increasing n_r by 1, $i_r^{(-)}$ for the state (defined only when $n_r > 0$) obtained by decreasing n_r by 1, and $i_{r,s}$ for the state (defined only when $n_r > 0$) obtained by decreasing n_r by 1 and increasing n_s by 1. It follows that the possible transitions are

$$\begin{array}{cccl} & \nearrow & i_{r,s} & n_r t_{rs} \\ i & \rightarrow & i_r^{(-)} & n_r t_r \\ & \searrow & i_r^{(+)} & \beta \alpha_r \end{array} \tag{3.1}$$

with the right column giving the intensities; the first row correspond to a phase change of some call, the second to a call being completed and the last to an arrival of customer.

The loss probability is then

$$E_K = \sum_{n_1 + \cdots + n_p = K} \pi_{n_1 \ldots n_p},$$

where $\boldsymbol{\pi}$ is the stationary distribution, and we will show:

Theorem 3.1 *Let $\mu_B = -\boldsymbol{\alpha}\boldsymbol{T}^{-1}\mathbf{1}$ denote the mean of B and let $\eta = \beta\mu_B$. Then*

$$E_K = \frac{\eta^K/K!}{1 + \eta + \cdots + \eta^K/K!}\,. \tag{3.2}$$

Proof. We will undertake the program of Kelly's lemma II.5.4, i.e. come up with trials π_i, $\widetilde{\lambda}(i,j)$ for the stationary probabilities and the intensities of the reversed process, and verify $\pi_i \lambda(i,j) = \pi_j \widetilde{\lambda}(j,i)$.

Let $\boldsymbol{\nu} = -\boldsymbol{\alpha} \boldsymbol{T}^{-1}/\mu_B$. Since (cf. III.5) $\boldsymbol{\nu}$ is the stationary distribution of a renewal process with interarrival distribution B, we can interpret ν_r as the probability that the current phase of a randomly selected call is r. Therefore a plausible candidate for the stationary distribution is

$$\pi_i \;=\; E_{|i|} \binom{|i|}{n_1 \, \ldots \, n_p} \nu_1^{n_1} \cdots \nu_p^{n_p}, \tag{3.3}$$

where $E_k = (\eta^k/k!)/(1+\eta+\cdots+\eta^K/K!)$ (note that everything following $E_{|i|}$ is just the probability of getting $n_1 \ldots n_p$ by multinomial sampling of $|i|$ objects with probabilities $\nu_1, \ldots, \nu_p$). Once (3.3) is verified to be the correct π_i, we get $\sum_{|i|=k} \pi_i = E_k$, and taking $k=K$, the proof is complete.

A plausible guess for the time–reversed system is another loss system with the same K and β, but with the given phase–type representation replaced by the time–reversed one, which according to III.5.7 is given by

$$\widetilde{\alpha}_r = \mu_B t_r \nu_r, \;\; \widetilde{t}_{rs} = \frac{\nu_s t_{sr}}{\nu_r}, \;\; \widetilde{t}_r = \frac{\alpha_r}{\mu_B \nu_r}.$$

That then indeed $\pi_i \lambda(i,j) = \pi_j \widetilde{\lambda}(j,i)$ is now obtained as follows. First

$$\begin{aligned} \frac{\pi_i \lambda\big(i, i_r^{(+)}\big)}{\pi_{i_r^{(+)}} \widetilde{\lambda}\big(i_r^{(+)}, i\big)} \;&=\; \frac{E_{|i|}(n_r+1)\lambda\big(i, i_r^{(+)}\big)}{E_{|i|+1}(|i|+1)\nu_r \widetilde{\lambda}\big(i_r^{(+)}, i\big)} \;=\; \frac{(n_r+1)\beta\alpha_r}{\eta\nu_r(n_r+1)\widetilde{t}_r} \\ &=\; \frac{\beta\alpha_r}{\beta\mu_B\nu_r(\alpha_r/\mu_B\nu_r)} \;=\; 1. \end{aligned}$$

That $\pi_i \lambda(i, i_r^{(-)}) \;=\; \pi_{i_r^{(-)}} \widetilde{\lambda}(i_r^{(-)}, i)$ then follows by a symmetry argument. Finally,

$$\frac{\pi_i \lambda(i, i_{rs})}{\pi_{i_{rs}} \widetilde{\lambda}(i_{rs}, i)} \;=\; \frac{\nu_r n_s \lambda(i, i_{rs})}{\nu_s n_r \widetilde{\lambda}(i_{rs}, i)} \;=\; \frac{\nu_r n_s n_r t_{rs}}{\nu_s n_r n_s \widetilde{t}_{sr}} \;=\; \frac{\nu_r t_{rs}}{\nu_s(\nu_r t_{rs}/\nu_s)} \;=\; 1. \qquad \square$$

Notes The fact that the loss probability depends on the service time distribution only through its mean μ_B is referred to as *insensitivity.* Some further examples will be given in the next section where it is shown that the stationary queue length distributions in the preemptive LCFS $M/G/1$ queue and the PS $M/G/1$ queue are insensitive (the same as the geometric distribution in the $M/M/1$ queue with the same traffic intensity). Some general treatments of insensitivity are in Whittle (1986) and Miyazawa (1993), where references to earlier work can also be found.

Kelly (1991) is a standard reference for loss systems.

4 Quasi–Reversibility and Single–Node Symmetric Queues

We consider a simple queue with multiple customer classes $c \in \mathscr{C}$ (the set $\mathscr{C}$ of classes is assumed finite or countable). A simple example is several types of customers with each their service requirements, say cars or trucks that arrive at a gas station. However, a main motivation that first becomes apparent when we proceed to networks is non–Markovian routing; cf. Example 5.1.

We assume that the time evolution of the queue can be modelled by an ergodic Markov process $\{X_t\}_{t\geq 0}$ with a finite or countable state space E. We assume given subsets A_c, D_c, $c \in \mathscr{C}$, of $E \times E$ such that $i \neq j$ when $ij \in A_c$ or $ij \in D_c$ and that $A_c \cap D_c = \emptyset$. Most often, the interpretation is that $X_{t-} = i, X_t = j$ corresponds to an arrival of a customer of class c when $ij \in A_c$ and to a departure when $ij \in D_c$, and one has $ij \in A_c \iff ji \in D_c$ (but see Problem 4.2).

Example 4.1 For a simple example, let customers of class c have Poisson arrivals at rate β_c and exponential services of rate δ_c, and let the queueing discipline be FIFO with no priorities among classes. Then we can take

$$E \;=\; \{i = c_1c_2\ldots c_n : n \in \mathbb{N},\; c_1,\ldots,c_n \in \mathscr{C}\},$$

with $X_t = c_1c_2\ldots c_n$ indicating that there are n customers in the system, with the one currently being served of class c_1, the next waiting in line of class c_2 and so on (the empty word corresponding to $n = 0$ is the idle state). Then $ij \in A_c$ precisely when i, j are of the form $i = c_1c_2\ldots c_n$, $j = c_1c_2\ldots c_nc$, and $ij \in D_c$ precisely when i, j are of the form $i = cc_1\ldots c_n$, $j = c_1\ldots c_n$. We will see more complicated examples later on, say with phase–type service times where E also contains information on phases. □

The random sets $N_c^{(+)}(t) = \bigl\{s \geq t : (X_{s-}X_s) \in A_c\bigr\}$, $N_c^{(-)}(t) = \bigl\{0 \leq s \leq t : (X_{s-}X_s) \in D_c\bigr\}$ represent the arrival process of class c customers after t, resp. the departure process prior to t (they can be viewed as measurable elements of the space of counting measures on the appropriate intervals). The queue is called *quasi–reversible* if X_t and all $N_c^{(+)}(t)$, $N_c^{(-)}(t)$, $c \in \mathscr{C}$, are independent in the steady state for any $t \geq 0$.

Proposition 4.2 *For a stationary quasi–reversible queue, the arrival processes $N_c^{(+)}(0)$, $c \in \mathscr{C}$, are independent Poisson processes, and so are the departure processes $N_c^{(-)}(\infty)$, $c \in \mathscr{C}$.*

Proof. It is clear by stationarity that $N_c^{(+)}(0)$ has stationary increments so to see that it is Poisson, it suffices to show that the increments, say $I_1, \ldots, I_n$, over intervals $[t_0, t_1), [t_1, t_2), \ldots, [t_{n-1}, t_n)$ with $0 < t_0 < t_1 <$

$\cdots < t_n$ are independent. However,

$$\begin{aligned}\mathbb{E}\big[f_1(I_1)\cdots f_n(I_n)\big] &= \mathbb{E}\,\mathbb{E}\big[f_1(I_1)\cdots f_n(I_n)\,\big|\,\mathscr{F}_{t_1}\big]\\ &= \mathbb{E}\,\big\{f_1(I_1)\,\mathbb{E}\big[f_2(I_2)\cdots f_n(I_n)\,\big|\,\mathscr{F}_{t_1}\big]\big\}\\ &= \mathbb{E}\,\big\{f_1(I_1)\,\mathbb{E}\big[f_2(I_2)\cdots f_n(I_n)\,\big|\,X_{t_1}\big]\big\}\\ &= \mathbb{E}f_1(I_1)\,\mathbb{E}\big[f_2(I_2)\cdots f_n(I_n)\big],\end{aligned}$$

where we used the Markov property in the third step and quasi–reversibility in the last. Proceeding in the same way, the expectation becomes factorized as $\mathbb{E}f_1(I_1)\cdots\mathbb{E}f_2(I_2)$, proving independence. The independence for different c follows in just the same way, as does the case of the departure processes. □

The definition of quasi–reversibility immediately implies:

Corollary 4.3 *The timereverse $\{\widetilde{X}_t\}$ of a quasi–reversible queue is itself quasi–reversible corresponding to $\widetilde{A}_c = \{ij : ji \in D_c\}$, $\widetilde{D}_c = \{ij : ji \in A_c\}$.*

Write $\boldsymbol{\Lambda} = \big(\lambda(i,j)\big)$ for the intensity matrix of $\{X_t\}$ and $\boldsymbol{\pi} = (\pi_i)$ for the stationary distribution.

Corollary 4.4 *For a stationary quasi–reversible queue, the rates $\mu_c^{(+)}, \mu_c^{(-)}$ of $N_c^{(+)}(0)$, $N_c^{(-)}(\infty)$ are given by*

$$\mu_c^{(+)} = \sum_{j:\, ij\in A_c} \lambda(i,j), \tag{4.4}$$

$$\mu_c^{(-)} = \frac{1}{\pi_j}\sum_{i:\, ij\in D_c} \pi_i\lambda(i,j). \tag{4.5}$$

In particular, (4.4) *does not depend on i and* (4.5) *not on j.*

Proof. The naive way to compute $\mu_c^{(+)}$ is of course as $\sum_{ij\in A_c}\pi_i\lambda(i,j)$ (condition upon the state i just before the arrival). That the stronger (4.4) holds follows by quasi–reversibility implying independence of i. Corollary 4.3 then yields independence of $\sum_{i:\, ji\in\widetilde{A}_c}\widetilde{\lambda}(j,i)$ of j, but this sum is just the r.h.s. of (4.5). □

Conversely:

Proposition 4.5 *If the r.h.s. of* (4.4) *does not depend on i and the r.h.s. of* (4.5) *not on j, then the queue is quasi–reversible.*

Proof. Clearly, $N_c^{(+)}(t)$ is a Poisson process with stochastic intensity $\sum_{j:\, X_{s-}j\in A_c}\lambda(X_{s-},j)$, $s > t$. However, by assumption this intensity does not depend on s, in particular it is unaffected by X_t and therefore $N_c^{(+)}(t)$ is independent of X_t. Similarly, the whole set $(N_c^{(+)}(t))_{c\in\mathscr{C}}$ is independent of X_t, and considering the time–reversed process shows that the same is true for $(N_c^{(-)}(t))_{c\in\mathscr{C}}$. To complete the proof, just observe that the independence of X_t, $\big(N_c^{(+)}(t)\big)_{c\in\mathscr{C}}$, $\big(N_c^{(-)}(t)\big)_{c\in\mathscr{C}}$ then follows from $\big(N_c^{(+)}(t)\big)_{c\in\mathscr{C}}$, $\big(N_c^{(-)}(t)\big)_{c\in\mathscr{C}}$ being conditionally independent given X_t (the Markov property). □

Example 4.6 Consider first a single class queue modelled as a standard birth–death process $\{X_t\}$ with birth rate β_k in state $k = 0, 1, \ldots$ and death rate δ_k in state $k = 1, 2, \ldots$. We take $A = \{01, 12, \ldots\}$, $D = \{10, 21, \ldots\}$. For quasi–reversibility, the arrival process must be Poisson, i.e. $\beta_n = \beta$, and this is in fact also a sufficient condition. To see this, use Proposition 4.5. The r.h.s. of (4.4) is β_i, which we assumed independent of i. The r.h.s. of (4.5) is $\pi_j^{-1}\pi_{j+1}\delta_{j+1}$, which according to the local balance equations $\pi_j\beta_j = \pi_{j+1}\delta_{j+1}$ is β_j, again assumed to be independent of j. See also the proof of Proposition 1.1. □

Example 4.7 (MULTICLASS $M/M/1$ QUEUES) Consider the model of Example 4.1. Let $\beta = \sum_{c\in\mathscr{C}}\beta_c$ denote the overall arrival rate and assume that $\beta < \infty$ (e.g. that $\mathscr{C}$ is finite) and that the service intensity $\delta = \delta_c$ is independent of c. Denote by $p_c = \beta_c/\beta$ the probability that an arriving customer is of class c. The traffic intensity is $\rho = \beta/\delta$.

Our candidate for the stationary distribution is

$$\pi_{c_1\ldots c_n} \;=\; (1-\rho)\rho^n p_{c_1}\ldots p_{c_n},$$

and we first verify that this is indeed the correct one by means of Kelly's lemma. Since the possible transitions of $\{X_t\}$ out of state (word) $c_1\ldots c_n$ are to either the state obtaining by deleting the first letter c_1 in the word or to a state obtained by adding a letter in the end, the possible transitions of the time–reversed process out of state $c_1\ldots c_n$ are to the state obtaining by deleting the last letter c_n in the word or to a state obtained by adding a letter at the beginning. That is, informally the queueing system is similar except that c_n now is the customer being currently served, etc., and thus the trial candidate for the nonzero off-diagonal time–reversed intensities are

$$\widetilde{\lambda}(c_1\ldots c_n, c_1\ldots c_{n-1}) \;=\; \delta,\;\; \widetilde{\lambda}(c_1\ldots c_n, cc_1\ldots c_n) \;=\; \beta_c.$$

In this setting, we must verify $\pi_i\lambda(i,j) = \pi_j\widetilde{\lambda}(j,i)$ for $i \neq j$. The only cases where not both sides are 0 are (1) $i = c_1\ldots c_n$, $j = c_1\ldots c_n c$ and (2) $i = c_1\ldots c_n$, $j = c_2\ldots c_n$. In case (1),

$$\frac{\pi_i\lambda(i,j)}{\pi_j\widetilde{\lambda}(j,i)} \;=\; \frac{\pi_i\beta_c}{\pi_j\delta} \;=\; \frac{(1-\rho)\rho^n p_{c_1}\ldots p_{c_n}\beta_c}{(1-\rho)\rho^{n+1}p_{c_1}\ldots p_{c_n}p_c\delta} \;=\; \frac{\beta_c}{\rho p_c\delta} \;=\; \frac{\beta}{\rho\delta} \;=\; 1,$$

and in case (2),

$$\frac{\pi_i\lambda(i,j)}{\pi_j\widetilde{\lambda}(j,i)} \;=\; \frac{\pi_i\delta}{\pi_j\beta_{c_1}} \;=\; \frac{(1-\rho)\rho^n p_{c_1}\ldots p_{c_n}\delta}{(1-\rho)\rho^{n-1}p_{c_2}\ldots p_{c_n}\beta_{c_1}} \;=\; \frac{\rho p_{c_1}\delta}{\beta_{c_1}} \;=\; \frac{\rho\delta}{\beta} \;=\; 1.$$

It follows that for $j = c_1 \dots c_n$,

$$\begin{aligned}\frac{1}{\pi_j}\sum_{i:\, ij \in D_c} \pi_i \lambda(i,j) &= \frac{1}{(1-\rho)\rho^n p_{c_1} \dots p_{c_n}} \sum_{c\in\mathscr{C}} (1-\rho)\rho^{n+1} p_c p_{c_1} \dots p_{c_n} \delta \\ &= \rho\delta \sum_{c\in\mathscr{C}} p_c \;=\; \rho\delta \;=\; \beta\end{aligned}$$

is independent of j, whereas trivially $\sum_{j:\, ij\in A_c} \lambda(i,j)$ equals β for all i. Hence the system is quasi–reversible by Proposition 4.5. □

In most cases, of course $\mu_c^{(+)} = \mu_c^{(-)}$ (but see Problem 4.1). Proposition 4.4 immediately gives:

Corollary 4.8 *For a quasi–reversible queue with $\mu_c^{(+)} = \mu_c^{(-)}$ for all $c \in \mathscr{C}$, the stationary distribution satisfies*

$$\pi_i \sum_{j:\, ij\in A_c} \lambda(i,j) \;=\; \sum_{j:\, ji \in D_c} \pi_j \lambda(j,i). \tag{4.6}$$

We now introduce a general framework for the study of quasi–reversibility (and insensitivity), having the advantage that in many examples one can deal with general (or rather phase–type) service times, the so–called *symmetric queues* (more precisely, the system we consider could be called a symmetric multiclass $M/PH/\cdot$ queue). Different customer classes have independent Poisson arrivals with rate β_c for class c. Let $\beta = \sum_{c\in\mathscr{C}}\beta_c$ denote the overall arrival rate and assume $\beta < \infty$. We further assume that class c customers have a phase–type service time distribution B_c, say with initial vector $\boldsymbol{\alpha}_c = (\alpha_{cr})_{r=1,\dots,p_c}$ and phase generator $\boldsymbol{T}_c = (t_{crs})$. The exit rate vector is $\boldsymbol{t}_c = -\boldsymbol{T}_c\mathbf{1} = (t_{cr})$, the mean service time is $m_c = -\boldsymbol{\alpha}_c\boldsymbol{T}_c^{-1}\mathbf{1}$ and the overall traffic intensity is $\rho = \sum_{c\in\mathscr{C}}\rho_c$ where $\rho_c = \beta_c m_c$. Let further

$$\boldsymbol{\nu}_c \;=\; (\nu_{cr}) \;=\; -\frac{\boldsymbol{\alpha}_c\boldsymbol{T}_c^{-1}}{m_c}, \quad p_{cr} \;=\; \frac{\beta_c m_c}{\rho}\nu_{cr};$$

$\boldsymbol{\nu}_c$ is the equilibrium distribution of the phase of service of a class c customer and p_{cr} is the probability that a randomly selected customer in service is of class c and in phase r of service. This is an important intuition behind the formula (4.8) for the stationary distribution given below.

We think of the customers as ordered in positions $k = 1, 2, \dots$ in front of the server (only finitely many are present at a given time) and refer for brevity to the customer in position k as just customer k. A state of the system has the form $i = c_1 r_1 \dots c_n r_n$ and indicates that there are $n = n(i)$ customers, such that customer k is of class c_k and has current phase r_k of service. The server works at rate $\phi(n)$ when there are n customers in the systems and then devotes the fraction $\omega(n,k)$ of his capacity to customer k (thus $\bigl(\omega(n,k)\bigr)_{k=1}^{n}$ is a probability vector for each $n \geq 1$). Further, it is assumed that a customer arriving to a system with n customers takes position $k = 1, \dots, n+1$ w.p. $\omega(n+1,k)$ independently of his class.

Note that the ω function has a double interpretation: it is the same for position allocation as for service (hence the term *symmetric*). The "current phase" of a class c customer is chosen according to $\boldsymbol{\alpha}_c$ upon arrival, even if the customer may not physically start service at once. If he has been interrupted in service, the "current phase" is the one in which interruption occured.

To describe intensities and transitions, denote by $i(k)$ the state obtained by deleting the customer $k = 1, \ldots, n(i)$ (thus customers $1, \ldots, k-1$ maintain their position and customers $k+1, \ldots, n(i)$ move up one position). Similarly, $i(k, c, r)$ is the state where a customer of class c and in phase r of service is added at position k (thus customers $1, \ldots, k-1$ maintain their position and customers $k, \ldots, n(i)$ move down one position). Finally, $i(k, s)$ is the state where customer k (of class c_k) has changed the phase of service from r_k to s.

It follows that the possible transitions are

$$\begin{array}{lcrl} & \nearrow & i(k,s) & \phi(n)\omega(n,k)t_{c_k r_k s} \\ i = c_1 r_1 c_2 r_2 \ldots c_n r_n & \rightarrow & i(k) & \phi(n)\omega(n,k)t_{c_k r_k} \\ & \searrow & i(k,c,r) & \beta_c \omega(n+1,k)\alpha_{cr} \end{array} \tag{4.7}$$

with the right column giving the intensities, and we can take

$$\begin{aligned} A_c &= \bigl\{ij : j = i(k,c,r) \text{ for some } k, c, r\bigr\}, \\ D_c &= \bigl\{ij : j = i(k) \text{ for some } k\bigr\}. \end{aligned}$$

Theorem 4.9 *Let*

$$\Phi(n) \;=\; \prod_{k=1}^{n} \phi(k), \quad \delta \;=\; \sum_{n=0}^{\infty} \frac{\rho^n}{\Phi(n)}\,.$$

If $\delta < \infty$, then the symmetric multiclass $M/PH/\cdot$ queue is ergodic and the stationary distribution is given by

$$\pi_i \;=\; \pi_{c_1 r_1 \ldots c_n r_n} \;=\; \delta^{-1} \frac{\rho^n}{\Phi(n)} \prod_{k=1}^{n} p_{c_k r_k}\,. \tag{4.8}$$

Further the queue is quasi–reversible with $\mu_c^{(+)} = \mu_c^{(-)}$.

Proof. Since $\sum_{c_k r_k} p_{c_k r_k} = 1$, we have

$$\sum_{n,\, c_1 r_1 \ldots c_n r_n} \frac{\rho^n}{\Phi(k)} \cdot \prod_{k=1}^{n} p_{c_k r_k} \;=\; \delta.$$

Thus indeed (4.8) is a probability distribution when $\delta < \infty$, and since the process is nonexplosive by Problem II.2.4, we are again in a position to apply Kelly's lemma II.5.4. As in Section 3 our trial candidate for the time–reversed system is the given system with the phase representations reversed. That is, the changed parameters are

$$\widetilde{\alpha}_{cr} = m_c t_{cr} \nu_{cr}, \quad \widetilde{t}_{crs} = \frac{\nu_{cs} t_{csr}}{\nu_{cr}}, \quad \widetilde{t}_{cr} = \frac{\alpha_{cr}}{m_c \nu_{cr}}.$$

With this trial system (for the intensities, just add tildes in the right column of (4.7)), we must verify $\pi_i \lambda(i,j) = \pi_j \widetilde{\lambda}(j,i)$ for $i \neq j$.

Let first $j = i(k,s)$. Then $\pi_i/\pi_j = \nu_{c_k r_k}/\nu_{c_k s}$ so that (4.7) yields

$$\frac{\pi_i}{\pi_j} \cdot \frac{\lambda(i,j)}{\widetilde{\lambda}(j,i)} = \frac{\nu_{c_k r_k}}{\nu_{c_k s}} \cdot \frac{t_{c_k r_k s}}{\widetilde{t}_{c_k s r_k}} = \frac{\nu_{c_k r_k}}{\nu_{c_k s}} \cdot \frac{t_{c_k r_k s}}{\nu_{c_k r_k} t_{c_k r_k s}/\nu_{c_k s}} = 1.$$

If $j = i(k)$, then $i = j(k, c_k, r_k)$ and we get

$$\frac{\pi_i}{\pi_j} \cdot \frac{\lambda(i,j)}{\widetilde{\lambda}(j,i)} = \frac{\rho_{c_k} \nu_{c_k r_k}}{\phi(n)} \cdot \frac{\phi(n)\omega(n,k) t_{c_k r_k}}{\beta_{c_k} \omega(n,k) \widetilde{\alpha}_{c_k r_k}} = \frac{\rho_{c_k} \nu_{c_k r_k} t_{c_k r_k}}{\beta_{c_k} m_{c_k} t_{c_k r_k} \nu_{c_k r_k}} = 1.$$

Finally, the case $j = i(k,c,r)$ follows from the case $j = i(k)$ by interchanging the given system and the tilded one.

To verify quasi–reversibility, we have

$$\begin{aligned}\sum_{j:\, ij \in A_c} \lambda(i,j) &= \sum_{k,c,s} \lambda\big(i, i(k,c,s)\big) = \sum_{k,c,s} \beta_c \omega(n+1,k) \alpha_{c,s} \\ &= \sum_{c,s} \beta_c \alpha_{c,s} = \sum_c \beta_c = \beta.\end{aligned}$$

This is independent of i so that (4.4) does not depend on i. Therefore by symmetry, also (4.4) holds for the tilded system which is the same as saying that (4.5) holds for the given system. □

Example 4.10 (MULTICLASS PREEMPTIVE $LCFS$ $M/PH/1$ QUEUES) This corresponds to $\phi(n) = 1$, $\omega(n,1) = 1$, all other $\omega(n,k) = 1$. □

Example 4.11 (MULTICLASS PS $M/PH/1$ QUEUES) This corresponds to $\phi(n) = 1$, and all $\omega(n,k) = 1/n$. □

Example 4.12 (MULTICLASS $M/PH/\infty$ QUEUES) This corresponds to $\phi(n) = n$, and all $\omega(n,k) = 1/n$. □

Corollary 4.13 *The queue length distributions in the multiclass $M/PH/\cdot$ queues in Examples 4.10–4.12 are insensitive in the sense that the steady–state distributions only depend on the B_c through their means m_c.*

Proof. The probability of queue length n is $\rho^n/\delta\Phi(n)$, and here ρ, δ depend only on the B_c through the m_c. □

Problems

4.1 Consider the $M/M/1$ queue with two customer types 0,1 arriving with intensities β_0, β_1. The state of the system is $n0$ if n customers are present and the last arrival was of type 0, and similarly for $n1$. Interpreting both types to be of a single class, take $D = \{((n+1)i, ni) : i = 0,1\}$ and $A = \{(ni, (n+1)0) : i = 0,1\}$.

Show that the queue is quasi–reversible but that the arrival and departure rates are not the same.

4.2 Consider a finite birth–death process on $\{0, 1, \ldots, L\}$ with death rates δ_k and birth rate $\beta = \beta_k$ independent of $k = 0, \ldots, L-1$. (a) Show that the naive definition of A, D, etc. does not lead to quasi–stationarity. (b) Splitting state L into two, let $E = \{0, 1, \ldots, L-1, L_0, L_1\}$ and let an (dummy) arrival in state L_0 trigger a transition (flip) to L_1 and similarly for arrivals in state L_1. Show that appropriate definitions of A, D, etc. lead to quasi–stationarity.

5 Quasi–Reversibility in Networks

We now consider a queueing network of K nodes. Let a state space $E^{(k)}$ be associated with the kth, and let $A_c^{(k)}, D_c^{(k)}$ be disjoint classes of transitions that we think of as arrivals, resp. departures. The set $\mathscr{C}$ of classes is the same for all nodes and it is assumed that class c customers arrive at node k according to a Poisson process with intensity α_{kc}, and that these Poisson processes are independent for different kc. The routing is Markovian in the sense that if a transition $i_k \to j_k$ with $i_k j_k \in D_c^{(k)}$ occurs at node k at time t, then a node ℓ and and a possibly different class d is selected with probability $\gamma_{kc,\ell d}$, such that a class d arrival occurs at node ℓ at the same time; $\gamma_{kc,0} = 1 - \sum_{\ell c} \gamma_{kc,\ell d}$ then gives the probability of a departure from the network.

Example 5.1 A major restriction of Jackson networks as treated in Section 2 is that of Markovian routing: the node ℓ to which a customer goes after leaving node k is chosen with a probability depending only on the current node k visited and not on which nodes were earlier visited and how many times. However, in a number of applications the routing is not Markovian and incorporating this is indeed a major motivation of the multiclass set–up.

A particular case that is often met is that a customer selects among a set $\mathscr{R}_1, \mathscr{R}_2, \ldots$ of possible routes w.p. say $q(\mathscr{R}_r)$ for $\mathscr{R}_r$ (each route $\mathscr{R}$ is a finite string $k_1 \ldots k_t$ of nodes). The set $\mathscr{C}$ of classes is then the set of similar finite strings $c = \ell_1 \ldots \ell_t$, such that $\ell_1, \ldots, \ell_t$ gives the remaining set of nodes (including the current one) to be visited by the customer, which upon leaving node ℓ_1 becomes a $\ell_2 \ldots \ell_t$ customer; thus, $\gamma_{kc,\ell d} = 1$ when $c = \ell_1 \ldots \ell_t$, $d = \ell_2 \ldots \ell_t$, $k = \ell_1$, $\ell = \ell_2$, and otherwise $\gamma_{kc,\ell d} = 0$.

Markovian routing as for a Jackson network corresponds to

$$q_{k_1 \ldots k_t} = \frac{\alpha_{k_1}}{\sum_1^K \alpha_{k_i}} \gamma_{k_1 k_2} \cdots \gamma_{k_{t-1} k_t} \gamma_{k_t 0}.$$

Note that here indeed $\mathscr{C}$ is countable rather than finite. □

As for a Jackson network we arrive at the traffic (throughput) equations

$$\beta_{kc} \;=\; \alpha_{kc} + \sum_{\ell d} \beta_{\ell d}\gamma_{\ell d,kc}, \quad k = 1, \ldots, K, \; c \in \mathscr{C}. \tag{5.1}$$

In the following, we let $(\beta_{kc})_{k=1,\ldots,K,\; c\in\mathscr{C}}$ be some fixed solution (assuming existence; uniqueness is not essential).

The basic assumption about node k is now loosely that it is quasi–reversible when operating in isolation with Poisson input. The precise content of this is that we assume given an ergodic intensity matrix $\bigl(\lambda^{(k)}(i,j)\bigr)_{i,j\in E^{(k)}}$ with stationary distribution say $(\pi_i^{(k)})_{i\in E^{(k)}}$, such that the corresponding Markov process on $E^{(k)}$ is quasi–reversible when the common input and output rates are given by (5.1); according to (4.4), (4.5), this means

$$\beta_{kc} \;=\; \sum_{j_k\,:\, i_kj_k\in A_c^{(k)}} \lambda^{(k)}(i_k,j_k) \;=\; \frac{1}{\pi_{j_k}^{(k)}} \sum_{i_k\,:\, i_kj_k\in D_c^{(k)}} \pi_i^{(k)}\lambda^{(k)}(i_k,j_k)\,. \tag{5.2}$$

The network has state space $E = \prod_1^K E^{(k)}$, a typical state being denoted $\boldsymbol{i} = (i_k)_{k=1,\ldots,K}$ in the following, and for the Markov process $\{\boldsymbol{X}_t\}$ describing the network, we write similarly $\boldsymbol{X}_t = (X_t^{(1)},\ldots,X_t^{(K)})$. The nonzero off–diagonal intensities are defined as follows:

(a) $\lambda(\boldsymbol{i},\boldsymbol{j}) = \dfrac{\alpha_{kc}\lambda^{(k)}(i_k,j_k)}{\beta_{kc}}$ if $\boldsymbol{i},\boldsymbol{j}$ differ only at node k and $i_kj_k \in A_c^{(k)}$;
(b) $\lambda(\boldsymbol{i},\boldsymbol{j}) = \lambda^{(k)}(i_k,j_k)\gamma_{kc,0}$ if $\boldsymbol{i},\boldsymbol{j}$ differ only at node k and $i_kj_k \in D_c^{(k)}$;
(c) $\lambda(\boldsymbol{i},\boldsymbol{j}) = \lambda^{(k)}(i_k,j_k)$ if $\boldsymbol{i},\boldsymbol{j}$ differ only at node k and i_kj_k do not belong to any $A_c^{(k)}$ or $D_c^{(k)}$;
(d) $\lambda(\boldsymbol{i},\boldsymbol{j}) = \lambda^{(k)}(i_k,j_k)\gamma_{kc,\ell d}\dfrac{\lambda^{(\ell)}(i_\ell,j_\ell)}{\beta_{\ell d}}$ if $\boldsymbol{i},\boldsymbol{j}$ differ only at nodes k,ℓ and $i_kj_k \in D_c^{(k)}$, $i_\ell j_\ell \in A_c^{(k)}$.

Note in (a) that $\lambda^{(k)}(i_k,j_k)/\beta_{kc}$ is the probability (in node k in isolation or in the network) that a class c customer arriving to node k and seeing state i_k will trigger a transition to state j_k, and analogously for $\lambda^{(\ell)}(i_\ell,j_\ell)/\beta_{\ell d}$ in (d); that these probabilities should coincide for the nodes in isolation and in the network is the crux in how to build the nodes together to a network.

The main results follows: as for a Jackson network, the stationary distribution is of product–form.

Theorem 5.2 *The network of quasi–reversible nodes is ergodic with stationary distribution of product form, i.e.*

$$\pi_{\boldsymbol{i}} \;=\; \pi_{i_1\ldots i_K} \;=\; \pi_{i_1}^{(1)}\ldots\pi_{i_K}^{(K)} \tag{5.3}$$

Proof. We will provide trial values $\widetilde{\lambda}(\boldsymbol{i},\boldsymbol{j})$ of the time–reversed intensities and verify that $\pi_{\boldsymbol{i}}\lambda(\boldsymbol{i},\boldsymbol{j}) = \pi_{\boldsymbol{j}}\widetilde{\lambda}(\boldsymbol{j},\boldsymbol{i})$ for all $\boldsymbol{i},\boldsymbol{j}$. The trial is another network of quasi–reversible nodes, with $A_c^{(k)}$ replaced by $\widetilde{A}_c^{(k)} = \bigl\{i_kj_k : j_ki_k \in D_c^{(k)}\bigr\}$, $D_c^{(k)}$ by $\widetilde{D}_c^{(k)} = \bigl\{i_kj_k : j_ki_k \in A_c^{(k)}\bigr\}$, $\lambda^{(k)}(i_k, j_k)$ by $\widetilde{\lambda}^{(k)}(i_k, j_k) = \pi_{j_k}^{(k)}\lambda^{(k)}(j_k, i_k)/\pi_{i_k}^{(k)}$ and arrival intensities and routing probabilities by

$$\widetilde{\alpha}_{kc} = \beta_{kc}\gamma_{kc,0}, \quad \widetilde{\gamma}_{kc,\ell d} = \frac{\beta_{\ell d}\gamma_{\ell d,kc}}{\beta_{kc}}, \quad \widetilde{\gamma}_{kc,0} = \frac{\alpha_{kc}}{\beta_{kc}}.$$

To see that these are reasonable guesses, note for example that the flow of ℓd customers to kc customers is $\beta_{\ell d}\gamma_{\ell d,kc}$ and must coincide with the flow $\widetilde{\beta}_{kc}\widetilde{\gamma}_{kc,\ell d}$ of kc customers to ℓd customers in the time–reversed network; however, since the input and output rate of class c customers at node k were assumed equal, we must have $\widetilde{\beta}_{kc} = \beta_{kc}$. Note also that $1 = \widetilde{\gamma}_{kc,0} + \sum_{\ell d}\widetilde{\gamma}_{kc,\ell d}$ is a consequence of the throughput equations (5.1).

To verify $\pi_{\boldsymbol{i}}\lambda(\boldsymbol{i}, \boldsymbol{j}) = \pi_{\boldsymbol{j}}\widetilde{\lambda}(\boldsymbol{j}, \boldsymbol{i})$, let first $\boldsymbol{i}, \boldsymbol{j}$ be as in (a) above. Then a transition from $\boldsymbol{j}$ to $\boldsymbol{i}$ in the tilded system corresponds to a departure from node k accompanied of a transition from j_k to i_k. The intensity is given by (b) with tildes added and hence

$$\frac{\pi_{\boldsymbol{i}}\lambda(\boldsymbol{i}, \boldsymbol{j})}{\pi_{\boldsymbol{j}}\widetilde{\lambda}(\boldsymbol{j}, \boldsymbol{i})} = \frac{\pi_{i_k}^{(k)}\alpha_{kc}\lambda^{(k)}(i_k, j_k)/\beta_{kc}}{\pi_{j_k}^{(k)}\widetilde{\lambda}^{(k)}(j_k, i_k)\widetilde{\gamma}_{kc,0}} = \frac{\alpha_{kc}/\beta_{kc}}{\widetilde{\gamma}_{kc,0}} = 1.$$

Case (b) is symmetric, whereas (c) follows trivially once one notices that $i_kj_k \notin A_c^{(k)} \cup D_c^{(k)}$ if and only if $j_ki_k \notin \widetilde{A}_c^{(k)} \cup \widetilde{D}_c^{(k)}$. Finally in (d),

$$\begin{aligned}\frac{\pi_{\boldsymbol{i}}\lambda(\boldsymbol{i}, \boldsymbol{j})}{\pi_{\boldsymbol{j}}\widetilde{\lambda}(\boldsymbol{j}, \boldsymbol{i})} &= \frac{\pi_{i_k}^{(k)}\pi_{i_\ell}^{(\ell)}\lambda^{(k)}(i_k, j_k)\gamma_{kc,\ell d}\lambda^{(\ell)}(i_\ell, j_\ell)/\beta_{\ell d}}{\pi_{j_k}^{(k)}\pi_{j_\ell}^{(\ell)}\widetilde{\lambda}^{(\ell)}(j_\ell, i_\ell)\widetilde{\gamma}_{\ell d,kc}\widetilde{\lambda}^{(k)}(j_k, i_k)/\beta_{kc}} \\ &= \frac{\gamma_{kc,\ell d}/\beta_{\ell d}}{\widetilde{\gamma}_{\ell d,kc}/\beta_{kc}} = 1.\end{aligned}$$

□

6 The Arrival Theorem

In a single–node queue with Poisson arrivals, the PASTA property to be formalized in VII.6 implies that the steady–state distribution of the state of the queue seen by a customer just before his arrival is the same as the steady–state distribution of the state of the queue at an arbitrary point of time. We will show here that results of rather similar form hold for networks.

In the network we have considered, a customer going to node ℓ after being served at node k does so instantaneously. However, we may imagine that he observes the state of the rest of the network during the infinitesimal interval where the transition takes place. Typically, these transition instants do not form a Poisson process so PASTA does not apply. Nevertheless, we will show here that results of rather similar form hold. First:

Theorem 6.1 *In the model of Section 5, the steady–state distribution $\boldsymbol{\eta}^{kc,\ell d}$ of the state seen by a customer in transition from node k as a class c customer to node ℓ as a class d customer coincides with the time–stationary distribution $\boldsymbol{\pi}$.*

Proof. Clearly, the rate of a kc customer making a transition to a ℓd customer is $\beta_{kc}\gamma_{kc,\ell d}$. To determine $\eta_{\boldsymbol{j}}^{kc,\ell d}$, we note that if the state seen during the transition is $\boldsymbol{j}$, then the state $\boldsymbol{i}$ just before the transition must be in the set $E_{\boldsymbol{j}}^{kc,\ell d}$ of states with $i_m = j_m$, $m \neq k$, and $i_k j_k \in D_c^{(k)}$. Thus

$$\begin{aligned}
\eta_{\boldsymbol{j}}^{kc,\ell d} &= \frac{1}{\beta_{kc}\gamma_{kc,\ell d}} \sum_{\boldsymbol{i} \in E_{\boldsymbol{j}}^{kc,\ell d}} \boldsymbol{\pi}_{\boldsymbol{i}} \lambda(\boldsymbol{i},\boldsymbol{j}) \gamma_{kc,\ell d} \\
&= \frac{1}{\beta_{kc}} \prod_{m \neq k} \pi_{j_m}^{(m)} \cdot \sum_{i_k :\, i_k j_k \in D_c^{(k)}} \pi_{i_k}^{(k)} \lambda^{(k)}(i_k, j_k) \\
&= \frac{1}{\beta_{kc}} \prod_{m \neq k} \pi_{j_m}^{(m)} \cdot \pi_{j_k}^{(k)} \beta_{kc} = \prod_{m=1}^{K} \pi_{j_m}^{(m)} = \pi_{\boldsymbol{j}},
\end{aligned}$$

using (5.2) in the third step. □

Corollary 6.2 *Let $\mathscr{H} = \{kc\}$, $\mathscr{K} = \{\ell d\}$ be arbitrary subsets of $\{0, \ldots, K\} \times \mathscr{C}$. Then the steady–state distribution of a $\mathscr{H}$ customer in transition to become a $\mathscr{K}$ customer is $\boldsymbol{\pi}$.*

Proof. Letting $\theta_{kc,\ell d}$ denote the probability that a $\mathscr{H}$customer in transition to become a $\mathscr{K}$ customer is a kc customer just before the transition and a ℓd customer just after, the distribution in question has point mass

$$\sum_{kc \in \mathscr{H},\ \ell d \in \mathscr{K}} \theta_{kc,\ell d}\, \eta_{\boldsymbol{j}}^{kc,\ell d} = \sum_{kc \in \mathscr{H},\ \ell d \in \mathscr{K}} \theta_{kc,\ell d}\, \pi_{\boldsymbol{j}} = \pi_{\boldsymbol{j}}$$

at $\boldsymbol{j}$ (note that the case $k = 0$ is covered by PASTA). □

Remark 6.3 If $\mathscr{K} = \{\ell\} \times \mathscr{C}$, $\mathscr{H} = \{0, \ldots, K\} \times \mathscr{C}$, then the statement of Corollary 6.2 means that a customer arriving at node ℓ (as an external arrival or from some other node k) sees distribution $\boldsymbol{\pi}$ just at the arrival instant (in the above sense of being in transition). Similarly, taking $\mathscr{H} = \{k\} \times \mathscr{C}$, $\mathscr{K} = \{0, \ldots, K\} \times \mathscr{C}$, shows that a customer departing node k sees distribution $\boldsymbol{\pi}$ just at the departure instant. □

The arrival theorem takes a striking form for closed Jackson networks. In the notation of Theorem 2.3:

Theorem 6.4 *For a closed Jackson network with N customers, the steady–state distribution $\boldsymbol{\eta}^{(N;k,\ell)}$ of the state seen by a customer in transi-*

tion from node k to node ℓ coincides with the time–stationary distribution $\boldsymbol{\pi}^{(N-1)}$ of the network with one customer removed. That is, $\eta_{\boldsymbol{n}}^{(N;k,\ell)} = C_{N-1}\prod_{m=1}^{K}\rho_m^{n_m}$ when $n_1+\cdots+n_K = N-1$.

Proof. In a closed network, the β_k are only determined up to a constant and can therefore not a priori be identified with the throughput rates, say $\beta_1^*,\ldots,\beta_K^*$. Instead, we can identify β_k^* with the output rate $\delta_k\theta_k =$ the service rate δ_k times the probability θ_k of node k being busy where

$$\theta_k = C_N \sum_{n_1+\cdots+n_K=N,n_k>0}\prod_{\ell=1}^{K}\rho_k^{n_k} = C_N\rho_k \sum_{n_1+\cdots+n_K=N-1}\prod_{m=1}^{K}\rho_m^{n_\ell} = \frac{C_N\rho_k}{C_{N-1}}$$

where the β_m (and accordingly $\rho_1,\ldots,\rho_K$ and C_N, C_{N-1}) are calculated based upon some fixed solution of $\boldsymbol{\beta} = \boldsymbol{\beta}\boldsymbol{\Gamma}$.

It follows that the rate of transitions of customers from node k to node ℓ is $\beta_k^*\gamma_{kl} = \beta_k\gamma_{kl}C_N/C_{N-1}$. Hence for $\boldsymbol{n}\in E_{N-1}$,

$$\eta_{\boldsymbol{n}}^{(N;k,\ell)} = \frac{C_{N-1}}{\beta_k\gamma_{kl}C_N}\pi_{\boldsymbol{n}_k^{(+)}}\delta_k\gamma_{kl} = \frac{C_{N-1}}{\beta_k}\rho_k\delta_k\prod_{m=1}^{K}\rho_m^{n_m} = C_{N-1}\prod_{m=1}^{K}\rho_m^{n_m}$$

when $\boldsymbol{n}\in E_{N-1}$. □

Notes Among many texts on queueing networks, we mention in particular Kelly (1979), Walrand (1988), Serfozo (1999), Chao *et al.* (1999) and Chen and Yao (2001); see also the volumes edited by Kelly and Williams (1995) and Kelly *et al.* (1996). Buzacott and Shantikumar (1993) contains a large number of applications to manufacturing system.

Extensive lists of references can be found in these texts. Some milestones in the theory exposed in Sections 1–2 and 4–6 are Jackson (1957), Gordon and Newell (1967), Baskett *et al.* (1975) and Kelly (1979). One often meets the terms *BCMP network*, meaning the models of Baskett *et al.* (which are special cases of the model of Section 5), and *Kelly network*, denoting the special case where customer classes are the routing schemes.

Queueing networks form maybe the most active and challenging area of research in queueing theory. The developments go in several directions. One is characterization of conditions for product–form solutions and the study of non-product form networks in a Markovian setting; see Serfozo (1999) and Chao *et al.* (1999). Another is stability theory (when do ergodic limits exist?) where certain unexpected phenomena occur. A classical example is the network in Fig. 5.1, dicussed in length in Chen and Yao (2001).

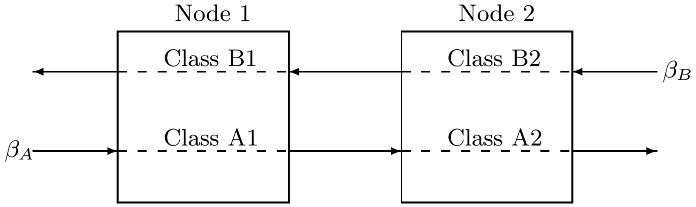

Figure 5.1

Here we have two arrival streams A,B, such that a type A customer enters node 1 as a class A1 customer and goes to node 2 as a class A2 customer after being served, and a type B customer enters node 2 as a class B2 customer and goes to node 1 as a class B1 customer after being served. Class B2 customers have preemptive priority over class A2 customers and class A1 over class B1. With μ_{A1}, etc. the mean service times, the natural stability is that the traffic intensity at each station is < 1, i.e.

$$\rho_1 = \beta_A \mu_{\mathrm{A1}} + \beta_B \mu_{\mathrm{B1}} < 1, \quad \rho_2 = \beta_A \mu_{\mathrm{A2}} + \beta_B \mu_{\mathrm{B2}} < 1. \tag{6.1}$$

However, assuming (6.1), we have $\beta_A \mu_{\mathrm{A2}} < 1$ and A2 customers block B1 customers in a fraction $\beta_A \mu_{\mathrm{A2}}$ of time in a stable system. Similarly, B2 customers block A1 customers in a fraction $\beta_B \mu_{\mathrm{B1}}$ of time, so that if $\beta_A \mu_{\mathrm{A1}} + \beta_B \mu_{\mathrm{B2}} > 1$, blocking occurs all the time and the number in system must build up, contradicting stability (obviously there are examples where this extra inequality and (6.1) hold at the same time).

Even just to show that $\rho_k < 1$ at each node in a *generalized Jackson network* (the model of Section 2 with nonexponential service times) is sufficient for stability presents considerable difficulties; see e.g. Baccelli and Foss (1994). A survey of the literature on stability of queueing networks is given p. 341 of Chao *et al.* (1999). An interesting direction not mentioned there is the connection between stability and the existence of *fluid limits* (also called *hydrodynamical limits*) (LLN type limits approximating, e.g., the netput process for the $M/G/1$ workload by the straight line $(\rho - 1)t$); see e.g. Dai (1995a,b). See also Cohen (1992), Fayolle *et al.* (1995) and Fayolle *et al.* (1999) for further stability issues.

A further active area is heavy-traffic limit theorems. With K nodes, the limit here is typically Brownian motion in $[0, \infty)^K$ and so–called *oblique reflection* at the boundary (see the Notes to IX.2 for more details). In contrast to the case $K = 1$, few characteristics of this process can be found in closed form. One reason that heavy-traffic limits are nevertheless useful is *state–space collapse*: the distinction between the customers classes, say there are M of them, vanishes so that the dimension of the state space is reduced by a factor of M. Some selected papers in this area containing further references are Reiman (1984), Harrison and Williams (1987, 1992), Bramson (1998) and Williams (1998).

Finally, we mention large deviations studies such as Glasserman and Kou (1995b), Atar and Dupuis (1999), McDonald (1999), Ignatiouk–Robert (2000) and Miyazawa (2003).

Part B: Some General Tools and Methods

V

Renewal Theory

1 Renewal Processes

Let $0 \leq S_0 \leq S_1 < S_2 < \cdots$ be the times of occurrences of some phenomenon and $Y_n = S_n - S_{n-1}$, $Y_0 = S_0$; see Fig. 1.1. Then $\{S_n\}_{n\in\mathbb{N}}$ is called a *renewal process* if $Y_0, Y_1, Y_2, \ldots$ are independent and $Y_1, Y_2, \ldots$ (but not necessarily Y_0) have the same distribution.

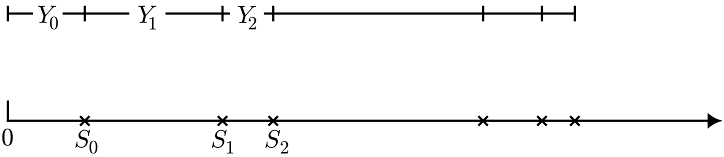

Figure 1.1

The S_n are called the *renewals* or the *epochs* of the renewal process. The common distribution F of $Y_1, Y_2, \ldots$ is the *interarrival* or *waiting–time distribution*. To avoid more than one renewal at a time we always assume that the Y_k have no mass at zero, $F(0) = 0$. The renewal process is *pure* or *zero–delayed* if $Y_0 = S_0 = 0$ a.s.. Otherwise it is *delayed* and the *delay distribution* is the distribution of Y_0. One also sometimes considers *terminating* or *transient* renewal processes, where the interarrival distribution is *defective*, i.e. may have an atom at $+\infty$, $\|F\| = \lim_{t\to\infty} F(t) < 1$ and in which case $S_n = \infty$ eventually. If $\|F\| = 1$, the renewal process is *proper*.

A main case of a renewal process is, of course, the Poisson process, where the interarrival distribution is exponential. For example, the Poisson process provides an adequate description of the emission of particles from a radioactive source. We next list some further phenomena, which it has been suggested can be modelled by renewal processes.

Example 1.1 Consider an item, say an electric bulb, that fails at times $S_0, S_1, \ldots$ and is replaced at the time of failure by a new item of the same sort. Then F is the distribution of the lifetime of an item. The process is delayed if the item present at time $t = 0$ is not new, so that its lifetime need not have distribution F. □

Example 1.2 Consider a road on which vehicles are driving in one direction only and all with the same constant velocity. Two interpretations are possible: (i) the S_n are the instants when vehicles pass a certain point on the road, (ii) the timescale $[0, \infty)$ is a map of the road and the S_n the positions of the vehicles at a certain instant. In both cases, the form of the interarrival distribution to be expected depends in an essential manner on whether there is little or much traffic on the road. In the first case (say on a rural road) F might be taken to be exponential, while in the second case (say on a main street in a city) the vehicles would rather be equally spaced, i.e. F concentrated at one point. □

Example 1.3 Consider a continuous–time Markov process $\{X_t\}_{t\geq 0}$ or a discrete–time Markov chain $\{X_n\}_{n\in\mathbb{N}}$ with discrete or continuous state space, and let i be some fixed state. Then the instants S_n of visits to i form a renewal process (provided in continuous time that the strong Markov property holds and that $\mathbb{P}_i(\tau(i) = 0) = 0$; the last requirement fails say for Brownian motion). The process is pure if and only if $X_0 = i$ and transient if and only if state i is visited only finitely often (which is the definition of transience in the case of a discrete state space).

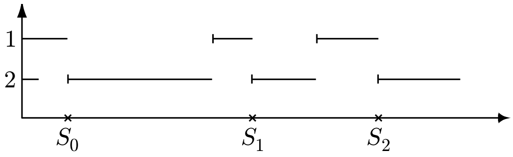

Figure 1.2

Fig. 1.2 illustrates the case of a two–state Markov jump process with $i = 2$. If the exponential holding time of state k has rate $\lambda(k)$, then F is the convolution of two exponential distributions with rates $\lambda(1)$ and $\lambda(2)$, respectively. □

Though the main probabilistic object describing a renewal process is nothing but a sum of nonnegative i.i.d. random variables, the point of view of renewal theory is somewhat different from the one usually taken when studying such sums. In particular, rather than in the behaviour of S_n as a function of n we are interested in the number of steps needed to reach size

t, i.e.

$$N_t \;=\; \#\{n=0,1,2,\ldots : S_n \le t\} \;=\; \inf\{n : S_n > t\}$$

the number of renewals up to time t. Note that

$$N_t \le n \iff S_n > t, \tag{1.1}$$

$$S_{N_t-1} \le t < S_{N_t}, \tag{1.2}$$

$$\{N_t = n\} \;=\; \{S_{n-1} \le t < S_n\} \tag{1.3}$$

These equations state, loosely speaking, that $t \to N_t$ is the inverse function of $n \to S_n$, and suggest that classical results on $\{S_n\}_{n\in\mathbb{N}}$ could be converted to results on $\{N_t\}_{t\ge 0}$. For example:

Proposition 1.4 *Let* $\mu = \mathbb{E}Y_1 = \int_0^\infty x\,F(\mathrm{d}x)$ *be the mean of the interarrival distribution. Then* (irrespective of the distribution of Y_0 or whether $\mu < \infty$ or $\mu = \infty$)

$$N_t/t \overset{\text{a.s.}}{\to} \mu^{-1}, \quad t \to \infty, \tag{1.4}$$

$$\mathbb{E}N_t/t \;\to\; \mu^{-1}, \quad t \to \infty, \tag{1.5}$$

Proof. Since $S_n/n \overset{\text{a.s.}}{\to} \mu$ and $N_t \overset{\text{a.s.}}{\to} \infty$, it follows by dividing (1.2) by N_t that $t/N_t \overset{\text{a.s.}}{\to} \mu$, i.e. $N_t/t \overset{\text{a.s.}}{\to} \mu^{-1}$. From this also the $\underline{\lim} \ge$ part of (1.5) follows by Fatou's lemma. To get $\overline{\lim} \le$, consider a renewal process $\{\widetilde{S}_n\}_{n\in\mathbb{N}}$ where the interarrival times are $\widetilde{Y}_0 = Y_0$, $\widetilde{Y}_n = Y_n \wedge a$, and let $\widetilde{N}_t, \widetilde{\mu}$, etc. be defined in the obvious way. Now by (1.3), $\widetilde{N}_t$ is a stopping time w.r.t. $\widetilde{Y}_1, \widetilde{Y}_2, \ldots$ for any fixed value of Y_0. Hence we may apply Wald's identity conditionally upon Y_0 and get

$$\mathbb{E}\bigl(\widetilde{Y}_1+\cdots+\widetilde{Y}_{\widetilde{N}_t}\bigr) \;=\; \mathbb{E}\bigl[\mathbb{E}\bigl(\widetilde{Y}_1+\cdots+\widetilde{Y}_{\widetilde{N}_t}\,\big|\,Y_0\bigr)\bigr] \;=\; \mathbb{E}\bigl[\widetilde{\mu}\mathbb{E}(\widetilde{N}_t\,|\,Y_0)\bigr] \;=\; \widetilde{\mu}\mathbb{E}\widetilde{N}_t.$$

Clearly, $\widetilde{N}_t \ge N_t$ and by (1.2), $\tilde{S}_{\widetilde{N}_t} = \widetilde{S}_{\widetilde{N}_t-1} + \widetilde{Y}_{\widetilde{N}_t} \le t + a$. Thus

$$\begin{aligned}\overline{\lim_{t\to\infty}}\,\mathbb{E}N_t/t \;&\le\; \overline{\lim_{t\to\infty}}\,\mathbb{E}\widetilde{N}_t/t \;=\; \overline{\lim_{t\to\infty}}\,\widetilde{\mu}^{-1}\mathbb{E}(\widetilde{Y}_1+\cdots+\widetilde{Y}_{\widetilde{N}_t})/t\\ &\le\; \overline{\lim_{t\to\infty}}\,\widetilde{\mu}^{-1}\mathbb{E}\widetilde{S}_{\widetilde{N}_t}/t \;\le\; \widetilde{\mu}^{-1}.\end{aligned}$$

Now $\tilde{\mu} \uparrow \mu$ as $a \uparrow \infty$ and (1.5) follows. □

A CLT analogue is given in Section 6.

One of the main points of renewal theory turns out to be obtaining refinements of (1.5). For this reason, (1.5) is sometimes called the *elementary renewal theorem.*

In the same way as in I.2, we shall now define the *backward recurrence time process* $\{A_t\}_{t\ge 0}$ and the *forward recurrence time process* $\{B_t\}_{t\ge 0}$ associated with the renewal process. In the language of Example 1.1, A_t is the *age* of the current item and B_t its *residual* or *excess lifetime.* That is,

A_t is the time elapsed since the last renewal and B_t the waiting time until the next renewal epoch $> t$,

$$A_t \;=\; t - S_{N_t - 1}, \quad B_t \;=\; S_{N_t} - t$$

(note that $B_0 = Y_0$ on $\{Y_0 > 0\}$ and $B_0 = Y_1$ on $\{Y_0 = 0\}$). The paths have the form illustrated in Fig. 1.3 and are right–continuous by definition.

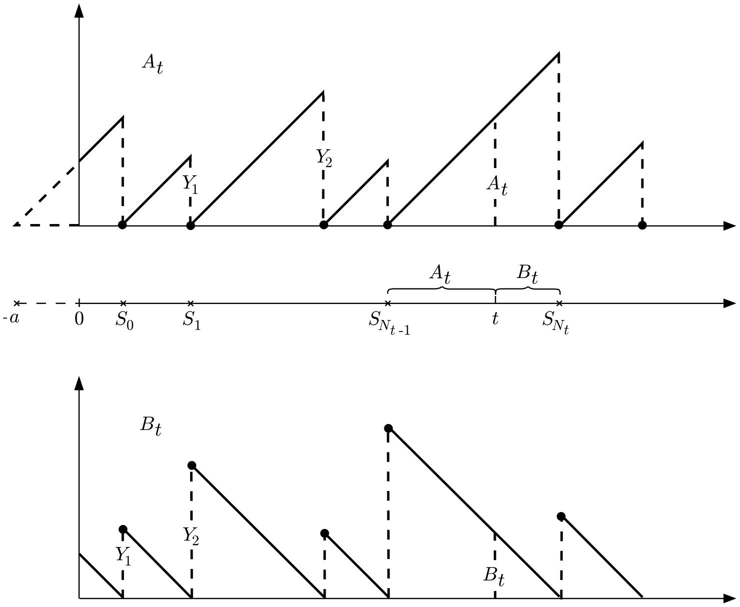

Figure 1.3

For a given renewal process, it is only possible to attach sense to A_t if $t \geq Y_0$. However, for any $a \geq 0$ with $\overline{F}(a) = 1 - F(a) > 0$ we can define a renewal process by "starting with a renewal at $-a$," i.e. letting Y_0 have the conditional distribution of Y_1 given $Y_1 > a$,

$$\mathbb{P}(Y_0 > y) \;=\; \mathbb{P}(Y_1 > y + a \,|\, Y_1 > a) \;=\; \frac{\overline{F}(a+y)}{\overline{F}(a)}. \tag{1.6}$$

Letting $A_t = a + t$, $t < Y_0$, we then get a version of $\{A_t\}$ that is defined for all $t \geq 0$ and has $A_0 = a$. In fact, we shall show:

Proposition 1.5 *The processes $\{A_t\}_{t\geq 0}$ and $\{B_t\}_{t\geq 0}$ are time–homogeneous strong Markov processes.*

Proof. The Markov property is intuitively obvious from the construction in the same way as in discrete time in I.2: B_t decreases linearly (and deterministically) until 0 is hit, then jumps to Y_1, decreases linearly to 0, jumps to Y_2 and so on, and this motion clearly has the asserted properties. The motion of A_t is also linear (but the jump times are not predictable). Given $\{A_s\}_{0\leq s\leq t}$, the evolution of the process after time t depends on the past only through the distribution G of the waiting time until the next jump. But the tail $\overline{G}(y)$ is given by (1.6) with $a = A_t$, which implies the Markov property.

For the strong Markov property for $\{B_t\}$, let f be continuous and bounded and let $g(b) = \mathbb{E}_b f(B_s)$. An inspection of the paths immediately shows that $g(b-t) = \mathbb{E}_b f(B_{t+s})$, $0 < t < b$. As $t \downarrow 0$, we have $f(B_{t+s}) \to f(B_s)$ and hence by dominated convergence $g(b-t) \to g(b)$, i.e. g is left–continuous. For $u \downarrow t$ we have $B_u \uparrow B_t$ so that $g(B_u) \to g(B_t)$. Thus $\{g(B_t)\}$ has right–continuous paths, and the strong Markov property for $\{B_t\}$ follows by I.8.3. For $\{A_t\}$, let $h(a) = \mathbb{E}_a f(A_s)$. Using first the right–continuity of $\{A_t\}$, we get for $t \downarrow 0$ that

$$\begin{aligned} h(a) &= \mathbb{E}_a f(A_{s+t}) + \mathrm{o}(1) \\ &= \frac{1}{\overline{F}(a)} \left\{ \overline{F}(a+t)\mathbb{E}_{a+t} f(A_s) + \int_{a+}^{a+t} \mathbb{E}_0 f(A_{s+t-y}) F(\mathrm{d}y) \right\} + \mathrm{o}(1) \\ &= \mathbb{E}_{a+t} f(A_s) + \mathrm{o}(1) \;=\; h(a+t) + \mathrm{o}(1). \end{aligned}$$

Therefore $h(a)$ is right–continuous, which in view of $A_u \downarrow A_t$, $u \downarrow t$, implies the paths of $\{h(A_t)\}$ to be so. □

We note that a number of Markov processes associated with queues and related models (see e.g. Problems 1.3, X.3.2, XIV.1.1) have paths of a similar shape as $\{A_t\}$, $\{B_t\}$ and that the strong Markov property in such cases follows by small variants of the proof of Proposition 1.5

It was remarked in Example 1.3 that the recurrence times of a point i in a Markov process $\{X_t\}$ form a renewal process. Proposition 1.5 shows that *any* renewal process is of this type (with $X_t = A_t$ and $i = 0$).

Problems

1.1 (THE TYPE I COUNTER) The incoming particles constitute a Poisson process, but the registrations do not, since for technical reasons the counter cannot register the second of two particles emitted at almost the same time. Suppose that each *registered* particle locks the counter for a time with distribution G, that particles arriving in a locked period have no effect and that locking times of different particles are independent, both mutually and of the Poisson process. Show that the registrations constitute a renewal process and find the interarrival distribution.

1.2 (THE PEDESTRIAN DELAY PROBLEM) At time 0, a pedestrian arrives at a road and wants to cross. Crossing is possible when the gap beween two cars is at least ξ. Find the distribution of the waiting time until crossing is performed.

1.3 Show that $\{(A_t, B_t)\}_{t \geq 0}$ has the strong Markov property. [*Hint:* Consider $\mathbb{E}_{a+t,b-t} f(A_s) g(B_s)$.]

Notes A classical reference for renewal theory is Chapter XI of Feller (1971).

In view of the basic importance of renewal theory, it is not surprising that several generalizations have been considered, in particular renewal theory for

Markov chains (VII.4 and references therein), for more general dependent sequences (Berbee, 1979; Lalley, 1986; Alsmeyer, 1994), for random walks (surveyed in Gut, 1988), nonlinear renewal theory (Woodroofe, 1982; Siegmund, 1985; Zhang, 1988), multivariate renewal theory, (Carlsson and Wainger, 1984, and references therein), and finally renewal theory with infinite mean where a recent paper with reference to older literature is Doney (1997).

2 Renewal Equations and the Renewal Measure

The *renewal equation* is the convolution equation $Z = z + F * Z$ (for the convolution notation, see the Notes at the end of this section), i.e.

$$Z(t) \;=\; z(t) + \int_0^t Z(t-u)\,F(\mathrm{d}u), \quad t \geq 0. \tag{2.1}$$

Here one thinks of Z as an unknown function on $[0,\infty)$, z as a known function on $[0,\infty)$ and F as a known nonnegative (Radon) measure on $[0,\infty)$. It is often assumed that F is a probability, i.e. $\|F\| = 1$, in which case (2.1) is *proper.* If $\|F\| < 1$, the renewal equation (2.1) is *defective*, but we shall also consider the *excessive* case $\|F\| > 1$. We always assume that $F(0) = 0$. We shall first give some examples.

Example 2.1 Consider a pure renewal process with interarrival distribution F and the recurrence times A_t, B_t defined as in Section 1. Let $\xi \geq 0$ be fixed and define $Z_A(t) = \mathbb{P}(A_t \leq \xi)$, $Z_B(t) = \mathbb{P}(B_t \leq \xi)$. Then Z_A, Z_B satisfy the renewal equations

$$Z_A = z_A + F * Z_A, z_A(t) \;=\; \mathbb{P}(A_t \leq \xi, Y_1 > t) \;=\; I(t \leq \xi)\overline{F}(t), \quad (2.2)$$
$$Z_B = z_B + F * Z_B, z_B(t) \;=\; \mathbb{P}(B_t \leq \xi, Y_1 > t) \;=\; F(t+\xi) - F(t)(2.3)$$

The proof of this is carried out by the *renewal argument*, i.e. (i) conditioning on the value u of Y_1, which yields

$$Z_A(t) \;=\; \mathbb{P}(A_t \leq \xi, Y_1 > t) \;+\; \int_0^t \mathbb{P}\big(A_t \leq \xi \,\big|\, Y_1 = u\big)\,F(\mathrm{d}u), \tag{2.4}$$

and (ii) remarking that the process starts from scratch at time Y_1, which yields $\mathbb{P}(A_t \leq \xi \,|\, Y_1 = u) = \mathbb{P}(A_{t-u} \leq \xi)$ for $u \leq t$. Thus, since $A_t = t$ on $\{Y_1 > t\}$, (2.2) and (2.4) are the same equation. Equation (2.3) is derived in a similar manner using $B_t = Y_1 - t$ on $\{Y_1 > t\}$. □

Example 2.2 (LOTKA'S INTEGRAL EQUATION) This comes from classical deterministic or semi–deterministic population theory associated with the names of Sharpe and Lotka. Consider the female part of a population, where women aged a give birth (to a single daughter) at rate $\lambda(a)$ and survive to age $a + t$ in a proportion of ${}_tp_a$ (in traditional demographic notation). We are interested in $Z(t)$, the overall birthrate at time t, which can be split into the rates $z(t)$, $Z_0(t)$ of births from women born before, resp. after, time $t = 0$. To determine $z(t)$, we must know the age structure of the population at time zero, which will be represented by its density $f_0(a)$ (thus $\int_0^\infty f_0(a)\,da$ is the initial population size, not necessarily $= 1$). Then women aged a at $t = 0$ have density $f_0(a)_t p_a$ at time t and are aged $a + t$, and hence

$$z(t) \;=\; \int_0^\infty f_0(a)_t p_a \lambda(a+t)\,\mathrm{d}a$$

Similarly, women born at time $t-s$ provide a contribution $Z(t-s)_s p_0 \lambda(s)\,\mathrm{d}s$ to $Z_0(t)$ so that

$$Z(t) \;=\; z(t) + \int_0^t Z(t-s)_s p_0 \lambda(s)\,\mathrm{d}s \qquad (2.5)$$

and we have a renewal equation with $F(\mathrm{d}s) \;=\; {}_s p_0 \lambda(s)\,\mathrm{d}s$. Note that $\|F\| = \int_0^\infty {}_s p_0 \lambda(s)\,\mathrm{d}s$ is the average number of daughters born to a woman, the so–called *net reproduction rate*. This could have values both < 1, $= 1$ and > 1, but the typical case is that of a growing population with $\|F\| > 1$, where (2.5) is thus excessive. Note also that other quantities of interest, such as the density

$$\begin{cases} f_0(a-t)_t p_{a-t} & t \le a \\ Z(t-a)_a p_0 & t > a \end{cases}$$

of women aged a at time t and the total population size

$$N(t) \;=\; \int_0^t Z(t-a)_a p_0\,\mathrm{d}a \;+\; \int_0^\infty f_0(a)_{t-a} p_a\,\mathrm{d}a \qquad (2.6)$$

are readily expressed in terms of Z. □

Example 2.3 (THE RUIN PROBLEM OF INSURANCE MATHEMATICS) Assume that the claims incurred by an insurance company arrive according to a Poisson process $\{N_t\}$ with intensity λ, that the sizes of the claims are i.i.d. nonnegative random variables $X_1, X_2, \ldots$, with common distribution say G, and that the inflow of premium up to time t is ct. Thus the risk reserve at time t is

$$U_t \;=\; u + ct - \sum_{n=1}^{N_t} X_n,$$

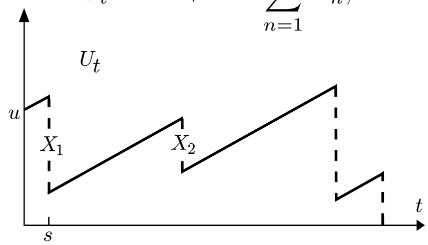

Figure 2.1

with $u = U_0$ the initial value (cf. Fig. 2.1). We are interested in the probabilities

$$Z(u) \;=\; \mathbb{P}\bigl(U_t \geq 0 \;\forall t \,\big|\, U_0 = u\bigr), \quad 1 - Z(u) \;=\; \mathbb{P}\Bigl(\inf_{0\leq t<\infty} U_t < 0 \,\big|\, U_0 = u\Bigr),$$

of ultimate survival and ultimate ruin of the company, say for the purpose of assessing whether c has been chosen sufficiently large compared to u, and we will see that

$$Z(t) \;=\; Z(0) + \frac{\lambda}{c}\int_0^t Z(t-x)\overline{G}(x)\,\mathrm{d}x. \tag{2.7}$$

This is of the form (2.1), with $F(\mathrm{d}x) = (\lambda/c)\overline{G}(x)\,\mathrm{d}x$. We note that $\|F\| = \lambda\nu/c$, where

$$\nu \;=\; \int_0^\infty \overline{G}(x)\,\mathrm{d}x \;=\; \int_0^\infty y\,G(\mathrm{d}y)$$

is the mean claim size, and that $\lambda\nu$ is the mean size of the claims received per unit time, c the inflow of premium per unit time; in practical situations, the company will typically have chosen $c > \lambda\nu$ such that $\|F\| < 1$.

The shortest proof of (2.7) exploits one of the most basic explicit formulas in Wiener–Hopf theory, to be proved in VIII.5.7: if $\sigma = \inf\{t > 0 : U_t < u = U_0\}$, then $u - U_\sigma$ has the (defective) density $(\lambda/c)\overline{G}(x)$ when $\lambda\nu/c < 1$. From this, (2.7) follows immediately by conditioning upon $x = u - U_\sigma$ and noting that $Z(0) = \mathbb{P}(\sigma = \infty)$.

A longer and more naive (but classical!) argument uses instead conditioning upon the time s of the first claim where the process renews itself, holding the new initial fortune $u + cs - X_1$ (terminates if $X_1 > u + cs$). Therefore

$$Z(u) \;=\; \int_0^\infty \lambda \mathrm{e}^{-\lambda s}\,\mathrm{d}s \int_0^{u+cs} Z(u+cs-x)\,G(\mathrm{d}x).$$

Letting $t = u + cs$, we get

$$Z(u)\mathrm{e}^{-\lambda u/c} \;=\; \frac{\lambda}{c}\int_u^\infty \mathrm{e}^{-\lambda t/c}\,\mathrm{d}t \int_0^t Z(t-x)\,G(\mathrm{d}x).$$

This representation shows that Z is differentiable, and differentiating w.r.t. u yields

$$\begin{aligned}
\mathrm{e}^{-\lambda u/c}\Bigl(Z'(u) - \frac{\lambda}{c}Z(u)\Bigr) \;&=\; -\frac{\lambda}{c}\mathrm{e}^{-\lambda u/c}\int_0^u Z(u-x)\,G(\mathrm{d}x),\\
Z'(u) \;&=\; \frac{\lambda}{c}Z(u) - \frac{\lambda}{c}\int_0^u Z(u-x)\,G(\mathrm{d}x).
\end{aligned}$$

Integrating w.r.t. $\mathrm{d}u$ from 0 to t and letting

$$h(y) \;=\; \int_0^{t-y} Z(u)\,\mathrm{d}u, \;\; 0 \leq y \leq t, \quad h(y) \;=\; 0, \;\; y > t,$$

yields

$$\begin{aligned}
Z(t) - Z(0) - \frac{\lambda}{c}h(0) \;&=\; -\frac{\lambda}{c}\int_0^t \mathrm{d}u \int_0^u Z(u-x)\,G(\mathrm{d}x) \\
=\; -\frac{\lambda}{c}\int_0^t h(x)\,G(\mathrm{d}x) \;&=\; -\frac{\lambda}{c}\int_0^\infty h(x)\,G(\mathrm{d}x) \\
=\; -\frac{\lambda}{c}h(0) - \frac{\lambda}{c}\int_0^\infty h'(x)\overline{G}(x)\,\mathrm{d}x,
\end{aligned}$$

which is the same as (2.7). □

We shall now study questions of existence and uniqueness of solutions. Asymptotic estimates will be derived in Sections 4 and 5 for the case $\|F\| = 1$ and in Section 7 for $\|F\| \neq 1$.

Given F, we define *the renewal measure* by $U(\mathrm{d}x) = \sum_0^\infty F^{*n}(\mathrm{d}x)$ and *the renewal function* U by $U(t) = \sum_0^\infty F^{*n}(t)$ (see again the Notes at the end of this section for notation).

Theorem 2.4 (i) *The renewal function $U(t)$ is finite for all $t < \infty$;*
(ii) *if the function z in the renewal equation* (2.1) *is bounded on finite intervals* (i.e. $\sup_{0\le t\le T}|z(t)| < \infty$ for all $T < \infty$), *then $Z = U * z$* (i.e. $Z(t) = \int_0^t z(t-x)\,U(\mathrm{d}x)$) *is well defined, a solution to* (2.1) *and the unique solution to* (2.1) *which is bounded on finite intervals;*
(iii) *if $\|F\| = 1$ then $U(t)$ is the expected number $\mathbb{E}N_t$ of renewals up to time t in a pure renewal process with interarrival distribution F. More generally, in any renewal process with interarrival distribution F, the expected number of renewals in $(t, t+a]$ is*

$$\mathbb{E}(N_{t+a} - N_t) \;=\; \int_0^a U(a-\xi)\,G_t(\mathrm{d}\xi) \;=\; G_t * U(a) \;=\; U * G_t(a) \quad (2.8)$$

where $G_t(\xi) = \mathbb{P}(B_t \le \xi)$. Further, the expression (2.8) *cannot exceed $U(a)$.*

Lemma 2.5 *If F is a measure on $(0, \infty)$ with $F(a) < \infty$ for all $a < \infty$, then for any $t < \infty$ and $\delta < 1$ there exists $C_{\delta,t} < \infty$ such that $F^{*n}(t) \le C_{\delta,t}\delta^n$ for all n.*

Proof. Since $F^{*n}(t)$ does not involve the restriction of F to (t, ∞), we may put $F(\mathrm{d}x) = 0$, $x > t$ if necessary to ensure that the Laplace transform $\widehat{F}[\beta] = \int_0^\infty \mathrm{e}^{-\beta x}F(\mathrm{d}x)$ is finite for all $\beta < \infty$ and that $\widehat{F}[\beta] \to 0$, $\beta \to \infty$. Choose β with $\widehat{F}[\beta] < \delta$ and note that

$$F^{*n}(t) \;\le\; \mathrm{e}^{\beta t}\int_0^t \mathrm{e}^{-\beta s}F^{*n}(\mathrm{d}s) \;\le\; \mathrm{e}^{\beta t}\widehat{F}[\beta]^n. \qquad \square$$

Proof of Theorem 2.4. (i) follows immediately by Lemma 2.5 (alternatively, if $\|F\| = 1$, use (iii) and (1.5)). For (ii), it is now obvious that $Z = U * z$ is well defined and bounded on finite intervals. Defining $U_N = \sum_0^N F^{*n}$,

$Z_N = U_N * z$, we have $Z_{N+1} = z + F * Z_N$ and that $Z = \lim Z_N$ is a solution follows as $N \to \infty$. Given two solutions of the type considered, their difference V satisfies $V = F * V = \cdots = F^{*n} * V$ so that

$$|V(t)| \;=\; \left|\int_0^t V(t-x)\, F^{*n}(\mathrm{d}x)\right| \;\le\; \sup_{0\le y\le t} |V(y)| \cdot F^{*n}(t),$$

and $V(t) = 0$ follows as $n \to \infty$. For (iii), it follows from (1.1) that in a pure renewal process with interarrival distribution F

$$\begin{aligned}\mathbb{E}N_t \;&=\; \sum_{n=0}^{\infty} \mathbb{P}(N_t > n) \;=\; \sum_{n=0}^{\infty} \mathbb{P}(S_n \le t)\\ &=\; \sum_{n=0}^{\infty} \mathbb{P}(Y_1 + \cdots + Y_n \le t) \;=\; \sum_{n=0}^{\infty} F^{*n}(t) \;=\; U(t),\end{aligned}$$

and the more general (2.8) then follows by conditioning on $B_t = \xi$ and noting that a pure renewal process starts at time $t + \xi$. Finally, an upper bound for (2.8) is obtained by replacing G_t with the distribution degenerate at zero and this yields $U(a)$. □

Example 2.6 In many examples, the form $Z = U * z$ of the solution to $Z = z + F * Z$ can be seen directly and sometimes this is even the most natural approach. Consider as an example a *shot–noise process*

$$W_t \;=\; \sum_{n=0}^{N_t - 1} f(t - S_n, X_n)$$

where $X_0, X_1, \ldots$ are i.i.d. and independent of the renewal process. Then $Z(t) = \mathbb{E}W_t$ satisfies

$$\begin{aligned}Z(t) \;&=\; \sum_{n=0}^{\infty} \mathbb{E}\big[f(t - S_n, X_n);\; S_n \le t\big]\\ &=\; \sum_{n=0}^{\infty} \int_0^t z(t-u)\, F^{*n}(\mathrm{d}u) \;=\; U * z(t),\end{aligned}$$

where $z(t) = \mathbb{E}f(t, X_1)$.

The shot–noise process is used to describe certain electrical tubes, where primary impulses of sizes $X_0, X_1, \ldots$ are emitted at the epochs of a renewal process. An impulse of size x then creates secondary effects which are of size $f(t, x)$ after time t. Similar phenomena occur in road traffic noise, where the renewal process describes the passing of the cars, X_n is the noise level of the nth car and $f(t, x)$ is the actual noise at a distance of t. □

If U is absolute continuous on $(0, \infty)$, we call the density $u(x) = \mathrm{d}U/\mathrm{d}x$ the *renewal density*.

Proposition 2.7 *The renewal density u exists if and only if F has a density f. Then $u = \sum_1^\infty f^{*n}$ or, equivalently, u is the solution of the renewal equation $u = f + F * u$.*

Proof. It is clear that if f exists, then so does u and is given by $u = \sum_1^\infty f^{*n}$; that this is the same as the solution $U * f$ of $u = f + F * u$ follows from $u(0) = f(0) = 0$ (equivalently, $u = f + F * u$ follows at once from the renewal argument). □

Example 2.8 The main examples where the renewal function, the renewal density, or the distribution of A_t, B_t, etc. can be found explicitly are the phase–type ones given in III.5. For an additional one, assume that F is uniform on $(0, a)$. Then the renewal density $u(x)$ exists and we will show that it is given by

$$u(x) \;=\; \frac{1}{a}\mathrm{e}^{x/a}\sum_{k=0}^{\lfloor x/a\rfloor}\mathrm{e}^{-k}\frac{(k-x/a)^k}{k!}. \tag{2.9}$$

To this end, let $v(x)$ denote the r.h.s. of (2.9), and assume w.l.o.g. that $a = 1$. Then the renewal equation $u = f + F * u$ for u means

$$u(x) \;=\; I(x<1) + \int_0^x u(x-y)I(y<1)\,\mathrm{d}y \;=\; I(x<1) + \int_{(x-1)^+}^x u(y)\,\mathrm{d}y$$

which implies

$$u'(x) \;=\; \begin{cases} u(x) & x<1 \\ u(x)-u(x-1) & x>1 \end{cases}.$$

In particular, from $u(0+) = f(0+) = 1$ we get $u(x) = \mathrm{e}^x = v(x)$, $x \le 1$. Thus the relation $u(x) = v(x)$, $n < x \le n+1$, holds for $n = 0$. Assuming it shown for $n-1$, we get

$$\begin{aligned} v'(x) \;&=\; v(x) \;-\; \mathrm{e}^x\sum_{k=1}^{\lfloor x\rfloor}\mathrm{e}^{-k}\frac{(k-x)^{k-1}}{(k-1)!} \\ &=\; v(x) \;-\; \mathrm{e}^{x-1}\sum_{\ell=0}^{\lfloor x-1\rfloor}\mathrm{e}^{-\ell}\frac{(\ell-x+1)^\ell}{\ell!} \\ &=\; v(x)-v(x-1) \;=\; v(x)-u(x-1), \quad n<x\le n+1, \end{aligned}$$

which together with $v(n) = u(n)$ and $u'(x) = u(x) - u(x-1)$ implies $u(x) = v(x)$, $n < x \le n+1$. Thus $u(x) = v(x)$ for all x. □

It is clear by the same argument as in the proof of Theorem 2.4 that if the renewal process is terminating, $\|F\| < 1$, then $U(t)$ can still be interpreted as the expected number of renewals in $[0, t]$. In particular, the expected number of renewals within finite time is

$$\|U\| \;=\; \lim_{t\to\infty} U(t) \;=\; \sum_{n=0}^\infty \|F\|^n \;=\; \frac{1}{1-\|F\|}, \tag{2.10}$$

cf. also I.2.4.

However, the renewal measure has a different important interpretation in the terminating case. Define the *lifetime* or *maximum* by $M = \sup\{S_n : S_n < \infty\}$. Then:

Proposition 2.9 *In the zero–delayed case with* $\|F\| < 1$, *the distribution of* M *is* $(1-\|F\|)U$.

Proof. We give two arguments, (a) and (b). In (a), put $\sigma = \inf\bigl\{n \geq 0 : Y_{n+1} = \infty\bigr\}$. Then

$$M \;=\; Y_1 + \cdots + Y_\sigma, \quad \mathbb{P}(\sigma = n) \;=\; \bigl(1-\|F\|\bigr)\|F\|^n,$$

and conditionally upon σ, we have $\mathbb{P}(Y_k \leq y \,|\, \sigma) = G(y)$ for $k \leq \sigma$, where $G = F/\|F\|$. Hence

$$\begin{aligned}
\mathbb{P}(M \leq y) \;&=\; \sum_{n=0}^{\infty} \mathbb{P}(\sigma = n)\mathbb{P}\bigl(Y_1 + \cdots + Y_n \leq y \,\big|\, \sigma = n\bigr) \\
&=\; \sum_{n=0}^{\infty} \bigl(1-\|F\|\bigr)\|F\|^n G^{*n}(y) \;=\; \bigl(1-\|F\|\bigr)\sum_{n=0}^{\infty} F^{*n}(y) \\
&=\; \bigl(1-\|F\|\bigr)U(y).
\end{aligned}$$

In (b), let $Z(x) = \mathbb{P}(M \leq x)$. Now if $Y_1 = \infty$, then $M = 0$ and hence $M \leq x$, whereas if $0 < Y_1 = y < \infty$, then for $\{M \leq x\}$ to occur we must have $y \leq x$ and that the lifetime of the renewal process starting at y is at most $x - y$. Hence

$$Z(x) \;=\; 1 - \|F\| + \int_0^x Z(x-y)\,F(\mathrm{d}y),$$

which implies $Z(x) \;=\; U * \bigl(1-\|F\|\bigr)(x) \;=\; \bigl(1-\|F\|\bigr)U(x)$. □

Problems

2.1 Find a renewal equation for the joint distribution of the recurrence times (A_t, B_t).

2.2 (THE TYPE II COUNTER) As in Problem 1.1, we assume that the particles arrive at the counter according to a Poisson process with intensity λ, but use a different model for the locking mechanism, namely that locking times of different particles are i.i.d. with common distribution G and that each particle arriving at the counter cancels the aftereffects (if any) of its predecessors. Show that the probability $Z(t)$ of the duration of the locked period to exceed t satisfies the renewal equation

$$Z(t) \;=\; \overline{G}(t)\mathrm{e}^{-\lambda t} + \int_0^t Z(t-x)\overline{G}(x)\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}x.$$

2.3 Show that $Z(t) = Z(0)U(t) = 0$ in the ruin problem with $c \leq \lambda\nu$.

2.4 Show that if the Laplace transform $\widehat{F}$ is well defined (e.g. if $\|F\| < \infty$) then $\widehat{U} = (1-\widehat{F})^{-1}$ and that in (2.2) $\widehat{Z} = \widehat{z}/(1-\widehat{F})$.

Notes Notationally, F^{*n} denotes the nth convolution power of F, i.e. the probability distribution degenerate at zero for $n = 0$, while for $n = 0, 1, 2, \ldots$

$$F^{*(n+1)}(t) \;=\; \int_0^t F^{*n}(t-u)\,F(\mathrm{d}u) \;=\; \int_0^t F(t-u)\,F^{*n}(\mathrm{d}u).$$

In particular, $\|F^{*n}\| = \|F\|^n$. Further, we have used the convention that the convolution $F*Z$ of a function Z and a measure F is a function. When identifying the measure $F(\mathrm{d}x)$ with the function $F(t) = \int_0^t F(\mathrm{d}x)$, this is consistent with the usual convolution of measures; indeed, if Z itself corresponds to $Z(\mathrm{d}x)$, then $F*Z(t) = \int_0^t F*Z(\mathrm{d}x)$. The proof of this is elementary as well as that of formulas like $F*G = G*F$, $F*(G*Z) = (F*G)*Z$ used without further notice in the text (here G is another measure).

More material on the demographic model in Example 2.2 is in Pollard (1973) and Preston *et al.* (2001). We return to ruin probabilities in XIV.5–6.

3 Stationary Renewal Processes

The definition that a renewal process (or a general point process on $[0,\infty)$) be stationary is the obvious one: if for any $t > 0$ we shift the origin to t, the distributions of the epochs should be left unchanged; formally, $\{N_{s+t} - N_t\}_{s\geq 0} \overset{\mathscr{D}}{=} \{N_s\}_{s\geq 0}$. Clearly, this will hold if the distribution of the forward recurrence time B_t does not depend on t, i.e. if the Markov process $\{B_t\}$ is stationary. Conversely, this is also necessary since otherwise the first epoch of $\{N_{s+t} - N_t\}$ has a distribution depending on t.

The form of the stationary distribution F_0 for $\{B_t\}$ is easily guessed by a level crossing argument. Namely, in a stationary situation the average number of upcrossings of level $x > 0$ should be the same as the average number of downcrossings, which in turn leads to the rate of upcrossings being equal to the rate of downcrossings. Assume that F_0 exists and has a density f_0. An upcrossing of a stationary version of $\{B_t\}$ in $[0,h]$ occurs if $B_0 \in (0,h]$ and the jump out of 0 at time B_0 exceeds x so that the rate is $f_0(0)\overline{F}(x)h + \mathrm{o}(h)$. For $x > h$, a downcrossing occurs precisely when $B_0 \in (x, x+h]$ so that the rate is $f_0(x)h + \mathrm{o}(h)$. This shows that $f_0(x) = f_0(0)\overline{F}(x)$ and from $1 = \int f_0 = f_0(0)\mu$, we then get $f_0(x) = \overline{F}(x)/\mu$.

A similar argument for $\{A_t\}$ is in principle possible but more complicated (see Problem 3.1), and it is easier to note that

$$\{B_t \leq \xi\} \;=\; \bigl\{\text{renewal in } (t, t+\xi]\bigr\} \;=\; \{A_{t+\xi} < \xi\} \tag{3.1}$$

so that (since F_0 was found to have a density so that $\mathbb{P}(B_t = \xi) = 0$) the stationary distributions for $\{A_t\}$, $\{B_t\}$ must be the same.

The formal justification for the level crossing argument is provided by the rate conservation law considered in VII.6, but in the rest of this section, we will give a rigorous direct treatment of the above discussion as well as look into some further topics. We let F_0 denote the measure with density

$\overline{F}(x)/\mu$, assuming for the rest of the section that $\mu < \infty$ since otherwise there is no hope for stationarity properties.

Considering $\{A_t\}$, $\{B_t\}$ jointly, we have

$$\mathbb{P}\big(B_t > y \,\big|\, A_t = x\big) \;=\; \mathbb{P}\big(Y_1 > x + y \,\big|\, Y_1 > x\big) \;=\; \frac{\overline{F}(x+y)}{\overline{F}(x)}, \tag{3.2}$$

which leads to:

Lemma 3.1 (i) *If $\{A_t\}$ is stationary, then so is $\{B_t\}$.* (ii) *If A_t has distribution F_0, then so has B_t, and $\mathbb{P}(A_t > x, B_t > y) = \overline{F}_0(x+y)$.*

Proof. Part (i) follows immediately from (3.2) which shows that the distribution of B_t is a function of that of A_t. If indeed as in (ii) A_t has distribution F_0, (3.2) yields

$$\begin{aligned}\mathbb{P}(A_t > x, B_t > y) \;&=\; \int_x^\infty \frac{\overline{F}(z+y)}{\overline{F}(z)} F_0(\mathrm{d}z) \;=\; \frac{1}{\mu}\int_x^\infty \overline{F}(z+y)\,\mathrm{d}z \\ &=\; \frac{1}{\mu}\int_{x+y}^\infty \overline{F}(z)\,\mathrm{d}z \;=\; \overline{F}_0(x+y);\end{aligned}$$

that the distribution of B_t is F_0 then follows by taking $x = 0$. □

The joint distribution in (ii) can be described in a simple intuitive way. Let $C_t = A_t + B_t$ denote (in the terminology of Example 1.1) the *current lifetime* of the item at time t.

Lemma 3.2 *Let F_1 be the distribution with density x/μ w.r.t. F, let C be a r.v. with distribution F_1, let U be independent of C and uniform on* $(0,1)$, *and define $A = CU$, $B = C(1-U)$. Then $\mathbb{P}(A > x, B > y) = \overline{F}_0(x+y)$.*

Proof. Since $A > x, B > y$ is equivalent to $C > x + y$, $U \in \big(x/C, 1 - y/C\big)$, we get

$$\begin{aligned}\mathbb{P}(A > x, B > y) \;&=\; \int_{x+y}^\infty \Big[1 - \frac{x+y}{c}\Big]\frac{c}{\mu}F(\mathrm{d}c) \\ =\; \frac{1}{\mu}\int_0^\infty [c - x - y]^+ F(\mathrm{d}c) \;&=\; \frac{1}{\mu}\int_{x+y}^\infty \overline{F}(z)\,\mathrm{d}z \;=\; \overline{F}_0(x+y),\end{aligned}$$

using integration by parts in the third step. □

Theorem 3.3 *Let C be a r.v. with distribution F_1 and let U be independent of C and uniform on* $(0,1)$. *Then the version of the Markov process $\{(A_t, B_t, C_t)\}$ obtained from the initial values $A_0 = CU$, $B_0 = C(1-U)$, $C_0 = C$ is strictly stationary. Further, the point process whose set of epochs is*

$$\{t \geq 0 : A_{t-} \neq A_t\} \;=\; \{t \geq 0 : B_{t-} \neq B_t\}$$

is a stationary renewal process with interarrival distribution F.

Lemma 3.4 *A Radon measure H on* $[0,\infty)$ *satisfies* $U * H(\mathrm{d}x) = \mathrm{d}x$ (Lebesgue measure) *if and only if H has density* $\overline{F}(x)$*.*

Proof. If H has density $h(x) = \overline{F}(x)$, then $U * H$ has density $U * h = U - U * F = F^{*0} = 1$, i.e. $U * H(\mathrm{d}x) = \mathrm{d}x$. If also $U * H_1(\mathrm{d}x) = \mathrm{d}x$, then $U * H(a) = a = U * H_1(a)$. Thus the solutions of $Z = H + F * Z$, $Z = H_1 + F * Z$ are the same, and this immediately implies $H = H_1$. □

Proof of Theorem 3.3. In view of Lemma 3.1(i), all that needs to be shown is that $\{A_t\}$ is stationary. By Lemma 3.2, A_0 has distribution F_0 and we must show that $\mathbb{P}(A_t \le \xi) = F_0(\xi)$ for all t, ξ. If $t \ge \xi$, $A_t \le \xi$ occurs precisely when there is a renewal in $[t-\xi, t]$. Since B_0 also has distribution F_0 by Lemma 3.1(i), the intensity measure (cf. A3) of renewals is $U * F_0(\mathrm{d}x)$ which by Lemma 3.4 is $\mathrm{d}x/\mu$. Thus conditioning upon the time x of the last renewal we get

$$\mathbb{P}(A_t \le \xi) \;=\; \int_{t-\xi}^{t} \overline{F}(t-x)\,\mathrm{d}x/\mu \;=\; \int_0^{\xi} \overline{F}(y)\,\mathrm{d}y/\mu \;=\; F_0(\xi).$$

If $t \le \xi$, $A_t > \xi$ occurs when the initial item has age a at least $\xi - t$ and survives to time t. Thus

$$\begin{aligned}\mathbb{P}(A_t > \xi) \;&=\; \int_{\xi-t}^{\infty} \frac{\overline{F}(a+t)}{\overline{F}(a)}\, F_0(\mathrm{d}a)\\ &=\; \int_{\xi-t}^{\infty} \overline{F}(a+t)\,\mathrm{d}a/\mu \;=\; \int_{\xi}^{\infty} \overline{F}(a)\,\mathrm{d}a/\mu \;=\; \overline{F}_0(\xi).\end{aligned}$$

□

The fact that the stationary distribution of C_t is F_1, not F, is known as the *inspection* or *waiting time paradox*, stating that the item at time t is not typical in the sense of having distribution F. The reason is loosely that sampling at a fixed time favours items with long lifetimes, and therefore one also speaks of *length–biasing.* The paradox is important not only for its own sake but also as a warning for intuition in many similar situations.

Having constructed a stationary renewal process, we finally consider the uniqueness question. In addition to the characterizations studied above, we will also consider the intensity measure, counting the expected number of renewals. The intensity mesure is stationary if it is translation invariant, i.e. equal to Lebesgue measure times a constant (necessarily μ^{-1} by the elementary renewal theorem).

Proposition 3.5 *Let G be a distribution on* $(0,\infty)$*, such that either* (i) *G is stationary for* $\{A_t\}$*,* (ii) *G is stationary for* $\{B_t\}$ (iii) *a renewal process with delay distribution G is stationary, or* (iv) *a renewal process with delay distribution G has stationary intensity measure. Then* $G = F_0$*.*

Proof. It is by now clear that (i)⇒(ii)⇒(iii)⇒(iv) so it suffices to show that (iv)⇒ $G = F_0$. But the intensity measure $\mathrm{d}x/\mu$ equals $U * G(\mathrm{d}x)$, so this follows at once from Lemma 3.4. □

Corollary 3.6 *A delayed renewal process is stationary if and only if the distribution of the initial delay B_0 is F_0.*

The proof is immediately clear from the above discussion. However, note that when replacing B_0 by A_0 one only gets "if" ("only if" fails say for the Poisson process).

Problems

3.1 Assume that F has a density f. Show by a level crossing argument that a stationary density f_0 for $\{A_t\}$ must satisfy $f_0'(x) = -f_0(x)f(x)/\overline{F}(x)$ and that therefore $f_0(x) = \overline{F}(x)/\mu$.

3.2 Evaluate F_0 if F is degenerate at 1.

3.3 Evaluate F_0 for the cases $F = E_k$ and $F = H_k$; cf. III.4.

3.4 Find the density of $F_0 * F^{*n}$.

3.5 Show that the current life distribution F_1 is stochastically larger than F, $\overline{F}_1(x) \geq \overline{F}(x)$.

Notes There has been some work on properties of the mapping sending F into F_1; see Brown (2004) for a recent contribution and references.

4 The Renewal Theorem in Its Equivalent Versions

The renewal theorem is one of the most fundamental results of probability theory, perhaps not so much because of its intrinsic interest but rather because of the applicability to, and strong implications for, a number of other areas. It has several versions, one analytical giving asymptotic estimates for the solutions of (proper) renewal equation, and various probabilistic ones which all in some way state that as $t \to \infty$, then a (possibly delayed) renewal process asymptotically behaves like a stationary one if $\mu < \infty$ and has a behaviour like null recurrence if $\mu = \infty$. In the present section, we state the various versions and prove their equivalence. The classical analytical proof of the renewal theorem is then in Section 5 and a more recent coupling proof in VII.2.

From now on it becomes necessary to distinguish between F being lattice (concentrated on a set of the form $\{\delta, 2\delta, \ldots\}$) or nonlattice. In the lattice case, one may rescale time so as to make F aperiodic on $\mathbb{N}$ and a number of aspects of renewal theory for that case have already been studied in I.2 (including the Problems). We shall therefore almost entirely concentrate on the nonlattice case and only state a few selected results for the lattice case.

Before being able to state all versions of the renewal theorem, we need a definition. Suppose for a while that z in the renewal equation $Z = z + F * Z$ is nonnegative, and for $h > 0$ define

$$\overline{z}_h(x) = \sup_{y \in I_n^h} z(y), \quad \underline{z}_h(x) = \inf_{y \in I_n^h} z(y), \quad x \in I_n^h = (nh, (n+1)h],$$

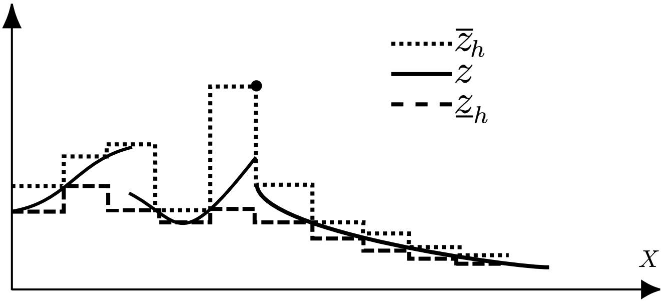

Figure 4.1

cf. Fig. 4.1. Then we call z *directly Riemann integrable* (d.R.i.) if $\int \overline{z}_h = \int_0^\infty \overline{z}_h(x)\mathrm{d}x$ is finite for some (and then all) h, and $\int \overline{z}_h - \int \underline{z}_h \to 0$ as $h \to 0$. For functions with compact support this concept is the same as Riemann integrability. If z can attain also negative values, we say that z is d.R.i. if both z^+ and z^- are so.

Proposition 4.1 *Suppose $z \geq 0$. Then if z is d.R.i., z is also Lebesgue integrable and $\int \overline{z}_h$, $\int \underline{z}_h$ have the common limit $\int z$ as $h \downarrow 0$. A necessary condition for z being d.R.i. is*
(i) *z is bounded and continuous a.e. w.r.t. Lebesgue measure.*
Sufficient conditions are:
(ii) *$\int \overline{z}_h < \infty$ for some h and* (i) *holds;*
(iii) *z has bounded support and* (i) *holds;*
(iv) *$z \leq z^*$ with z^* d.R.i. and* (i) *holds for z;*
(v) *z is nonincreasing and Lebesgue integrable.*

Proof. Boundedness is necessary for $\int \overline{z}_h < \infty$. Suppose that z is bounded but not a.e. continuous. Then if we let $\overline{z}(x) = \overline{\lim}_{y \to x} z(y)$, $\underline{z}(x) = \underline{\lim}_{y \to x} z(y)$, we have for some $\epsilon > 0$ that the Lebesgue measure δ, say, of $\{z : \overline{z}(x) > \underline{z}(x) + \epsilon\}$ is nonzero. But except possibly for $x = nh$ we have

$$\overline{z}_h(x) \geq \overline{z} \geq \underline{z} \geq \overline{z}_h \quad \text{so that} \quad \int \overline{z}_h - \int \underline{z}_h \geq \int \overline{z} - \int \underline{z} \geq \epsilon\delta,$$

and the necessity of (i) follows. In particular, if z is d.R.i., then by (i) $\overline{z}_h(x) \downarrow z(x)$ a.e. and $\lim_{h \downarrow 0} \overline{z}_h$ is Lebesgue integrable by monotone convergence. Hence z is so too, and $\int \overline{z}_h \to \int z$. Similarly, $\int \underline{z}_h \to \int z$. The same argument gives the sufficiency of (ii), and obviously (iii)⇒(ii). If (iv) holds, then

$\int \overline{z}_h \le \int \overline{z}^*_h < \infty$ and (ii) holds. Finally by a standard argument (v)⇒(ii). □

Example 4.2 We will verify that $z_A(t) = I(t \le \xi)\overline{F}(t)$, $z_B(t) = F(t + \xi) - F(t)$ in Example 2.1 are d.R.i. Here F, having a countable number of jumps, is continuous a.e., and thus the assertion for z_A follows from (iii). If $\mu < \infty$, we may apply (iv) to $z = z_B$, with $z^*(t) = \overline{F}(t)$ being d.R.i. according to (v). If $\mu = \infty$, let $m \in \mathbb{N}$ satisfy $\xi \le (m-1)h$. Then

$$\begin{aligned}\int \overline{z}_h &\le h\sum_{n=0}^{\infty}\bigl[F\bigl((n+1)h+\xi\bigr) - F(nh)\bigr]\\ &\le h\sum_{n=0}^{\infty}\bigl[F\bigl((n+m)h\bigr) - F(nh)\bigr]\\ &= h\lim_{N\to\infty}\sum_{n=0}^{N}\bigl[F\bigl((n+m)h\bigr) - F(nh)\bigr] \;\le\; h\lim_{N\to\infty}\sum_{k=N+1}^{N+m} F(kh)\end{aligned}$$

which is bounded by hm. Now use (ii). □

We can now state four different version 4.4–4.7 of the renewal theorem:

Theorem 4.3 *Suppose that F is nonlattice and proper* ($\|F\| = 1$) *and let* $\mu = \int_0^\infty x\,F(\mathrm{d}x)$, $F_0(t) = \mu^{-1}\int_0^t \overline{F}(y)\,\mathrm{d}y$ (i.e. $F_0 \equiv 0$ when $\mu = \infty$). *Then:*

4.4 (BLACKWELL'S RENEWAL THEOREM) *Let* $U = \sum_0^\infty F^{*n}$ *be the renewal function. Then for all a,*

$$U(t+a) - U(t) \;\to\; \frac{a}{\mu}, \quad t \to \infty.$$

More generally, in any (possibly delayed) *renewal process with interarrival distribution F the expected number $V_t(a)$ of renewals in $(t, t+a]$ tends to a/μ as $t \to \infty$.*

4.5 Let $\{A_t\}_{t\ge 0}$ be the backward recurrence time process in a (possibly delayed) *renewal process with interarrival distribution F. Then $\mathbb{P}(A_t \le \xi) \to F_0(\xi)$ for all ξ. In particular, if $\mu < \infty$ then $A_t \xrightarrow{\mathscr{D}} F_0$.*

4.6 Let $\{B_t\}_{t\ge 0}$ be the forward recurrence time process in a (possibly delayed) *renewal process with interarrival distribution F. Then $\mathbb{P}(B_t \le \xi) \to F_0(\xi)$ for all ξ. In particular, if $\mu < \infty$ then $B_t \xrightarrow{\mathscr{D}} F_0$.*

4.7 (KEY RENEWAL THEOREM) *Suppose that the function z in the renewal equation $Z = z + F * Z$ is d.R.i. Then*

$$Z(t) \;=\; U * z(t) \;\to\; \frac{1}{\mu}\int_0^\infty z(x)\,\mathrm{d}x, \quad t \to \infty.$$

We note that in the case $\mu < \infty$, 4.4–4.6 state that the renewal process becomes asymptotically stationary as $t \to \infty$. If $\mu = \infty$, 4.5 and 4.6 state that the mass in the distributions of A_t and B_t drifts off to ∞.

Proof of 4.7 ⇒ 4.5. Consider first the case of a zero–delayed renewal process, where according to Example 2.1 we have to show $Z_A(t) = U * z_A(t) \to F_0(\xi)$. But it was shown in Example 4.2 that z_A is d.R.i., hence by 4.7 the limit of $Z_A(t)$ exists and equals $\mu^{-1}\int z_A = F_0(\xi)$. In the case of a general delay distribution F_0^*, replace F by F_0^* in (2.4), let $t \to \infty$ and note that the first term tends to 0 and $F_0^* * Z_A(t)$ to $\lim Z_A(t) = F_0(\xi)$ by dominated convergence. □

Proof of 4.5 $\Longleftrightarrow$ 4.6. The equivalence is an immediate consequence of the identity $\{B_t \le \xi\} = \{A_{t+\xi} < \xi\}$ noted in (3.1). □

Proof of 4.6 ⇒ 4.4. Let $h(\xi) = U(a-\xi)I(\xi \le a)$, $G_t(\xi) = \mathbb{P}(B_t \le \xi)$. Then h is bounded and continuous a.e. w.r.t. dξ, hence a.e. w.r.t. $F_0(\mathrm{d}\xi)$, and since $G_t \to F_0$ in the sense of vague convergence (see A1 and Remark 4.9), (2.8) yields

$$\begin{aligned} V_t(a) \;&=\; (U * G_t)(a) \;=\; \int_0^\infty h(\xi)\, G_t(\mathrm{d}\xi) \\ &\to\; \int_0^\infty h(\xi)\, F_0(\mathrm{d}\xi) \;=\; U * F_0(a) \;=\; \frac{a}{\mu}. \end{aligned}$$

□

Proof of 4.4 ⇒ 4.7. Assume w.l.o.g. that $z \ge 0$. Let $nh < x \le (n+1)h$ and define $I_n = I_n(x) = \bigl(x-(k+1)h, x-kh\bigr]$. Then

$$\begin{aligned} Z(x) \;&=\; \int_0^x z(x-y)\,U(\mathrm{d}y) \\ &=\; \int_0^{x-nh} z(x-y)\,U(\mathrm{d}y) + \sum_{k=0}^{n-1}\int_{I_k} z(x-y)\,U(\mathrm{d}y)\,. \end{aligned}$$

Since $z(t) \to 0$ as $t \to \infty$, the first term tends to 0. The second is at most

$$\begin{aligned} &\sum_{k=0}^{n-1} \overline{z}_h(kh)\bigl[U(x-kh) - U\bigl(x-(k+1)h\bigr)\bigr] \\ &\quad\le\; \sum_{k=0}^{M} \overline{z}_h(kh)\bigl[U(x-kh) - U\bigl(x-(k+1)h\bigr)\bigr] + U(h)\sum_{k=M+1}^{n-1}\overline{z}_h(kh) \end{aligned}$$

since $U(t+h) - U(t) \le U(h)$ by Theorem 2.4(iii). Letting first $n, x \to \infty$ with M fixed, next $M \to \infty$ and finally $h \to 0$ yields

$$\begin{aligned} \overline{\lim_{x\to\infty}}\, Z(x) \;&\le\; \frac{h}{\mu}\sum_{k=0}^{M}\overline{z}_h(kh) + U(h)\sum_{k=M+1}^{\infty}\overline{z}_h(kh), \\ \overline{\lim_{x\to\infty}}\, Z(x) \;&\le\; \frac{1}{\mu}\int \overline{z}_h + 0, \quad \overline{\lim_{x\to\infty}}\, Z(x) \;\le\; \frac{1}{\mu}\int_0^\infty z(t)\,\mathrm{d}t. \end{aligned}$$

$\underline{\lim} \ge$ is proved similarly. □

Suitable versions of the renewal theorem also exist in the lattice case. Mathematically, this is somewhat easier and has to a large extent already been treated in I.2. For example:

Proposition 4.8 *Suppose that F is lattice with span δ and let $0 \le y < \delta$. Then if $\varphi(y) = \sum_0^\infty z(y+n\delta)$ converges absolutely, it holds that $U*z(y+n\delta) \to (\delta/\mu)\varphi(y)$ as $n \to \infty$.*

Proof. The renewal measure U is supported by $\{0, \delta, 2\delta, \ldots\}$, with mass say $u_{k\delta}$ at $k\delta$. By I.2.2, $u_{k\delta} \to \delta/\mu$ as $k \to \infty$, and hence by dominated convergence

$$\begin{aligned} Z(y+n\delta) &= (U*z)(y+n\delta) = \sum_{k=0}^{n} z(y+n\delta-k\delta)u_{k\delta} \\ &= \sum_{k=0}^{n} z(y+k\delta)u_{(n-k)\delta} \to \frac{\delta}{\mu}\varphi(y)\,. \end{aligned}$$

□

Remark 4.9 The connection between Blackwell's renewal theorem and the key renewal theorem may in more abstract terms be rephrased as follows. Consider for each t the measure ν_t on $[0,t]$ obtained by time reversion of the renewal measure restricted to $[0,t]$, i.e. $\int_0^\infty f(x)\,\nu_t(\mathrm{d}x) = \int_0^t f(t-y)\,U(\mathrm{d}y)$. Then Blackwell's theorem asserts that $\nu_t[0,a) \to a/\mu$ for all a, which by general results from measure theory is equivalent to $\nu_t(\mathrm{d}a) \to \mathrm{d}a/\mu$ vaguely (i.e. $\int f(a)\,\nu_t(\mathrm{d}a) \to \int f(a)\,\mathrm{d}a/\mu$ whenever f has compact support and is continuous or, more generally, bounded and a.e. continuous). Any such f is d.R.i., and hence we may view the key renewal theorem as an extension of Blackwell's theorem to also cover certain f with unbounded support, a case of major importance for applications. □

Problems

4.1 Show that the stationary distribution F_1 of the current life in Section 3 is also a limiting distribution.

4.2 Show that if F has a d.R.i. density f so that the renewal density u exists, then $u(x) \to 1/\mu$.

4.3 Show that the z in Problem 2.2 is d.R.i. and express $\lim Z(t)$ in terms of the Laplace transform of G.

4.4 Find examples of functions that are Lebesgue integrable but not d.R.i.

4.5 Give a simplified proof of Theorem 3.5 by invoking the renewal theorem.

4.6 Show that Blackwell's theorem remains valid if F is allowed to have an atom at 0. [*Hint:* Show first that B_t converges in distribution and apply next Wald's identity to S_{N_t+1}.]

4.7 Show that if z in the renewal equation is not necessarily d.R.i. but only bounded with $z(x) \to 0$, $x \to \infty$, then $Z(x) = \mathrm{o}(x)$.

4.8 Show that if $z(x) \sim cx^\alpha$ with $\alpha > 0$ in the renewal equation, then $Z(x) \sim cx^{\alpha+1}/\mu\alpha$.

5 Proof of the Renewal Theorem

The formulations 4.5–4.6 of the renewal theorem in terms of the recurrence time processes shows that for the lattice case the situation is essentially settled by the analysis of Chapter I. The nonlattice case is considerably more involved, and no really short and elementary approach is known. We give in this section a standard analytical proof developed largely by Feller (1971). A parallel more probabilistic proof is then given in VII.2. Some ingredients are common, in particular the way the nonlattice property comes in:

Lemma 5.1 *Suppose that F is nonlattice on $(0,\infty)$ and define $U = \sum_0^\infty F^{*n}$, $S = \mathrm{supp}(U)$. Then S is asymptotically dense at ∞ in the sense that $d(x,S) = \inf_{y\in S}|x-y| \to 0$, $x\to\infty$.*

Proof. S is the closure of $\cup_0^\infty \mathrm{supp} F^{*n}$ which is asymptotically dense by A7.1 (take $\mathscr{G} = \mathrm{supp}(F)$ there). □

Proposition 5.2 (CHOQUET–DENY) *If F is a nonlattice distribution on $(0,\infty)$ and φ a bounded continuous function on $\mathbb{R}$ satisfying $\varphi = F*\varphi$, then φ is necessarily constant.*

Proof. Since

$$\begin{aligned}\mathbb{E}[\varphi(x-S_{n+1})\,|\,Y_1,\ldots,Y_n] &= \int_0^\infty \varphi(x-S_n-y)\,F(\mathrm{d}y)\\ &= (F*\varphi)(x-S_n) \;=\; \varphi(x-S_n),\end{aligned}$$

the sequence $\{\varphi(x-S_n)\}$ is a bounded martingale and hence converges a.s. and in L_2. By the Hewitt–Savage 0–1 law, the limit is almost surely constant which by L_2–theory for martingales implies that $\varphi(x) = \varphi(x-S_1) = \cdots = \varphi(x-S_n)\cdots$ a.s. Thus $\varphi(x-u) = \varphi(x)$ for F^{*n}–a.a. u, which by the continuity of φ shows that $\varphi(x-u) = \varphi(x)$ for all $u \in \mathrm{supp}(F^{*n})$. Now let a,b be given. Then by Lemma 5.1 we can choose sequences $\{a_n\}$, $\{b_n\}$ with $n-a_n \to a$, $n-b_n\to b$ and $a_n,b_n \in \cup_1^\infty \mathrm{supp}(F^{*n})$. Then

$$\varphi(a) \;=\; \lim\varphi(n-a_n) \;=\; \lim\varphi(n) \;=\; \lim\varphi(n-b_n) \;=\; \varphi(b)\,. \qquad \square$$

Now let $\lambda^{(t)}(\mathrm{d}y) = U(t-\mathrm{d}y)$, $\lambda(\mathrm{d}y) = \mathrm{d}y/\mu$, $t\ge 0$, so that the renewal theorem is equivalent to $\lambda^{(t)} \to \lambda$ vaguely; cf. Remark 4.9. To show this, it is sufficient to show that each sequence $\{s_n\}$ with $s_n\to\infty$ has a subsequence $\{t_n\}$ with $\lambda^{(t_n)} \to \lambda$. By Theorem 2.4(iii), $\sup_n \lambda^{(s_n)}(K) < \infty$ for any compact set K which by standard facts from measure theory implies that $\{\lambda^{(s_n)}\}$ is vaguely compact. Thus $\lambda^{(t_n)} \to \nu$ for some subsequence $\{t_n\}$ and some ν, and we have to show $\nu = \lambda$.

Lemma 5.3 $\lambda^{(t_n+x)} \to \nu$ *for all $x\in\mathbb{R}$.*

Proof. With $\nu^{(x)}(\mathrm{d}y) = \nu(\mathrm{d}y - x)$, it is clear that $\lambda^{(t_n+x)} \to \nu^{(x)}$ for all x. To get $\nu^{(x)} = \nu$, it is sufficient to show that continuous functions z with compact support have the same integral, i.e. that

$$\varphi(x) \;=\; \int z(y)\nu^{(x)}(\mathrm{d}y) \;=\; \lim_{n\to\infty} \int z(y)\lambda^{(t_n+x)}(\mathrm{d}y) \;=\; \lim_{n\to\infty} (U*z)(t_n+x)$$

is independent of x. But $Z = U * z$ being a solution of a renewal equation, this implies that $\varphi(x)$ equals

$$\lim_{n\to\infty} \left\{ z(t_n+x) \;+\; \int_0^{t_n+x} Z(t_n+x-y)\,F(\mathrm{d}y) \right\} \;=\; 0 + \int_0^{\infty} \varphi(x-y)\,F(\mathrm{d}y)$$

(using dominated convergence). Hence by the Choquet–Deny theorem, we have only to show that φ is continuous which will follow if Z is uniformly continuous. But let z be supported by $[0,T]$ and define $\kappa(\epsilon) = \sup_{|x-y|\le\epsilon} |z(x)-z(y)|$. Then for $|x_1-x_2| \le \epsilon$,

$$\begin{aligned} Z(x_1) - Z(x_2) \;&=\; \left| \int_0^{x_1} z(x_1-y)\,U(\mathrm{d}y) \;-\; \int_0^{x_2} z(x_2-y)\,U(\mathrm{d}y) \right| \\ &\le\; \kappa(\epsilon) U(T+\epsilon) \;\to\; 0, \quad \epsilon \to 0. \end{aligned}$$

□

Proof of the renewal theorem. The conclusion of Lemma 5.3 means that ν is translation invariant, i.e. $\nu(\mathrm{d}y) = \gamma\,\mathrm{d}y$ for some γ, and we have to show $\gamma = 1/\mu$. From

$$U(t_n-a) - U(t_n-a-h) \;=\; \lambda^{(t_n)}[a, a+h) \;\to\; \gamma h$$

for all $h > 0$ it follows by the same arguments as in the proof of 4.4⇒4.7 that $U * z(t_n) \to \gamma \int z$ whenever z is d.R.i. If $\mu < \infty$, let $z(x) = \overline{F}(x)$. Then $U * z \equiv 1$, cf. Lemma 3.4 , and $\int z = \mu$ so that $1 = \gamma\mu$. If $\mu = \infty$, let $z(x) = \overline{F}(x) I(x \le x_0)$. Then similarly we get $\gamma \int z \le 1$ which, letting $x_0 \to \infty$, yields $\gamma\mu = \gamma \cdot \infty \le 1$ and $\gamma = 0 = 1/\mu$. □

Notes The present proof of the renewal theorem follows Feller (1971) closely. For alternative proofs, we mention in particular the coupling proof in VII.2, a Markov chain proof due to McDonald (1975) and finally the Fourier analytic proofs that can be found e.g. in Woodroofe (1982).

6 Second–Moment Results

We are concerned with certain refinements of renewal theory, which require the existence of the second moment $\mathbb{E}Y^2$ of the interarrival distribution F, or equivalently that $\sigma^2 = \mathbb{V}arY < \infty$. For simplicity, only the nonlattice case is considered.

The first (and simplest) problem to be studied is to look for expansions of the renewal function $U(t)$ more detailed than the one $U(t) \sim t/\mu$ provided by the elementary renewal theorem. This can be obtained by noting that

$$S_{N_t} \;=\; t+B_t \tag{6.1}$$

and taking expectations: by Wald's identity,

$$\mathbb{E}S_{N_t} \;=\; \mu\mathbb{E}N_t \;=\; \mu U(t), \quad U(t) \;=\; \frac{t}{\mu}+\frac{\mathbb{E}B_t}{\mu}. \tag{6.2}$$

Furthermore, from $B_t \xrightarrow{\mathscr{D}} F_0$ one expects that

$$\mathbb{E}B_t \;\to\; \int_0^\infty x\,F_0(\mathrm{d}x) \;=\; \frac{1}{\mu}\int_0^\infty x\overline{F}(x)\,\mathrm{d}x \;=\; \frac{\mathbb{E}Y^2}{2\mu} \;=\; \frac{\sigma^2+\mu^2}{2\mu}. \tag{6.3}$$

To see that this is indeed the case, evaluate e.g. $Z(t)=\mathbb{E}B_t$ by the renewal argument and check that

$$z(t) \;=\; \mathbb{E}[B_t;\,Y_1>t] \;=\; \mathbb{E}[Y_1-t;\,Y_1>t]$$

is directly Riemann integrable with integral $\mathbb{E}Y^2/2$. Combining (6.1)–(6.3), we have proved:

Proposition 6.1 $U(t) \;=\; \dfrac{t}{\mu}+\dfrac{\mathbb{E}Y^2}{2\mu^2}+\mathrm{o}(1), \quad t\to\infty.$

One might expect that (assuming higher order moments) the o(1) term could be further expanded as $c_1/t+c_2/t^2+\cdots$. However, under suitable regularity conditions the rate of decay is in fact exponentially fast; cf. Problem VII.2.2.

Ignoring B_t in (6.1) yields the lower bound $U(t)\ge t/\mu$. We shall next find an upper bound somewhat related to the asymptotic expression in Proposition 6.1:

Proposition 6.2 (LORDEN'S INEQUALITY) $U(t) \;\le\; \dfrac{t}{\mu}+\dfrac{\mathbb{E}Y^2}{\mu^2}.$

Proof. According to (6.2) we must show that $\mathbb{E}B_t\le\mathbb{E}Y^2/\mu$ for all t. By (2.8), $U(t+s)-U(s)\le U(t)$ so that (6.1) and (6.2) yield

$$\begin{aligned}\mathbb{E}B_t+\mathbb{E}B_s \;&=\; \mu[U(t)+U(s)]-t-s \;\ge\; \mu U(t+s)-t-s \;=\; \mathbb{E}B_{t+s},\\ \mathbb{E}B_t \;&\le\; \inf\bigl\{\mathbb{E}B_s+\mathbb{E}B_{t-s}\,:\,0\le s\le t/2\bigr\}\\ &\le\; \frac{2}{t}\int_0^{t/2}\bigl[\mathbb{E}B_s+\mathbb{E}B_{t-s}\bigr]\,\mathrm{d}s \;=\; \frac{2}{t}\int_0^t \mathbb{E}B_s\,\mathrm{d}s\,. \end{aligned}\tag{6.4}$$

Now an inspection of the paths shows that

$$\int_0^t B_s\,\mathrm{d}s \;=\; \frac{1}{2}\sum_{n=1}^{N_t}Y_n^2-\frac{1}{2}B_t^2.$$

Thus

$$\int_0^t \mathbb{E}B_s\,\mathrm{d}s \;=\; \frac{1}{2}\mathbb{E}N_t\mathbb{E}Y^2-\frac{1}{2}\mathbb{E}B_t^2 \;=\; \frac{1}{2}U(t)\mathbb{E}Y^2-\frac{1}{2}\mathbb{E}B_t^2$$

$$= \frac{1}{2}(t+\mathbb{E}B_t)\mathbb{E}Y^2/\mu - \frac{1}{2}\mathbb{E}B_t^2. \tag{6.5}$$

Letting $\alpha = \mathbb{E}B_t$ we have $\mathbb{E}B_t^2 \geq \alpha^2$ and combining (6.4), (6.5) yields

$$t\alpha \;\leq\; (t+\alpha)\mathbb{E}Y^2/\mu - \alpha^2,$$

i.e.

$$\alpha^2 + \alpha(t - \mathbb{E}Y^2/\mu) - t\mathbb{E}Y^2/\mu \;\leq\; 0. \tag{6.6}$$

But the l.h.s. of (6.6) is a quadratic in α with roots $-t$ and $\mathbb{E}Y^2/\mu$. Thus $-t \leq \alpha \leq \mathbb{E}Y^2/\mu$. □

Our next objective is to establish the CLT for the number of renewals and the corresponding expansion of the variance:

Proposition 6.3 (a) *As $t \to \infty$, N_t is asymptotically normal with mean t/μ and variance $t\sigma^2/\mu^3$;*
(b) $\mathbb{V}ar N_t \;=\; \dfrac{t\sigma^2}{\mu^3} + \mathrm{o}(t)$.

Proof. The results are also valid for general delay distributions, but will for simplicity only be proved for the zero–delayed case. Here (a) can easily be shown by applying Anscombe's theorem to

$$U_t \;=\; \frac{S_{N_t} - N_t\mu}{t^{1/2}} \;=\; \frac{B_t + t - N_t\mu}{t^{1/2}} \;\approx\; \frac{t - N_t\mu}{t^{1/2}} \tag{6.7}$$

(see VI.3.2 for details). An elementary direct argument is as follows: let y be fixed and let $n = n(t)$ depend on t in such a way that

$$t/\mu + (t\sigma^2/\mu^3)^{1/2}y \in (n-1, n].$$

Then $t(1+\mathrm{o}(1)) = n\mu$, from which we get $t = n\mu + \mathrm{o}(n)$ and (by Taylor expansion) $t^{1/2} = (n\mu)^{1/2} + \mathrm{o}(n^{1/2})$ so that

$$t \;=\; n\mu + \mathrm{O}(1) - (t\sigma^2/\mu)^{1/2}y \;=\; n\mu - \sigma y n^{1/2} + \mathrm{o}(n^{1/2}).$$

Therefore the CLT for S_n yields

$$\begin{aligned}\mathbb{P}\Bigl(\frac{N_t - t/\mu}{(t\sigma^2/\mu^3)^{1/2}} \leq y\Bigr) \;&=\; \mathbb{P}(N_t \leq n) \;=\; \mathbb{P}(S_n > t)\\ \;&=\; \mathbb{P}\Bigl(\frac{S_n - n\mu}{\sigma n^{1/2}} > -y + \mathrm{o}(1)\Bigr) \;\to\; 1 - \Phi(-y) \;=\; \Phi(y),\end{aligned}$$

proving (a). The proof of (b) can be carried out in a number of ways, none of which are entirely brief. Recalling (6.1), (6.2) and (6.7), we let

$$V_t \;=\; \mu\frac{\mathbb{E}N_t - N_t}{t^{1/2}}, \quad W_t \;=\; U_t - V_t \;=\; \frac{B_t - \mathbb{E}B_t}{t^{1/2}}$$

and have to prove $\mathbb{E}V_t^2 \to \sigma^2/\mu$. We first note that a renewal argument shows that $Z(t) = \mathbb{E}B_t^2$ satisfies $Z = z + F * Z$ where $z(t) = \mathbb{E}[B_t^2; t < Y_1]$. Since $B_t \leq Y_1$ on $\{t < Y_1\}$, we have $z(t) \to 0$ so that Problem 4.7

yields $\mathbb{E}B_t^2 = \mathrm{o}(t)$ and therefore $\mathbb{E}W_t^2 \to 0$. Now by Wald's second moment identity A10.2(b), $\mathbb{E}U_t^2 = \sigma^2\mathbb{E}N_t/t \to \sigma^2/\mu$. Hence by the Cauchy–Schwarz inequality, $\mathbb{E}[U_tW_t] \to 0$ and

$$\mathbb{E}V_t^2 \;=\; \mathbb{E}U_t^2 + \mathbb{E}W_t^2 - 2\mathbb{E}[U_tW_t] \;\to\; \sigma^2/\mu + 0 - 0\,. \qquad \square$$

Problems

6.1 Check the details in the proof of (6.3) outlined in the text.
6.2 Give an alternative proof of Proposition 6.1 by showing that $Z(t) = U(t) - t/\mu$ satisfies a renewal equation with $z(t) = \overline{F}_0(t)$.
6.3 Show that $\mathbb{E}N_t^2 = 2(U*U)(t) - U(t)$, that $(U*U)(t) = t^2/2\mu^2 + t\mathbb{E}Y^2/2\mu^3 + \mathrm{o}(t)$, and give hereby a different derivation of the asymptotic form of $\mathbb{V}arN_t$.

7 Excessive and Defective Renewal Equations

Recall that the renewal equation $Z = z + F*Z$ is called *excessive* if $\|F\| > 1$ and *defective* if $\|F\| < 1$. We still have $Z = U*z$ (provided Z, z are bounded on finite intervals), but Blackwell's renewal theorem does not apply to determine the asymptotic behaviour of U and thereby Z. However, by a transformation we may often reduce to the case $\|F\| = 1$:

Theorem 7.1 *Assume that $Z = z + F*Z$ with Z, z bounded on finite intervals and that for some real β*

$$\widehat{F}[\beta] \;=\; \int_0^\infty \mathrm{e}^{\beta x}\, F(\mathrm{d}x) \;=\; 1. \tag{7.1}$$

Define

$$\widetilde{Z}(x) \;=\; \mathrm{e}^{\beta x} Z(x), \quad \widetilde{z}(x) \;=\; \mathrm{e}^{\beta x} z(x), \quad \widetilde{F}(dx) \;=\; \mathrm{e}^{\beta x} F(\mathrm{d}x).$$

Then $\widetilde{Z} = \widetilde{z} + \widetilde{F} * \widetilde{Z}$ and if $\widetilde{z}$ is directly Riemann integrable (d.R.i.)

$$\lim_{x\to\infty} \mathrm{e}^{\beta x} Z(x) \;=\; \lim_{x\to\infty} \widetilde{Z}(x) \;=\; \frac{1}{\widetilde{\mu}}\int_0^\infty \widetilde{z}(t)\,\mathrm{d}t \;=\; \frac{\int_0^\infty \mathrm{e}^{\beta t} z(t)\,\mathrm{d}t}{\int_0^\infty t\mathrm{e}^{\beta t}\,F(\mathrm{d}t)} \tag{7.2}$$

Proof. Clearly $\widetilde{Z}$, $\widetilde{z}$ are bounded on finite intervals and $\|\widetilde{F}\| = 1$ by (7.1). Also

$$\begin{aligned}\widetilde{Z}(x) \;&=\; \mathrm{e}^{\beta x}\left\{z(x) + \int_0^x Z(x-y)\,F(\mathrm{d}y)\right\}\\ &=\; \widetilde{z}(x) + \int_0^x \mathrm{e}^{\beta(x-y)} Z(x-y)\,\mathrm{e}^{\beta y} F(\mathrm{d}y) \;=\; \widetilde{z}(x) + (\widetilde{F} * \widetilde{Z})(x)\end{aligned}$$

and the remaining statements are clear from results for the case $\|F\| = 1$. $\square$

For a closer study of the assumption (7.1) and its implication (7.2), we need to treat the excessive and the defective case separately.

Proposition 7.2 *Consider the excessive case* $1 < \|F\| \le \infty$. *Then a solution* β *to* (7.1) *is necessarily strictly negative,* $\beta < 0$. *A sufficient condition for the existence of* β *is* $1 < \widehat{F}[\delta] < \infty$ *for some* δ *and then always* $\widetilde{\mu} < \infty$. *This holds in particular if* $\|F\| < \infty$.

Proof. Since $\widehat{F}[\beta] \ge \|F\|$ for $\beta \ge 0$, it is clear that (7.1) implies $\beta < 0$. If δ exists, then by monotone convergence $\widehat{F}[\beta]$ is a continuous function of β with limits 0 as $\beta \to -\infty$ and $\widehat{F}[\delta] > 1$ as $\beta \uparrow \delta$. Hence the value 1 is attained and the F–integrability of $\mathrm{e}^{\delta x}$ implies that of $x\mathrm{e}^{\beta x}$, i.e. $\widetilde{\mu} < \infty$. Finally, if $\|F\| < \infty$, we can just take $\delta = 0$. □

Example 7.3 Consider Lotka's integral equation (Example 2.2) for the density $Z(t)$ of births in a population at time t and assume that, as will typically be the case, that the net reproduction rate $\|F\|$ is > 1. The assumption $\|F\| < \infty$ is innocent from the demographic point of view and hence we may conclude that β exists, is < 0 and that

$$\widetilde{\mu} \;=\; \int_0^\infty s\mathrm{e}^{\beta s}\,F(\mathrm{d}s) \;=\; \int_0^\infty s\mathrm{e}^{\beta s}{}_sp_0\lambda(s)\,\mathrm{d}s \;<\; \infty.$$

Also the assumption of $\widetilde{z}$ being d.R.i. is innocent. In fact, inspection of the expression for z shows this to hold if only the birth intensity $\lambda(u)$ is bounded and continuous and the survival rate ${}_tp_a$ is continuous (then $z(t)$ is bounded and continuous, hence $\widetilde{z}(t) = \mathrm{e}^{\beta t}z(t)$ is d.R.i. because of Proposition 4.1(iv)–(v) and $\beta < 0$). Thus under these assumptions, $Z(t)$ grows asymptotically exponentially, $Z(t) \sim C\mathrm{e}^{-\beta t}$, where

$$C \;=\; \frac{1}{\widetilde{\mu}}\int_0^\infty \widetilde{z}(t)\,\mathrm{d}t \;=\; \frac{1}{\widetilde{\mu}}\int_0^\infty\int_0^\infty \mathrm{e}^{\beta t}f_0(a){}_tp_a\lambda(a+t)\,\mathrm{d}t\,\mathrm{d}a$$

(the rate $-\beta > 0$ is usually called the *Malthusian parameter* of the population). From this the limiting behaviour of other quantities is easily obtained. For example, for the total population size $N(t)$ we easily obtain from (2.6) that

$$\begin{aligned}\mathrm{e}^{\beta t}N(t) \;&=\; \int_0^t \mathrm{e}^{\beta(t-a)}Z(t-a)\mathrm{e}^{\beta a}{}_ap_0\,\mathrm{d}a \,+\, \mathrm{e}^{\beta t}\int_0^t f_0(a){}_{t-a}p_a\,\mathrm{d}a \\ &\to\; C\int_0^\infty \mathrm{e}^{\beta a}{}_ap_0\,\mathrm{d}a \,+\, 0.\end{aligned}$$

□

In the defective case, a simple conclusion can be obtained without reference to condition (7.1):

Proposition 7.4 *If in the defective case* z *is bounded and* $z(\infty) = \lim_{t\to\infty} z(t)$ *exists, then* $Z(t) \to z(\infty)/(1-\|F\|) = Z(\infty)$ (say).

Proof. Using dominated convergence and (2.10) we get

$$Z(t) \;=\; \int_0^t z(t-y)\,U(\mathrm{d}y) \;\to\; \int_0^\infty z(\infty)\,U(\mathrm{d}y) \;=\; \frac{z(\infty)}{1-\|F\|}. \qquad \square$$

If $z(\infty)=0$, this result is rather imprecise, and also in some cases with $z(\infty)\neq 0$ it is of substantial interest to estimate the rate of convergence of $Z(t)$ to $Z(\infty)$. To this end (7.1) comes in. However, as (7.1) already shows, the conditions for the existence of β are rather stronger than in the excessive case and require the existence of exponential moments. We have the following analogue of Proposition 7.2:

Proposition 7.5 *Consider the defective case $\|F\|<1$. Then a solution β to* (7.1) *is necessarily strictly positive, $\beta>0$. A sufficient condition for the existence of β is $1<\widehat{F}[\delta]<\infty$ for some δ, and then always $\widetilde{\mu}<\infty$. This holds in particular if $\widehat{F}[\delta]<\infty$ for all $\delta\in\mathbb{R}$.*

Proof. Exactly as for Proposition 7.2, except that for the last step one notes that if $\widehat{F}[\delta]<\infty$ for all δ, then $\widehat{F}[\delta]\to\infty$ as $\delta\to\infty$. □

Proposition 7.6 *Suppose that in the defective case $\|F\|<1$ a solution β to* (7.1) *satisfying $\widetilde{\mu}<\infty$ exists. If $z(\infty)=\lim_{t\to\infty}z(t)$ exists and $\mathrm{e}^{\beta t}(z(t)-z(\infty))$ is d.R.i., then*

$$Z(t)-Z(\infty) \;\sim\; \mathrm{e}^{-\beta t}\frac{1}{\widetilde{\mu}}\Bigl\{\int_0^\infty \mathrm{e}^{\beta s}\bigl[z(s)-z(\infty)\bigr]\,\mathrm{d}s \;-\; \frac{z(\infty)}{\beta}\Bigr\}. \qquad (7.3)$$

Proof. Define $Z_1(t)=Z(t)-Z(\infty)$,

$$z_1(t) \;=\; z(t)-z(\infty)+z(\infty)\frac{F(t)-\|F\|}{1-\|F\|} \;=\; z(t)-z(\infty)\frac{\overline{F}(t)}{1-\|F\|}\,.$$

Since $U*F=U-1$, we get

$$U*z_1 \;=\; U*z-\frac{z(\infty)}{1-\|F\|} \;=\; Z-Z(\infty) \;=\; Z_1.$$

Now since $\beta>0$,

$$\mathrm{e}^{\beta t}\bigl[\|F\|-F(t)\bigr] \;=\; \mathrm{e}^{\beta t}\int_t^\infty F(\mathrm{d}s) \;\le\; \int_t^\infty \mathrm{e}^{\beta s}\,F(\mathrm{d}s) \;=\; 1-\widetilde{F}(t)\,.$$

The r.h.s. is nonincreasing with integral $\widetilde{\mu}$. Thus the l.h.s. is d.R.i., hence $\mathrm{e}^{\beta t}z_1(t)$ is so and (7.2) yields

$$\mathrm{e}^{\beta t}Z_1(t) \;\to\; \frac{1}{\widetilde{\mu}}\int_0^\infty \mathrm{e}^{\beta s}z_1(s)\,\mathrm{d}s\,.$$

Thus (7.3) follows from

$$\int_0^\infty \mathrm{e}^{\beta t}[\|F\|-F(t)]\,\mathrm{d}t \;=\; \int_0^\infty \mathrm{e}^{\beta t}\,\mathrm{d}t\int_t^\infty F(\mathrm{d}s)$$
$$=\; \int_0^\infty \frac{1}{\beta}(\mathrm{e}^{\beta s}-1)\,F(\mathrm{d}s) \;=\; \frac{1}{\beta}(1-\|F\|).$$

□

Example 7.7 In Proposition 7.6, take $z \equiv 1$. Then if β exists and $\widetilde{\mu} < \infty$, we obtain

$$\|U\| - U(t) \;\sim\; \frac{1}{\beta\widetilde{\mu}}\mathrm{e}^{-\beta t}. \tag{7.4}$$

□

Example 7.8 Consider the ruin problem (Example 2.3) and assume that, as will typically be the case, the premium exceeds the expected claims. Recalling that the probability $Z(u)$ of ultimate survival with initial reserve u satisfies $Z = z + F * Z$ with $z(u) = Z(0)$, $F(\mathrm{d}x) = (\lambda/c)\overline{G}(x)\,\mathrm{d}x$, where λ is the arrival intensity, c the premium per unit time and G the claim size distribution, this amounts to $\lambda\nu/c < 1$ where $\nu = \int_0^\infty x\,G(\mathrm{d}x)$.

It remains to evaluate $Z(0)$. First note that

$$ct - \sum_{n=1}^{N_t} X_n \;\approx\; t\Bigl(c - \frac{N_t}{t}\nu\Bigr) \;\approx\; t(c - \lambda\nu)$$

so that $ct - \sum_1^{N_t} X_n$ tends to infinity, hence attains a minimum $m > -\infty$ and thus $Z(u) = \mathbb{P}(u + m \geq 0) \to 1$ as $u \to \infty$. Using Proposition 7.4 we therefore get $1 = Z(0)/(1 - \lambda\nu/c)$, i.e. $Z(0) = 1 - \lambda\nu/c$. Since for $\beta > 0$

$$\widehat{F}[\beta] \;=\; \frac{\lambda}{c}\int_0^\infty \mathrm{e}^{\beta x}\overline{G}(x)\,\mathrm{d}x \;=\; \frac{\lambda}{c\beta}\int_0^\infty (\mathrm{e}^{\beta x} - 1)\,G(\mathrm{d}x) \;=\; \frac{\lambda}{c\beta}(\widehat{G}[\beta] - 1),$$

the assumption $\widehat{F}[\beta] = 1$, $\widetilde{\mu} < \infty$ amounts to

$$\widehat{G}[\beta] \;=\; 1 + \frac{c}{\lambda}\beta \tag{7.5}$$

for some $\beta > 0$ satisfying $\widehat{G}'[\beta] < \infty$; cf. Fig. 7.1. Since $z(t) = z(\infty)$ and $Z(\infty) = 1$, we have thus from Proposition 7.6 derived the celebrated *Cramér–Lundberg approximation* for the probability $1 - Z(u)$ of ultimate ruin,

$$1 - Z(u) \;\sim\; \mathrm{e}^{-\beta u}\frac{1 - \lambda\nu/c}{\widetilde{\mu}\beta} \tag{7.6}$$

The equation (7.5) is known as the *Lundberg equation.* □

If $\|F\| < 1$ but F is heavy–tailed, β will fail to exist. One has the following heavy–tailed counterpart of (7.4):

Proposition 7.9 *Assume $\|F\| < 1$ and that $G = F/\|F\|$ is subexponential* (cf. A5)*, and write $\overline{F}(t) = F(t,\infty)$. Then*

$$\|U\| - U(t) \;\sim\; \frac{1}{1 - \|F\|}\overline{F}(t)\,. \tag{7.7}$$

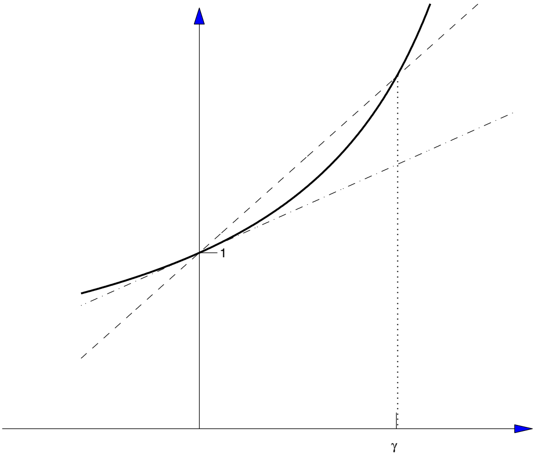

Figure 7.1

—— is the m.g.f. $\hat{G}[\alpha]$ of G, $- \cdot - \cdot$ is the tangent $1 + \hat{G}'[0]\alpha = 1 + \nu\alpha$ of $\hat{G}[\alpha]$ at 0, and $- - -$ is the line $1 + c\alpha/\lambda$

Proof. Recall from the proof of Proposition 2.9 that $\|U\| - U(t) = \mathbb{P}(Y_1 + \cdots + Y_\sigma > t)$ where $Y_1, Y_2, \ldots$ are i.i.d. with distribution G and N an independent r.v. with $\mathbb{P}(\sigma = n) = (1 - \|F\|)\|F\|^n$. By a general lemma on the tail of a random sum of subexponential r.v.'s, cf. X.9.2, we then obtain

$$\|U\| - U(t) \;\sim\; \mathbb{E}\sigma\,\overline{G}(t) \;=\; \frac{\|F\|}{1 - \|F\|}\frac{\overline{F}(t)}{\|F\|} \;=\; \frac{\overline{F}(t)}{1 - \|F\|}. \qquad \square$$

Notes The exponential transformations above are related to the general techniques studied in Ch. XIII, and in particular we will revisit the Cramér–Lundberg approximation in XIII.5 (for reasons to become clear there, we rewrite the Lundberg equation in the form $\lambda(\widehat{G}[\beta] - 1) - c\beta = 0$ there; cf. Problem XIII.1.2).

In the heavy–tailed case, the typical asymptotical counterpart of Proposition 7.6 for the case $z(\infty) = 0$ is

$$Z(t) \;\sim\; \begin{cases} \dfrac{1}{1 - \|F\|} \cdot z(x) & \text{if} \quad \dfrac{z(x)}{f(x)} \to \infty\,, \\[2ex] \left(\dfrac{\int_0^\infty z(y)\,\mathrm{d}y}{(1 - \|F\|)^2} + \dfrac{c}{1 - \|F\|}\right) g(x) & \text{if} \quad \dfrac{z(x)}{f(x)} \to c \in (0, \infty)\,, \\[2ex] \dfrac{\int_0^\infty z(y)\,\mathrm{d}y}{(1 - \|F\|)^2} \cdot f(x) & \text{if} \quad \dfrac{z(x)}{f(x)} \to 0 \end{cases} \tag{7.8}$$

(assuming for simplicity the existence of a density f of F). However, the precise formulation and proof requires some care and we refer to Asmussen *et al.* (2003). Heuristically, one arrives at the result by considering the distributions G, H where $G = U/\|U\|$ and H has density $z/\int z$. Then $Z = U * z = ck$ where k is the density of $G * H$ and $c = \|U\| \int z = \int z/(1 - \|F\|)$. In the case of at least one of G, H being subexponential, the asymptotics of k should be that of the heavier of the densities of G, H if one dominates and otherwise that of the sum (see A5.1(c)). For example, in the case $f/z \to \infty$, this combined with (7.7) leads to $Z \sim cu/\|U\| \sim f \int z/(1 - \|F\|)^2$ as in the last case in (7.8).

VI
Regenerative Processes

1 Basic Limit Theory

The classical definition of a stochastic process $\{X_t\}$ to be regenerative means in intuitive terms that the process can be split into i.i.d. *cycles*. A basic example is the $GI/G/1$ queue length process and its busy cycles, i.e. the time intervals separated by the instants S_n with a customer entering an empty systems; cf. Fig. 1.1. At each such instant the queue *regenerates*, i.e. starts completely from scratch independently of the past. Different cycles are independent and all governed by the same probability law. Similar statements hold for the workload or other processes associated with the system.

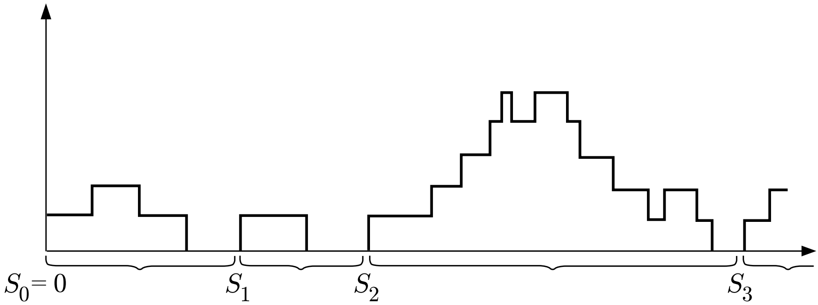

Figure 1.1

This structure with i.i.d. cycles is found in the majority of examples and can most often safely be used as a guide for intuition, but we shall use a slightly wider definition. Assume that $\{X_t\}$ has state space E and discrete or continuous time parameter $t \in \mathbb{T}$, where $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = [0, \infty)$. We then call $\{X_t\}_{t \in \mathbb{T}}$ *regenerative* (pure or delayed) if there exists a renewal process (pure or delayed) $\{S_n\} = \{Y_0 + \cdots + Y_n\}$ with the following property: for each $n \geq 0$, the post–S_n process

$$\theta_{S_n} X \;=\; \bigl(Y_{n+1}, Y_{n+2}, \ldots, \{X_{S_n+t}\}_{t \in \mathbb{T}}\bigr)$$

is independent of $S_0, \ldots, S_n$ (or, equivalently, of $Y_0, \ldots, Y_n$) and its distribution does not depend upon n. We call $\{S_n\}$ the *imbedded renewal process* and refer to the S_n as *regeneration points*. The kth cycle is $\{X_{t+S_k}\}_{0 \leq t < Y_{k+1}}$, a random element of the space $D_0(E)$ of E–valued D–functions with finite lifelengths; cf. A2 (in discrete time, consider instead the space of finite E–valued sequences).

Concerning the definition, we note the following points:

(i) Cycles are still welldefined and all governed by the same probability law, but some dependence between cycles may occur (for main examples, see Section 2e and VII.3).

(ii) The embedded renewal process is by no means unique. For example, we may as well use $\{S_{2n}\}$ or, in the $M/G/1$ case of the queueing example, take the regeneration points as the instants where idle periods start.

(iii) The S_n need not be given as functions of $\{X_t\}$. In particular, there are some examples (e.g. VII.3) where we need to enlarge the probability space and introduce randomization before the regenerative properties of $\{X_t\}$ can be recognized. It may, however, in some cases be convenient to have the S_n given as stopping times for $\{X_t\}$. This is no restriction since one can just enlarge the state space to $E \times (0, \infty)$ and consider $\{\widetilde{X}_t\} = \{(X_t, B_t)\}$ with B_t the forward recurrence time for $\{S_n\}$ at time t (clearly, $\{\widetilde{X}_t\}$ is regenerative with the same imbedded renewal process).

To a given delayed regenerative process, there clearly corresponds a zero–delayed one with a unique probability law (e.g. $\{X_{S_0+t}\}_{t \in \mathbb{T}}$). We let $\mathbb{P}_0, \mathbb{E}_0$ correspond to the zero–delayed case and then write $Y = Y_1$ for the length of the first cycle, $\mu = \mathbb{E}_0 Y$.

A trivial but noteworty property is that the regenerative property is preserved under mappings (nothing like that is true for say a Markov process):

Proposition 1.1 *If* $\{X_t\}_{t \in \mathbb{T}}$ *is regenerative and* $\varphi : E \to F$ *any measurable mapping, then* $\{\varphi(X_t)\}_{t \in \mathbb{T}}$ *is regenerative with the same embedded renewal process.*

The power of the concept of regenerative processes lies in the existence of a limiting distribution under conditions that are very mild and usually easy to verify. For example, in continuous time it is only required that $\mu < \infty$, that the cycle length distribution is nonlattice and that the sample paths satisfy some conditions that are automatic in any concrete example:

Theorem 1.2 *Assume that a* (possibly delayed) *regenerative process* $\{X_t\}_{t\in\mathbb{T}}$ *has metric state space, right–continuous paths and nonlattice cycle length distribution F with finite mean μ. Then the limiting distribution, say $\mathbb{P}_e$, of X_t exists and is given by*

$$\mathbb{E}_e f(X_t) \;=\; \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y f(X_s)\,\mathrm{d}s\,. \tag{1.1}$$

Proof. It is immediately checked that

$$A \;\to\; \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(X_s\in A)\,\mathrm{d}s$$

defines a probability measure on the Borel σ–algebra on E, and hence by standard facts on weak convergence it is sufficient to prove that $\mathbb{E}f(X_t) \to \mathbb{E}_e f(X_t)$ whenever f is continuous with $0 \le f \le 1$. Letting $Z(t) = \mathbb{E}_0 f(X_t)$, $z(t) = \mathbb{E}_0[f(X_t); t < Y]$, $F_0^*(x) = \mathbb{P}(Y_0 \le x)$, it follows by the usual renewal argument that

$$\mathbb{E}f(X_t) \;=\; \mathbb{E}[f(X_t); t<Y_0] + \int_0^t Z(t-x)\,F_0^*(\mathrm{d}x), \tag{1.2}$$

$$Z(t) \;=\; z(t) + \int_0^t Z(t-x)\,F(\mathrm{d}x). \tag{1.3}$$

Hence letting $t\to\infty$ in (1.2) shows that it is sufficient to show

$$Z(t) \;\to\; \mathbb{E}_e f(X_t) \;=\; \frac{1}{\mu}\int_0^\infty \mathbb{E}_0[f(X_t); t<Y]\,\mathrm{d}s \;=\; \frac{1}{\mu}\int_0^\infty z(s)\,\mathrm{d}s\,,$$

i.e. according to the key renewal theorem to show that z is d.R.i. But z is right–continuous, hence continuous a.e. by A2.1. Also $z(t) \le z^*(t) = \mathbb{P}_0(Y > t) = \overline{F}(t)$ where z^* is d.R.i. by V.4.1(v). Part (iv) of V.4.1 completes the proof. □

The basic renewal argument in the proof may be given in various ways. For example, the following representation is often useful:

Proposition 1.3 *Let* $\{X_t\}_{t\in\mathbb{T}}$ *be regenerative and* $\{A_t\}_{t\ge Y_0}$ *the backward recurrence time process of the imbedded renewal process. Further let $f : E\to\mathbb{R}$ be measurable and bounded, and define $g(t) = \mathbb{E}_0[f(X_t)\,|\,Y>t]$. Then*

$$\mathbb{E}f(X_t) \;=\; \mathbb{E}[g(A_t);\, Y_0\le t] \,+\, \mathbb{E}[f(X_t);\, Y_0>t]\,. \tag{1.4}$$

In particular, in the zero–delayed case $\mathbb{E}_0 f(X_t) = \mathbb{E}_0 g(A_t)$.

Proof. Conditioning upon Y_0 shows that it is sufficient to consider the zero–delayed case. Define $Z(t), z(t)$ as above and let $Z_1(t) = \mathbb{E}_0 g(A_t)$, $z_1(t) =$

$\mathbb{E}_0[g(A_t);\, Y > t]$. Then $Z = U * z$, $Z_1 = U * z_1$ and the desired conclusion follows since

$$\mathbb{E}_0[g(A_t);\, Y > t] \;=\; g(t)\mathbb{P}(Y > t) \;=\; \mathbb{E}_0[f(X_t);\, Y > t]$$

implies that $z_1 = z$. □

Proposition 1.3 yields an alternative proof of the limit result of Theorem 1.2, see Problem 1.3, and we also note the following strengthening (for total variation convergence, see A8, and recall that F_0 is the distribution with density $\overline{F}(x)/\mu$):

Corollary 1.4 *If A_t converges to F_0 in t.v., then also a t.v. limit of X_t exists and is given by* (1.1) *for $\mathbb{T} = [0, \infty)$, whereas for $\mathbb{T} = \mathbb{N}$*

$$\mathbb{E}_e f(X_t) \;=\; \frac{1}{\mu}\,\mathbb{E}_0 \sum_{k=0}^{Y-1} f(X_k) \;=\; \frac{1}{\mu}\,\mathbb{E}_0 \sum_{k=1}^{Y} f(X_k)\,. \qquad (1.5)$$

Proof. We must show that $\mathbb{E}f(X_t)$ converges to the asserted limit uniformly in the bounded measurable f with $\|f\|_\infty \le 1$. But since a uniform bound for the last term in (1.4) is $\mathbb{P}(Y_0 > t)$, the uniformity is immediate from the t.v. convergence of A_t and (1.4). Also for $\mathbb{T} = [0, \infty)$ (the case $\mathbb{T} = \mathbb{N}$ is entirely similar), the limit is given by

$$\begin{aligned}\mathbb{E}_e f(X_t) \;&=\; \int_0^\infty g(t)\, F_0(\mathrm{d}t) \;=\; \frac{1}{\mu}\int_0^\infty \mathbb{E}_0\big[f(X_t)\,\big|\, Y > t\big]\mathbb{P}_0(Y > t)\,\mathrm{d}t \\ &=\; \frac{1}{\mu}\mathbb{E}_0 \int_0^\infty f(X_t) I(Y > t)\,\mathrm{d}t \;=\; \frac{1}{\mu}\mathbb{E}_0 \int_0^Y f(X_t)\,\mathrm{d}t\,.\end{aligned}$$

□

Corollary 1.5 *Let $\{X_t\}_{t\in\mathbb{N}}$ be regenerative in discrete time with $\mu = \mathbb{E}_0 Y < \infty$, and let d be the period of the distribution F of the cycle length Y. Then:*
(i) *In the aperiodic case $d = 1$, a t.v. limit exists and is given by* (1.5)*.*
(ii) *If $d > 1$, then*

$$\frac{1}{d}\sum_{j=0}^{d-1} \mathbb{E}f(X_{nd+j}) \;\to\; \frac{1}{\mu}\,\mathbb{E}_0 \sum_{k=0}^{Y-1} f(X_k)\,. \qquad (1.6)$$

Proof. The process $\{A_n\}$ is Markov and if $\mu < \infty$, $d = 1$, it follows from I.2–4 that $A_n \to F_0$ weakly, hence also (since the state space is discrete) in t.v. Thus (i) follow from Corollary 1.4, whereas (ii) is a similar application of Proposition 1.3 and I.(4.2) (or Theorem 3.1 below). □

We return to t.v. convergence for $\mathbb{T} = [0, \infty)$ in VII.1–2.

Problems

1.1 Let $\{A_t\}$ be the backward recurrence time process of a renewal process with interarrival distribution F with $\mu < \infty$ and let $X_t = I(A_t \in \mathbb{Q})$. Show that $\{X_t\}$ is regenerative but that X_t needs not converge in distribution if say F is concentrated on $\mathbb{Q}$.

1.2 Show by an example that $\mu < \infty$ is not necessary for convergence in distribution of a regenerative process.

1.3 In Proposition 1.3, show that g is continuous a.e. provided f is continuous and the paths right–continuous. Give hereby an alternative proof of the limit result of Theorem 1.2.

1.4 Let $\{X_t\}_{t\in\mathbb{T}}$ be regenerative and satisfy the conditions for existence of a limit distribution π, and let $f : E \to \infty$ be π–integrable. Show that $\mathbb{E}f(X_t) \to \int f(x)\,\pi(\mathrm{d}x)$ holds always when $\mathbb{T} = \mathbb{N}$ but not always when $\mathbb{T} = [0, \infty)$. [*Hint:* Backward recurrence times, f chosen such that $z(x) = f(x)\overline{F}(x)$ is Lebesgue integrable with $\overline{\lim}\ z(x) = \infty$.]

Notes The pioneering paper on regenerative processes is Smith (1955). A main recent monograph treating the subject in depth and giving further history and many references is Thorisson (2000); note, however, that the flavour is somewhat different from the present book by emphasizing structure rather than asymptotic theory and applications.

Of concepts related to regenerative processes, we mention in particular regenerative phenomena (Kingman, 1972), regenerative sets (level sets of Markov processes or, equivalently, ranges of subordinators; Fristedt, 1996), renovating events (Borovkov, 1984) and Palm theory, which we study in more depth in VII.6.

2 First Examples and Applications

2a. Renewal Processes
2b. Alternating Renewal Processes
2c. Reflected Brownian Motion
2d. Regenerative Simulation
2e. Functionals of Regenerative Processes

Examples and applications of regenerative processes to queues and related models will abound in Part C, so here we shall only consider a few topics of a somewhat different flavour.

2a Renewal Processes

Consider a renewal process with nonlattice interarrival distribution F. If $\mu < \infty$, the stationary limiting distributions of the recurrence times A_t, B_t and of the current life $C_t = A_t + B_t$ have been found in V.3. Their particular form comes from the basic formula (1.1) as follows. For $0 \le t < Y$ we have

$A_t = t$, $B_t = Y - t$, $C_t = Y$. In particular,

$$\begin{aligned}
\mathbb{P}_e(A_t \le \xi) &= \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(A_t \le \xi)\,\mathrm{d}t \;=\; \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(t \le \xi)\,\mathrm{d}t \\
&= \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(Y - t \le \xi)\,\mathrm{d}t \;=\; \mathbb{P}_e(B_t \le \xi)
\end{aligned}$$

and the common value is

$$\begin{aligned}
\frac{1}{\mu}\,\mathbb{E}_0\int_0^\infty I(t \le \xi, t < Y)\,\mathrm{d}t &= \frac{1}{\mu}\int_0^\xi \mathbb{P}_0(t < Y)\,\mathrm{d}t \\
&= \frac{1}{\mu}\int_0^\xi \overline{F}(t)\,\mathrm{d}t \;=\; F_0(\xi)
\end{aligned}$$

(from results of VII.3 it will also follow that the density $\overline{F}(x)$ is stationary for $\{A_t\}$, $\{B_t\}$ even if $\mu = \infty$). Finally

$$\begin{aligned}
\mathbb{P}_e(C_t \le \xi) &= \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(C_t \le \xi)\,\mathrm{d}t \;=\; \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y I(Y \le \xi)\,\mathrm{d}t \\
&= \frac{1}{\mu}\,\mathbb{E}_0[Y;\, Y \le \xi] \;=\; \frac{1}{\mu}\int_0^\xi x\,F(\mathrm{d}x).
\end{aligned}$$

Problems

2.1 Show by similar arguments that the relative position A_t/C_t of the current item has a limiting uniform distribution; cf. V.3.3.

2b Alternating Renewal Processes

A point process on $[0, \infty)$ with first epoch at Y_0 and interarrival times $Y_1, Y_2, \ldots$ is called an *alternating renewal process* if all $Y_0, Y_1, \ldots$ are independent with distributions (say) G_0 of $Y_0, Y_2, \ldots$ and G_1 of $Y_1, Y_3, \ldots$. Such processes arise, for example, in reliability theory where Y_{2k-1} could be the lifetime of the kth item and Y_{2k} the time needed to replace it. Here one might ask, for example, for the probability $p(t)$ that the system is operating at time t, for the distribution of the remaining lifetime of the current item and so on. These quantities are easily obtained by observing that the system regenerates at every second renewal. For example, for $p(t)$ we can define $X_t \in \{0, 1\}$ by

$$X_t = \begin{cases} 0 & \text{if } Y_0 + \cdots + Y_{2k-1} \le t < Y_0 + \cdots + Y_{2k} \text{ for some } k \\ 1 & \text{if } Y_0 + \cdots + Y_{2k} \le t < Y_0 + \cdots + Y_{2k+1} \text{ for some } k \end{cases}$$

Then $p(t) = \mathbb{P}(X_t = 1)$ and the $Y_0 + \cdots + Y_{2k}$ are regeneration points for $\{X_t\}$. The cycle length distribution is the distribution $F = G_0 * G_1$ of $Y_1 + Y_2$, and if F is nonlattice with $\mu = \mathbb{E}Y_0 + \mathbb{E}Y_1 < \infty$ and $\mathbb{E}_0$ refers to the case $Y_0 = 0$, we get

$$\lim_{t\to\infty} p(t) \;=\; \frac{1}{\mu}\mathbb{E}_0\int_0^{Y_1+Y_2} I(X_t = 1)\,\mathrm{d}t \;=\; \frac{1}{\mu}\mathbb{E}Y_1 \;=\; \frac{\mathbb{E}Y_1}{\mathbb{E}Y_1+\mathbb{E}Y_2}.$$

Further characteristics of the system are easily computed in just the same manner; see Problem 2.2.

Problems

2.2 Consider the system conditioned to be operating ($X_t = 1$)- Show that the past life, the residual life and the total life of the current item all have the same limit distribution as in a renewal process with interarrival distribution G_1.

2c Reflected Brownian Motion

Let $\{X_t\}$ be reflected Brownian motion with drift $\mu < 0$ and variance 1 and starting from $X_0 = 0$. In order to view the process as regenerative, one would try to take the cycles as the excursions away from 0, leading to $Y = \inf\{t : X_t = 0\}$. However, the sample path structure of Brownian motion implies that the Y defined in this way is 0 and therefore useless (nevertheless, it makes sense to study the concept of a stationary excursion; for two different viewpoints, see Pitman, 1986, and Salminen and Norros, 2001).

Instead one may, e.g., take $Y = \inf\{t > \tau(1) : X_t = 0 \,|\, X_0 = 0\}$ where $\tau(1) = \inf\bigl\{t > 0 : X_t = 1 \,\big|\, X_0 = 0\bigr\}$ ("up to 1 and back to 0"). We have

$$\mu \;=\; \mathbb{E}_0\tau(1) + \mathbb{E}_0\bigl(Y - \tau(1)\bigr) \;=\; \frac{\mathrm{e}^{-2\mu} + 2\mu - 1}{2\mu^2} - \frac{1}{\mu} \;=\; \frac{\mathrm{e}^{-2\mu} - 1}{2\mu^2},$$

where the expression for $\mathbb{E}_0\tau(1)$ is shown in IX.3.8 and the one for $\mathbb{E}_0(Y - \tau(1))$ is just Wald's identity.

2d Regenerative Simulation

As explained in III.1, many practical situations call for numerical values of a parameter of the form $\theta = \mathbb{E}_e f(X_t)$. For example, $\{X_t\}$ could be a queue length process so that $f(x) = x$ would correspond to θ being the mean queue length in the steady state, $f(x) = I(x \geq N)$ to θ being the probability of queue length at least N in steady state and so on (similar remarks apply to waiting–time processes in discrete time). Now Theorem 1.2 states that θ is indeed welldefined, but to use formula (1.1) to express θ in terms of the interarrival and service time distributions may be difficult or impossible. Hence an alternative method could be required, and here we shall look at simulation.

The standard simulation (Monte Carlo) technique for estimating θ would be to design a simulation experiment giving as outcome a response variable

R having the $\mathbb{P}_e$–distribution of $f(X_t)$. One then would perform N replications giving i.i.d. variables $R_1, \ldots, R_N$ distributed as R, estimate θ by the empirical mean $\widehat{\theta} = \overline{R}_N = (R_1 + \cdots + R_N)/N$ and give the uncertainty on θ say in the form of asymptotic 95% confidence intervals $\widehat{\theta} \pm 1.96s/\sqrt{N}$, where

$$s^2 \;=\; \frac{1}{N-1}\sum_{n=1}^{N}(R_n - \overline{R}_N)^2$$

is the empirical variance. This method is not feasible here since we may well simulate the queue starting from any given set of initial conditions but usually not in the unknown steady state. A partial solution would be to simulate the queue in $[0, T]$ starting from, say, an empty queue and choose T so large that hopefully $R = f(X_T)$ would have a distribution close to the required steady–state distribution. However, with T large each replication of the experiment becomes timeconsuming and one is faced with the uncertainty inherent in the choice of T.

Instead we focus on the basic formula (1.1) and estimate the unknown $\mu = \mathbb{E}_0 Y$, $\nu = \mathbb{E}_0 \int_0^Y f(X_t)\,\mathrm{d}t$ by simulation of a regenerative cycle. That is, the simulation experiment consists in running one cycle and observing a two–dimensional response (column) vector $\boldsymbol{R} = \bigl(R(1)\ R(2)\bigr)^{\mathsf{T}}$ given by $R(1) = Y$, $R(2) = \int_0^Y f(X_t)\,\mathrm{d}t$. We then create i.i.d. replications $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ and estimate μ, ν and $\theta = \nu/\mu$ by

$$\widehat{\mu} \;=\; \frac{1}{N}\sum_{n=1}^{N} R_n(1), \quad \widehat{\nu} \;=\; \frac{1}{N}\sum_{n=1}^{N} R_n(2), \quad \widehat{\theta} \;=\; \frac{\widehat{\nu}}{\widehat{\mu}}.$$

By the LLN, $\widehat{\mu}$ and $\widehat{\nu}$ are strongly consistent for μ, ν ($\widehat{\mu} \overset{\text{a.s.}}{\to} \mu$, $\widehat{\nu} \overset{\text{a.s.}}{\to} \nu$ as $N \to \infty$) and hence $\widehat{\theta}$ is so for θ. Confidence intervals can also be obtained assuming i.i.d. cycles. To this end, let

$$\begin{aligned} \boldsymbol{\Sigma} \;&=\; \mathbb{V}ar_0\boldsymbol{R} \;=\; \mathbb{E}[\boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}] - \mathbb{E}\boldsymbol{R}[\mathbb{E}\boldsymbol{R}]^{\mathsf{T}} \\ &=\; \begin{pmatrix} \mathbb{V}ar_0 R(1) & \mathbb{C}ov_0\bigl(R(1), R(2)\bigr) \\ \mathbb{C}ov_0\bigl(R(1), R(2)\bigr) & \mathbb{V}ar_0 R(2) \end{pmatrix}. \end{aligned}$$

Then $(\widehat{\mu}\ \widehat{\nu})^{\mathsf{T}} = \overline{\boldsymbol{R}}_N$ is two–dimensional asymptotically normal with mean $(\widehat{\mu}\ \widehat{\nu})^{\mathsf{T}}$ and covariance matrix $\boldsymbol{\Sigma}/N$. Letting $\varphi(x, y) = y/x$, it thus follows by a standard transformation result that $\widehat{\theta} = \varphi(\widehat{\mu}, \widehat{\nu})$ is asymptotically normal with mean $\varphi(\mu, \nu) = \theta$ and variance σ^2/N, where

$$\sigma^2 \;=\; [\nabla\varphi]^{\mathsf{T}}\boldsymbol{\Sigma}\nabla\varphi \quad \text{with } \nabla\varphi \;=\; \begin{pmatrix} \partial\varphi/\partial x \\ \partial\varphi/\partial y \end{pmatrix} \;=\; \begin{pmatrix} -y/x^2 \\ 1/x \end{pmatrix}$$

(the gradient) evaluated at $x = \mu, y = \nu$. Now the empirical covariance matrix $\boldsymbol{S}$ with elements

$$s_{ij} \;=\; \frac{1}{N-1}\sum_{n=1}^{N}\bigl(R_n(i) - \overline{R}_N(i)\bigr)\bigl(R_n(j) - \overline{R}_N(j)\bigr), \quad i, j = 1, 2,$$

is strongly consistent for $\boldsymbol{\Sigma}$ and $(-\widehat{\nu}/\widehat{\mu}^2\ 1/\widehat{\mu})^{\mathsf{T}}$ so for $\Delta\varphi$. Hence

$$s^2 \;=\; \frac{\widehat{\nu}^2}{\widehat{\mu}^4}s_{11} + \frac{1}{\widehat{\mu}^2}s_{22} - 2\frac{\widehat{\nu}}{\widehat{\mu}^3}s_{12}$$

is strongly consistent for σ^2 and $\widehat{\theta} \pm 1.96s/\sqrt{N}$ is an asymptotic 95% confidence interval for θ.

Problems

2.3 The $M/M/1$ queue length process $\{X_t\}$ with $\beta = 70$, $\delta = 100$ was simulated in 11 busy cycles and the values

$R(1)$	0.1494	0.0320	0.0124	0.0114	0.0212	0.0271
$R(2)$	0.5023	0.0104	0.0036	0.0019	0.0046	0.0169
$R(1)$	0.0142	0.0145	0.0243	0.0122	0.1175	
$R(2)$	0.0103	0.0003	0.0094	0.0001	0.2332	

of the cycle length $R(1) = Y$ and $R(2) = \int_0^Y X_t\,\mathrm{d}t$ were recorded, giving

$$\sum_{n=1}^{11} R_n(i) \;=\; \left\{\begin{array}{ll} 0.4363 & i=1 \\ 0.7930 & i=2 \end{array}\right., \quad \left(\sum_{n=1}^{11} R_n(i)R_n(j)\right) \;=\; \begin{pmatrix} -0.0398 & 0.1038 \\ 0.1038 & 0.3073 \end{pmatrix}$$

Check whether the deviation of the corresponding estimate for $\mathbb{E}_e X_t = \rho/(1-\rho)$ is within the statistical uncertainty.

2.4 Show that the bias $\mathbb{E}\widehat{\theta} - \theta$ is of order $1/N$.

Notes For regenerative simulation, see e.g. Rubinstein and Melamed (1998).

2e Functionals of Regenerative Processes

In a variety of contexts, one is interested in more general functionals of the paths of a regenerative process $\{X_t\}_{t\in\mathbb{T}}$ than just the value of a single X_t. For example, for $\mathbb{T} = \mathbb{N}$, it would be of interest to say something not only about X_n but also about the dependence between consecutive values (X_n, X_{n+1}). Other examples could be $\max_{k=0,\dots,N} X_{n+k}$, $\int_t^{t+h} X_s\,\mathrm{d}s$ and so on. For such cases, the classical independent cycle property does not carry over to the functionals. For instance, for the (X_n, X_{n+1}) example (X_Y, X_{Y+1}) and (X_{Y-1}, X_Y) belong to distinct cycles but may clearly be dependent. However, *the slightly weaker definition that we have given of a regenerative process also includes such cases* since we have required the post–S_n process to be independent only of S_n, not of the whole pre–S_n process.

A convenient formalism for expressing this is "lifting of the regenerative process to function space." Let $\{X_t\}$ be regenerative (not necessarily with i.i.d. cycles) with imbedded renewal process $\{S_n\}$. If $\mathbb{T} = [0,\infty)$, we assume in addition that the state space E is Polish and that $\{X_t\}$ has paths in $\widetilde{E} = D([0,\infty), E)$. For $\mathbb{T} = \mathbb{N}$ we let $\widetilde{E} = E^{\mathbb{N}}$. It is then an immediate consequence of the regenerativity of $\{X_t\}$ that the $\widetilde{E}$–valued process $\{\theta_t X\}$ defined by $\theta_t X = \{X_{t+s}\}_{s\in\mathbb{T}}$ is again regenerative with the same imbedded renewal process $\{S_n\}$.

Theorem 2.1 *If* $\{X_t\}$ *satisfies the condition for existence of a limit of* X_t, *then also* $\theta_t X$ *has a limit* $X^{(e)} = \{X^{(e)}_t\}_{t\in\mathbb{T}}$ (in total variation for $\mathbb{T}=\mathbb{N}$ and weakly for $\mathbb{T}=[0,\infty)$) *given by*

$$\mathbb{E}_e\varphi(X^{(e)}) \;=\; \frac{1}{\mu}\,\mathbb{E}_0\int_0^Y \varphi(\theta_t X)\,\mathrm{d}t \;=\; \mathbb{E}_0\left[\frac{Y}{\mu}\varphi\left(\theta_{UY}X\right)\right] \tag{2.1}$$

($\sum_0^{Y-1}$ if $\mathbb{T}=\mathbb{N}$) *for any nonnegative or bounded* $\varphi:\widetilde{E}\to\mathbb{R}$ *where* U *is uniform on* $(0,1)$ *and independent of the regenerative process. Further,* $\{X^{(e)}_t\}_{t\in\mathbb{T}}$ *is strictly stationary.*

Proof. For $\mathbb{T}=[0,\infty)$ it is easily checked that the paths of $\{\theta_t X\}$ are right–continuous (they are not in $D([0,\infty),\widetilde{E})$, however, since $\lim_{s\uparrow t}\theta_s x$ will fail to exist in this space if x is a D–function with a jump at t). Hence the existence of a limit and the first expression for $\mathbb{E}\varphi(X^{(e)})$ in (2.1) follows immediately; that this is the same as the second follows since

$$\mathbb{E}_0\big[Y\varphi\left(\theta_{UY}X\right)\big] \;=\; \mathbb{E}_0\int_0^1 \varphi(\theta_{uY}X)Y\,\mathrm{d}u \;=\; \mathbb{E}_0\int_0^Y \varphi(\theta_t X)\,\mathrm{d}t.$$

Stationarity follows follows from

$$\begin{aligned}\theta_u X^{(e)} &= \theta_u\left(\lim_{t\to\infty}\theta_t X^{(e)}\right) \;=\; \lim_{t\to\infty}\theta_u\theta_t X^{(e)}\\ &= \lim_{t\to\infty}\theta_{u+t}X^{(e)} \;=\; \lim_{t\to\infty}\theta_t X^{(e)} \;=\; X^{(e)}.\end{aligned}$$

□

In view of Theorem 2.1, $X^{(e)}$ represents a strictly stationary version of the given regenerative process. Note the peculiarity of the process $X^{(e)}$ that it is deterministic given its initial value $X^{(e)}_0$: for any t, $X^{(e)}_t$ is a function of $X^{(e)}_0$. Note also that for $\mathbb{T}=[0,\infty)$ we obtain convergence of $\theta_t X$ in function space without as usual having to invoke tightness. Finally note that the final expression in (2.1) gives a similar description of a stationary regenerative process as the one for a renewal process given in V.3.3: the stationary version is obtained from a zero–delayed version by first length–biasing (using the length of the first cycle Y as likelihood ratio) and next shifting $t=0$ to a uniformly chosen point in the cycle.

3 Time–Average Properties

We shall state and prove the results only in continuous time, the modifications in discrete time being obvious.

A real–valued process $\{Z_t\}$ is called *cumulative* if $Z_0 = 0$ and there exists a renewal process $\{S_n\}$ such that for any n $\{Z_{S_n+t} - Z_{S_n}\}_{t\geq 0}$ is independent of $S_0, \ldots, S_n$ and $\{Z_t\}_{t<S_n}$, and the distribution is independent of n. An obvious example is $Z_t = \int_0^t f(X_s)\,\mathrm{d}s$ where $\{X_t\}$ is regenerative with i.i.d. cycles, but there are others (with possible jumps) like the Markov additive processes studied in XI.2.

The basic tool in the study is to write

$$Z_t \;=\; U_0(t) + U_1 + \cdots + U_{N_t-1} + \Delta_t \tag{3.1}$$

where $U_0(t) = Z_{t\wedge S_0}$, $U_n = Z_{S_n} - Z_{S_{n-1}}$, $\Delta_t = Z_t - Z_{S_{N_t-1}}$. Here $U_0(t) = Z_{S_0}$ eventually and becomes negligible in the limit $t \to \infty$, $U_1 + \cdots + U_{N_t-1}$ is a random sum of i.i.d. summands and can be handled by standard tools, and the only problem turns out to be to bound Δ_t. To this end, define $V = \max_{0\leq t<Y} |Z_t|$,

$$V_n \;=\; \max_{S_{n-1}\leq t<S_n} \left|Z_t - Z_{S_{n-1}}\right| \;=\; \max_{S_{n-1}\leq t<S_n} |\Delta_t|\,. \tag{3.2}$$

Then $V_1, V_2, \ldots$ are i.i.d. with $V_n \overset{\mathscr{D}}{=} V$. We will assume throughout that $V < \infty$ a.s. and that $\mu = \mathbb{E}_0 Y < \infty$ (as usual, Y is the generic cycle and $\mathbb{E}_0$ refers to the case $S_0 = 0$; we then write $U = U_1$).

We start by a LLN which contains as a special case results stated in I.4 and II.4.

Theorem 3.1 *Suppose $\mu = \mathbb{E}_0 Y < \infty$, $\mathbb{E}_0|U| < \infty$ and let $\overline{z} = \mathbb{E}_0 U/\mu$. Then $Z_t/t \overset{\text{a.s.}}{\to} \overline{z}$ if and only if $\mathbb{E}_0 V < \infty$.*

Note that in the regenerative example $Z_t = \int_0^t f(X_s)\mathrm{d}s$, we may write the limit $\mathbb{E}U_1/\mu$ as $\mathbb{E}_e f(X_t)$.

Proof. By the standard LLN and the elementary renewal theorem,

$$\frac{U_1 + \cdots + U_{N_t-1}}{t} \;\sim\; \frac{N_t - 1}{t}\mathbb{E}_0 U \sim\; \frac{1}{\mu}\mathbb{E}_0 U \;=\; \overline{z} \text{ a.s.}$$

Also obviously $U_0(t)/t \overset{\text{a.s.}}{\to} 0$ a.s. and hence the asserted convergence of Z_t/t holds if and only if $\Delta_t \to 0$. But $t \sim n\mu$ when n, t are connected by $S_{n-1} \leq t < S_n$ and hence by (3.1), $\Delta_t \to 0$ is equivalent to $V_n/n \overset{\text{a.s.}}{\to} 0$ which in turn (Borel–Cantelli!) is well known to hold if and only if $\mathbb{E}_0 V < \infty$. □

Note that in some applications, it is more convenient to identify the form of the limiting distribution by means of the LLN 3.1 than to use the formula (1.1). In particular, this may be the case when a discrete–time process is imbedded in a continuous–time one, and one wishes to relate the limiting distributions (an example is PASTA; see VII.6).

We next prove a CLT analogue:

Theorem 3.2 *Assume $\mathbb{V}ar_0 U < \infty$, $\mathbb{V}ar_0 Y < \infty$. Then the limiting distribution of $(Z_t - t\overline{z})/\sqrt{t}$ exists and is normal with mean zero and variance*

σ^2/μ, *where*

$$\sigma^2 \;=\; \mathbb{V}ar_0(U - \overline{z}Y) \;=\; \mathbb{V}ar_0 U + \overline{z}^2\,\mathbb{V}ar_0 Y - 2\overline{z}\,\mathbb{C}ov_0(U,Y).$$

Proof. We again use (3.1) where clearly $U_0(t)/\sqrt{t} \xrightarrow{\mathbb{P}} 0$. Also $\Delta_t/\sqrt{t} \xrightarrow{\mathbb{P}} 0$ is clear as above from $V_n/\sqrt{n} \xrightarrow{\mathbb{P}} 0$ (moments of V are not needed for this). Thus it remains to prove that

$$\frac{1}{\sqrt{t}}\,(U_1 + \cdots + U_{N_t-1} - t\overline{z}) \tag{3.3}$$

has the desired limit distribution. But letting

$$T(n) \;=\; U_1 + \cdots + U_n - (Y_1 + \cdots + Y_n)\overline{z},$$

we may write (3.3) as

$$\frac{T(N_t-1)}{\sqrt{t}} \;+\; \frac{1}{\sqrt{t}}(Y_1 + \cdots + Y_{N_t-1} - t)\overline{z}. \tag{3.4}$$

Now clearly $T(n)/\sqrt{n}$ is asymptotically normal with mean zero and variance σ^2. Since $(N_t-1)/t \xrightarrow{\mathbb{P}} \mu^{-1}$, it thus follows by Anscombe's theorem (Chung, 1974, p. 216) that the first term of (3.4) has the desired limit distribution, and it remains only to check that the second term vanishes in the limit. But

$$t - Y_1 - \cdots - Y_{N_t-1} \;=\; t - S_{N_t-1} + Y_0 \;=\; A_t + Y_0.$$

Since $\{A_t\}$ is always tight when $\mu < \infty$ (in fact convergent in distribution in the nonlattice case), we thus always have $A_t/\sqrt{t} \xrightarrow{\mathbb{P}} 0$ and since clearly $Y_0/\sqrt{t} \xrightarrow{\mathbb{P}} 0$, the proof is complete. □

Notes Beyond the (simpler!) regenerative setting, the traditional approach to the CLT in the presence of dependence is to assume stationarity and some mixing condition. The limiting variance constant for the time–average of $\{X_t\}_{t\in\mathbb{T}}$ then comes out as

$$2\int_0^\infty \mathbb{C}ov(X_0, X_t)\,\mathrm{d}t\,, \quad \mathbb{V}ar X_0 \;+\; 2\sum_{n=1}^{\infty}\mathbb{C}ov(X_0, X_n) \tag{3.5}$$

for $\mathbb{T} = [0,\infty)$, resp. $\mathbb{T} = \mathbb{N}$ (assuming, of course, that (3.5) is finite). See e.g. Durrett (1991), pp. 381, 384. It should be noted that the evaluation of the covariances in (3.5) is most often cumbersome also in the regenerative setting. For further expressions for variance constants, see I.7, II.4d and XI.2.8.

A sharp version of the CLT for regenerative processes is in Glynn and Whitt (1993).

4 Rare Events and Extreme Values

We consider a regenerative process $\{X_t\}_{t\in\mathbb{T}}$ with independent cycles and finite cycle mean μ. The aim is to obtain information on the first time certain rare events occur. For a typical example, assume that the state space is $E = [0, \infty)$ and that the rare event is exceedance of level x where x is large. The hitting time of the rare event is then

$$\tau(x) \; = \; \inf \{t > 0 : \, X_t > x\} \, . \tag{4.1}$$

Letting $\overline{X}_T = \max_{0 \le t \le T} X_t$, we have

$$\mathbb{P}(\tau(x) > T) \; = \; \mathbb{P}(\overline{X}_T \le x). \tag{4.2}$$

Therefore, once a limit theorem for $\tau(x)$ as $x \to \infty$ has been established, this can be translated into the classical goal of extreme value theory, viz. a limit theorem for $\overline{X}_T$ as $T \to \infty$.

The key feature of the regenerative setting is that the discussion of issues like these can be reduced to the study of the behaviour within a regenerative cycle. Generalizing the set–up somewhat, let $\{A(x)\}_{x \ge 0}$ be a family of cycle events, that is, measurable sets in $D_0(E)$ indexed by a parameter $x \ge 0$ and having the property that $A(x) \downarrow \emptyset$, $x \to \infty$; note that this implies

$$a(x) \; = \; \mathbb{P}_0\bigl(\{X_t\}_{0 \le t < Y} \in A(x)\bigr) \; \to \; 0, \quad x \to \infty.$$

Let further

$$M(x) \; = \; \inf \Bigl\{ n = 0, 1, \ldots : \, \bigl\{X_{t + Y_0 + \cdots + Y_{n-1}}\bigr\}_{0 \le t < Y_n} \in A(x) \Bigr\}$$

be the index of the first cycle in which $A(x)$ occurs and

$$\underline{\omega}(x) = Y_0 + Y_1 + \cdots + Y_{M(x)-1}, \quad \overline{\omega}(x) = Y_0 + Y_1 + \cdots + Y_{M(x)}.$$

Then, in intuitive terms, the rare event occurs for the first time somewhere between $\underline{\omega}(x)$ and $\overline{\omega}(x)$. We will see that both are approximately exponentially distributed with mean $\mu(x)$ where

$$\mu(x) \; = \; \mu \mathbb{E}_0 M(x) \; = \; \frac{\mu}{a(x)}$$

(the last identity follows since $M(x)$ is geometric w.r.t. $\mathbb{P}_0$ on $\{1, 2, \ldots\}$ with success parameter $a(x)$, i.e. $\mathbb{P}_0(M(x) = n) = (1 - a(x))^{n-1} a(x)$, $n = 1, 2, \ldots$). Consider first the means:

Proposition 4.1 *For all x, $\mathbb{E}_0 \overline{\omega}(x) = \mu(x)$. Further, for a delayed process with $\mathbb{E} Y_0 < \infty$, $a(x) \mathbb{E} \underline{\omega}(x)$ and $a(x) \mathbb{E} \overline{\omega}(x)$ both converge to μ as $x \to \infty$.*

Proof. The first statement follows immediately from Wald's identity. Since

$$a(x) \mathbb{E}_0 Y_{M(x)} \; = \; a(x) \mathbb{E}_0 [Y \,|\, A(x)] \; = \; \mathbb{E}_0 [Y ; \, A(x)] \; \to \; 0 \tag{4.3}$$

by monotone convergence, the second statement is then also clear in the zero–delayed case. In the delayed case, just appeal to the bounds

$$\mathbb{E} \underline{\omega}(x) \; \ge \; \mathbb{P}\bigl(\{X_t\}_{0 \le t < Y_0} \notin A(x)\bigr) \mathbb{E}_0 \underline{\omega}(x) \; \sim \; \mathbb{E}_0 \underline{\omega}(x),$$

$$\mathbb{E} \overline{\omega}(x) \; \le \; \mathbb{E} Y_0 + \mathbb{E}_0 \overline{\omega}(x). \qquad \square$$

Next consider convergence in distribution:

Theorem 4.2 *As $x \to \infty$, it holds that $a(x)\underline{\omega}(x) \xrightarrow{\mathscr{D}} \mu V$, $a(x)\overline{\omega}(x) \xrightarrow{\mathscr{D}} \mu V$, where V is standard exponential. Further, $a(x)[\overline{\omega}(x) - \underline{\omega}(x)] \xrightarrow{\mathbb{P}} 0$.*

Proof. It is straightforward from $a(x) \to 0$ that $a(x)M(x) \xrightarrow{\mathscr{D}} V$:

$$\mathbb{P}\bigl(a(x)M(x) \geq y\bigr) \;=\; \mathbb{P}\bigl(M(x) \geq \lfloor y/a(x) \rfloor\bigr) \;=\; \bigl(1 - a(x)\bigr)^{\lfloor y/a(x) \rfloor} \;\to\; \mathrm{e}^{-y}.$$

It then follows from $(Y_0 + Y_1 + \cdots + Y_n)/n \xrightarrow{\text{a.s.}} \mu$ a.s. that

$$a(x)\underline{\omega}(x) \;=\; a(x)M(x) \cdot \frac{Y_0 + Y_1 + \cdots + Y_{M(x)-1}}{M(x)} \;\xrightarrow{\mathscr{D}}\; V\mu.$$

The last statement follows from (4.3) and implies then the stated asymptotics of $\overline{\omega}(x)$. □

In practice, one is seldom given a family of rare events $\{A(x)\}_{x \geq 0}$ but rather a single cycle event A such that $a = \mathbb{P}(A)$ is small. The implication of the above results is then to use $\mu V/a$ as an approximation for the first occurence time of A, and μ/a as an approximation for the mean. In specific cases, the evaluation of a may well be nontrivial.

Example 4.3 Let $\{X_t\}$ be the $M/M/1$ queue length process, β the arrival intensity and δ the service intensity. Assume $\rho = \beta/\delta < 1$ and that we want the asymptotics of the time $\tau(x) = \inf\{n > 0 : X_t = x\}$ of the first visit to x. We then take the start of cycles as the instances of returns to 0 and let $A(x) = \{X_t = x \text{ for some } t < Y\}$. To evaluate $a(x) = \mathbb{P}_0(A(x))$ for $x = 2, 3, \ldots$, we note that $a(x)$ is the probability that $\{S_t\}$, the difference between two independent Poisson processes with rates β, δ, hits x before 0 when started from $S_0 = 1$. The Wald martingale for $\{S_t\}$ (see III.8.8) is

$$\frac{z^{S_t}}{\mathbb{E}\bigl[z^{S_t} \bigm| S_0 = 0\bigr]} \;=\; \frac{z^{S_t}}{\exp\bigl\{t[\beta(z-1) + \delta(z^{-1} - 1)]\bigr\}}\,.$$

Taking $z = \rho^{-1}$, the denominator becomes 1. Optional stopping at time $\sigma = \inf\{t : S_t = 0 \text{ or } x \,|\, S_0 = 1\}$ yields

$$\begin{aligned}
\frac{1}{\rho} \;&=\; \mathbb{E}_1 \rho^{-S_0} \;=\; \mathbb{E}_1 \rho^{-S_\sigma} \;=\; \rho^{-x}\mathbb{P}_1(S_\sigma = x) + \rho^{-0}\mathbb{P}_1(S_\sigma = 0) \\
&=\; \mathbb{P}_1(S_\sigma = x)\bigl\{\rho^{-x} - 1\bigr\} + 1, \\
a(x) \;&=\; \mathbb{P}_1(S_\sigma = x) \;=\; \frac{(1-\rho)\rho^{x-1}}{1 - \rho^x} \;\sim\; (1-\rho)\rho^{x-1}.
\end{aligned}$$

Since $\underline{\omega}(x) \leq \tau(x) \leq \overline{\omega}(x)$, it follows that $a(x)\tau(x)$ is asymptotically exponential with mean $1/\beta(1-\rho)$ (the mean busy cycle), and that $a(x)\mathbb{E}\tau(x) \to 1/\beta(1-\rho)$. □

Example 4.4 Assume that $\{X_n\}$ is a positive recurrent Markov chain with state space $\{0, 1, 2, \ldots\}$ and $p_{0x} > 0$ for all $x = 1, 2, \ldots$. Assume we want the asymptotics of the time $\tau(0, x) = \inf\{n > 0 : X_{n-1} = 0, X_n = x\}$ of the first jump from 0 to x. We then take the start of cycles as the instances of visits to 0 and let $A(x) = \{X_{n-1} = 0, X_n = x \text{ for some } n = 0, \ldots, Y-1\}$. However, the choice of Y implies that $A(x) = \{X_0 = 0, X_1 = x\}$ and hence $a(x) = \mathbb{P}_0(A(x)) = p_{0x}$ whereas $\mu = \mathbb{E}_0 Y = 1/\pi_0$, where $(\pi_0\, \pi_1 \ldots)$ is the stationary distribution. Since $\underline{\omega}(x) \le \tau(0,x) \le \overline{\omega}(x)$, it follows that $p_{0x}\tau(0,x)$ is asymptotically exponential with mean $1/\pi_0$, and that $p_{0x}\mathbb{E}\tau(0,x) \to 1/\pi_0$. □

Example 4.5 Assume that $\{X_t\}$ is an ergodic birth–death process on $\{0,1,2,\ldots\}$ with birth rates β_n and death rates δ_n, and that we want the asymptotics of the time $\tau(x) = \inf\{t > 0 : X_t = x\}$ of the first visit to 0. We then take the start of cycles as the instances of return to 0 (then $\mu = 1/\pi_0\beta_0$) and let $A(x) = \{X_t = x \text{ for some } t < Y\}$. Since $\underline{\omega}(x) \le \tau(x) \le \overline{\omega}(x)$, it follows that $\mathbb{E}\tau(x) \sim \mu/a(x)$ and that $\tau(x)/\mathbb{E}\tau(x)$ is asymptotically standard exponential.

To compute $a(x)$, we let $\{Y_n\}$ denote the imbedded Markov chain stopped at the time σ when 0 or x is hit (then $a(x) = \mathbb{P}_1(Y_\sigma = x)$) and put $\{Y_t\}$ on its natural scale φ. That is, $\varphi(0), \ldots, \varphi(x)$ are such that $\bigl\{\varphi(Y_n)\bigr\}$ is a martingale. With $\Delta_n = \varphi(n) - \varphi(n-1)$ this means

$$0 \;=\; \mathbb{E}_n\varphi(Y_1) - \varphi(n) \;=\; -\frac{\delta_n}{\beta_n + \delta_n}\Delta_n + \frac{\beta_n}{\beta_n + \delta_n}\Delta_{n+1}, \quad n = 1, 2, \ldots, x-1.$$

Taking $\varphi(0) = 0$, $\varphi(1) = \Delta_0 = 1$, it follows that $\Delta_{n+1} = (\delta_n/\beta_n)\Delta_n$ and

$$\varphi(n) \;=\; 1 + \Delta_1 + \cdots + \Delta_n \;=\; 1 + \frac{\delta_1}{\beta_1} + \cdots + \frac{\delta_1 \cdots \delta_{n-1}}{\beta_1 \cdots \beta_{n-1}}.$$

Finally $1 = \varphi(1) = \mathbb{E}_1\varphi(Y_\sigma) = a(x)\varphi(x)$ so that $a(x) = 1/\varphi(x)$. For example, for $M/M/\infty$ with $\beta_n = \beta$, $\delta_n = n\delta$, $\eta = \beta/\delta$,

$$\mathbb{E}\tau(x) \;\sim\; \frac{\mu}{a(x)} \;=\; \frac{1 + \eta + 2\eta + \cdots + (x-1)!\eta^{x-1}}{\beta e^{-\eta}} \;\sim\; \frac{(x-1)!\eta^{x-1}}{\beta e^{-\eta}}.$$

□

The approximations above apply to time intervals of order $T(x) \approx a(x)^{-1}$. On a shorter time scale, we have the following result (stated for simplicity only for the zero–delayed case and for $A(x) = \bigl\{\tau(x) < Y\bigr\}$):

Theorem 4.6 *Let $a(t;x) = \mathbb{P}_0\left(\tau(x) \le t < Y\right)$ and assume that $T(x) \in \mathbb{T}$ varies with x in such a way that*

$$\lim_{x\to\infty} a(x)T(x) \;=\; 0, \qquad \lim_{x\to\infty} \frac{a(\epsilon T(x);x)}{a(x)} \;=\; 1 \tag{4.4}$$

for all $\epsilon > 0$. Then $\mathbb{P}_0\bigl(\tau(x) \le T(x)\bigr) \sim a(x)T(x)/\mu$.

Proof. Let U be the renewal measure of the regeneration points and

$$U(A;x) \;=\; \sum_{k=0}^{\infty} \mathbb{P}_0\bigl(Y_0 + \cdots + Y_k \in A,\; \tau(x) > Y_0 + \cdots + Y_k\bigr).$$

If $m(x)$ denotes the expected number of cycles before $T(x)$ (including the one straddling $T(x)$) where $A(x)$ occurs, then $m(x) = a(x)U\bigl(T(x)\bigr)$ is of order $a(x)T(x)/\mu$ so that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\bigl(\tau(x) \le T(x)\bigr)}{a(x)T(x)/\mu} \ \le\ \limsup_{x\to\infty} \frac{m(x)}{a(x)T(x)/\mu} \ =\ 1.$$

Conversely, if x is so large that $a(\epsilon T(x);x) \ge (1-\epsilon)a(x)$, then

$$\begin{aligned}
\mathbb{P}(\tau(x) \le T(x)) \ &=\ \int_0^{T(x)} a(T(x)-t;x)U(\mathrm{d}t;x) \\
\ge\ & \int_0^{(1-\epsilon)T(x)} a(T(x)-t;x)U(\mathrm{d}t;x) \\
\ge\ & \int_0^{(1-\epsilon)T(x)} a(\epsilon T(x);x)U(\mathrm{d}t;x) \ \ge\ (1-\epsilon)a(x)U\bigl((1-\epsilon)T(x);x\bigr) \\
\ge\ & (1-\epsilon)a(x)(1-a(x))U\bigl((1-\epsilon)T(x)\bigr),
\end{aligned}$$

and $U(z) \sim z/\mu$ and $1-a(x) \to 1$ yields

$$\liminf_{x\to\infty} \frac{\mathbb{P}(\tau(x) \le T(x))}{a(x)T(x)/\mathbb{E}C} \quad \ge \quad = \ (1-\epsilon)^2.$$

Let $\epsilon \downarrow 0$. □

We now turn to the study of extreme values. Recall that in the real–valued case, $\overline{X}_T = \max_{0\le t\le T} X_t$, and let ξ_k denote the maximum over cycle k, i.e. $\xi_0 = \sup_{0\le t\le Y_0} X_t$, $\xi_k = \sup_{S_{k-1}\le t\le S_k} X_t$, $k = 1,2,\ldots$. Thus, the ξ_k are independent and $\xi_1, \xi_2, \ldots$ have the common tail $a(x) = \mathbb{P}_0\left(\max_{t<Y} X_t > x\right)$.

Proposition 4.7 *Assume that E is a real interval and define*

$$G(x) = 1 - a(x) = \mathbb{P}_0\Bigl(\max_{t<Y} X_t \le x\Bigr), \quad F_T(x) = \mathbb{P}(\overline{X}_T \le x).$$

Then $\lim_{T\to\infty} \|F_T - G^{T/\mu}\| = 0$ where $\|\cdot\|$ is the uniform norm. Here in the delayed case with G having finite support, one needs in addition the condition

$$\mathbb{P}\Bigl(\xi_0 > \max_{k=1,\ldots,n} \xi_k\Bigr) \ \to\ 0. \tag{4.5}$$

Proof. The function $z(1-z^\gamma)$, $z \in [0,1]$, attains it maximum $\gamma/(1+\gamma)^{1+1/\gamma}$ at $z = (1+\gamma)^{-1/\gamma}$, and is therefore bounded by γ. Hence for all T, x and ϵ

$$\bigl|G^T(x) - G^{T(1+\epsilon)}(x)\bigr| \ \le\ \epsilon. \tag{4.6}$$

Define $k_T^\pm = \lfloor T(1\pm\delta)/\mu \rfloor$, and let $N_t = \inf\bigl\{n : S_n > t\bigr\}$. Then

$$F_T(x) \ \ge\ \mathbb{P}(\overline{X}_{S_{N_T}} \le x) \ \ge\ \mathbb{P}\Bigl(N_T \le k_T^+,\ \max_{0\le k\le k_T^+} \xi_k \le x\Bigr)$$

$$\geq \ \mathbb{P}\Bigl(\max_{1\leq k\leq k_T^+} \xi_k \leq x\Bigr) \ - \ \mathbb{P}(N_T > k_T^+) \ - \ \mathbb{P}\Bigl(\xi_0 > \max_{k=1,\ldots,k_T^+} \xi_k\Bigr).$$

Here the the two last terms on the r.h.s. tend to zero and are independent of x, while the first term is $G^{k_T^+}(x)$, which according to (4.6) can be approximated by $G^{T/\mu}(x)$ uniformly in x. A similar but easier upper bound

$$F_T(x) \ \leq \ \mathbb{P}\left(N_T < k_T^-\right) + \mathbb{P}\Bigl(\max_{1\leq k\leq k_T^-} \xi_k \leq x\Bigr)$$

completes the proof. □

The classical goal of extreme value theory is to find constants s_T, r_T such that $s_T(\overline{X}_T - r_T)$ has a limit Γ, say, in distribution. It is well known in the i.i.d. case that (up to trivial scalings and translations) such a Γ must either be a Gumbel (type I) r.v. Γ_G with c.d.f. $\mathbb{P}(\Gamma_G \leq x) = \mathrm{e}^{-\mathrm{e}^{-x}}$, $x \in \mathbb{R}$, a Fréchet (type II) r.v. Γ_F with c.d.f. $\mathbb{P}(\Gamma_F \leq x) = \mathrm{e}^{-x^{-\alpha}}$, $x \geq 0$, or a Weibull (type III) r.v. The type III case can occur only in the case of a bounded support and will not be discussed in detail here. Proposition 4.7 immediately yields that these limits are also the only possible ones in the case of regenerative processes:

Corollary 4.8 *Assume that E is a real interval unbounded to the right and that $s_T(\overline{X}_T - r_T) \xrightarrow{\mathscr{D}} \Gamma$ for some r.v. Γ. Then Γ has one of the extremal types Gumbel, Fréchet or Weibull.*

The two following types of asymptotics occur in a large number of queueing models (Proposition 4.9 covering light–tailed service times, cf. XIII.5, and Proposition 4.10 heavy–tailed service times, cf. X.9). The results follows immediately by translating standard i.i.d. results via Proposition 4.7, but we give self–contained proofs.

Proposition 4.9 *Assume that E is a real interval unbounded to the right and that $a(x) = \mathbb{P}_0(\tau(x) < Y) \sim c_0\mathrm{e}^{-\gamma x}$ as $x \to \infty$ continuously for some $c_0 > 0$ and some $\gamma > 0$. Then*

$$\gamma\overline{X}_T - \log T - \log(c_0/\mu) \ \xrightarrow{\mathscr{D}} \ \Gamma_G, \quad T \to \infty,$$

where Γ_G is Gumbel.

Proof. Using (4.2), we get

$$\begin{aligned}
&\mathbb{P}\bigl(\gamma\overline{X}_T - \log T - \log(c_0/\mu) \ \leq \ x\bigr) \\
&\quad = \ \mathbb{P}\bigl(\tau([x + \log T + \log(c_0/\mu)]/\gamma) \ > \ T\bigr) \\
&\quad \sim \ \mathbb{P}\bigl(\mu V/a([x + \log T + \log(c_0/\mu)]/\gamma) \ > \ T\bigr) \\
&\quad \sim \ \mathbb{P}\bigl(V \ > \ Tc_0 \exp\{-x - \log T - \log(c_0/\mu)\}/\mu\bigr) \\
&\quad = \ \mathbb{P}(V \ > \ \mathrm{e}^{-x}) \ = \ \mathrm{e}^{-\mathrm{e}^{-x}} \ = \ \mathbb{P}(\Gamma_G \leq x).
\end{aligned}$$

□

Proposition 4.10 *Assume that E is a real interval unbounded to the right or that E contains a set of the form $\{n_0, n_0+1, \ldots\}$, and that $a(x) \sim L(x)/x^\alpha$, $x \to \infty$, where $\alpha > 0$ and $L(x)$ is slowly varying. Then*

$$\overline{X}_T / b(T/\mu) \xrightarrow{\mathscr{D}} \Gamma_F, \quad T \to \infty,$$

where $b(x)$ is determined by $L(b(x))/b(x)^\alpha \sim 1/x$ and Γ_F is Fréchet.

Proof. Using first the definition of a slowly varying function (see A5) and next the definition of $b(\cdot)$, we get

$$\frac{a\big(xb(T/\mu)\big)T}{\mu} \sim \frac{L\big(xb(T/\mu)\big)T}{x^\alpha b(T/\mu)^\alpha \mu} \sim \frac{L\big(b(T/\mu)\big)T}{x^\alpha b(T/\mu)^\alpha \mu} \sim \frac{1}{x^\alpha}.$$

Hence (4.2) yields

$$\begin{aligned} & \mathbb{P}\big(\overline{X}_T / b(T/\mu) \le x\big) \\ = \ & \mathbb{P}\big(\tau(xb(T/\mu)) > T\big) \sim \mathbb{P}\big(V > [a(xb(T/\mu))T]/\mu\big) \\ \sim \ & \mathbb{P}\left(V > x^{-\alpha}\right) = \mathrm{e}^{-x^{-\alpha}} = \mathbb{P}(\Gamma_F \le x). \end{aligned}$$

□

Problems

4.1 Assume that $a(x) \sim c_0 \mathrm{e}^{-c_1 x^\beta}$, $x \to \infty$, for some c_0, c_1 and some $\beta < 1$. Show that there exist constants s_T, r_T such that $s_T(\overline{X}_T - r_T)$ has a Gumbel limit.

4.2 Assume that E contains a set of the form $\{n_0, n_0+1, \ldots\}$ and that $a(n) \sim c_0 \mathrm{e}^{-\gamma n}$, $n \to \infty$, for some $c_0 > 0$ and some $\gamma > 0$. Show that

$$\mathbb{P}(\gamma \overline{X}_T - \log T - \log(c_0/\mu) \le x) \sim \exp\left\{-\alpha(\log T + x)\mathrm{e}^{-x}\right\}$$

for all $x \in \mathbb{R}$, where α is the periodic function

$$\alpha(y) = \exp\left\{y + \log(c_0/\mu) - \gamma\lfloor[y + \log(c_0/\mu)]/\gamma\rfloor\right\}.$$

Notes The results of this section are standard; see the author's survey paper, Asmussen (1998b), for references to earlier literature. Standard treatments of extreme value theory are Leadbetter *et al.* (1983) and (emphasizing the heavy–tailed case) Embrechts *et al.* (1997). A notable recent paper on the regenerative case is Glasserman and Kou (1995a).

Note that only certain specific asymptotic forms of $1 - G(x) = a(x)$ allow us to find a normalization of the form $s_T(\overline{X}_T - r_T)$ such that a limit exists. One important exception is many light–tailed discrete distributions; see e.g. Problem 4.2 and further Andersson (1970). However, once the asymptotic form of $a(x)$ is known, then according to Proposition 4.7 the asymptotic properties of F_T are completely known as well, only in a possibly different form (cf. again Problem 4.2).

VII

Further Topics in Renewal Theory and Regenerative Processes

1 Spread–Out Distributions

By a *component* of a distribution F on $\mathbb{R}$ we understand a nonnegative measure G with the property $0 \neq G \leq F$. We say that F is *spread out* if F^{*n} has a component G that is absolutely continuous (i.e. has a density g w.r.t. Lebesgue measure) for some n.

In applied contexts, situations where F is nonlattice and spread out (or even has a density) are virtually the same. Strengthening the nonlattice assumption of renewal theory and regenerative processes to F being spread out does not therefore appear terribly restrictive, and the theory then gains some simplifications and strengthenings, rather in the spirit of the discrete–time case. The basic tool is *Stone's decomposition* of the renewal measure $U = \sum_0^\infty F^{*n}$:

Theorem 1.1 *If the interarrival distribution F of a renewal process is spread out, then we can write $U = U_1 + U_2$, where U_1, U_2 are nonnegative measures on $[0,\infty)$, U_2 is bounded ($\|U_2\| < \infty$) and U_1 has a bounded continuous density $u_1(x) = \mathrm{d}U_1(x)/\mathrm{d}x$ satisfying $u_1(x) \to 1/\mu$ as $x \to \infty$.*

The proof is based on smoothness properties of the convolution. Most of these are easy to check and are used without further reference. However, we shall prove:

Lemma 1.2 *If F is spread out, then F^{*m} has a uniform component on $(a, a+b)$ for some m, $a, b > 0$.*

Proof. We can assume that g is bounded with compact support. Choose continuous bounded functions $g_k \in L_1$ with $\|g - g_k\|_1 = \int |g - g_k| \to 0$. Then $g_k * g(x) = \int g_k(x-y)g(y)\,\mathrm{d}y$ is continuous by dominated convergence. Furthermore, $\|g^{*2} - g_k * g\|_\infty \le \|g\|_\infty \|g - g_k\|_1 \to 0$, where $\|\cdot\|_\infty$ is the supremum norm. Thus g^{*2} is continuous as the uniform limit of continuous functions; hence there exists $a, b, \delta > 0$ such that $g^{*2}(x) \ge \delta$ for $x \in (a, a+b)$. Take $m = 2n$. □

Proof of Theorem 1.1. In Lemma 1.2, let G be the uniform component and $g(x) = I(a \le x < a+b)\|G\|/b$ its density.

Assume first that $m = 1$ and let $H = F - G$, $U_2 = \sum_0^\infty H^{*n}$. Then

$$\begin{aligned} F^{*n} &= G * \sum_{k=0}^{n-1} F^{*(n-k-1)} * H^{*k} + H^{*n}, \\ U &= G * \sum_{k=0}^{\infty} H^{*k} * \sum_{n=k+1}^{\infty} F^{*(n-k-1)} + U_2 \;=\; G * U_2 * U + U_2. \end{aligned}$$

Since $\|H\| = 1 - \|G\| < 1$, we have $\|U_2\| < \infty$, and we must show that $U_1 = G * U_2 * U$ has the desired properties. Now $G * U$ has density $U * g(x) = \int_0^x g(x-y)U(\mathrm{d}y)$ which is bounded and continuous by dominated convergence ($g(0) = 0$ is needed for this since otherwise a discontinuity at x arises when U has an atom at x). Also $U * g(0) = 0$, and hence by the same argument $U_1 = U_2 * (G * U)$ has the bounded continuous density $u_1 = U_2 * (U * g)$. We then get

$$U * g(x) \;=\; \frac{\|G\|}{b} U(x-a-b, x-a] \;\to\; \frac{\|G\|}{\mu}, \tag{1.1}$$

$$u_1(x) \;=\; \int_0^x U * g(x-y)\, U_2(\mathrm{d}y) \;\to\; \frac{\|G\|}{\mu}\|U_2\| \;=\; \frac{1}{\mu}, \tag{1.2}$$

using dominated convergence in the first step of (1.2) and $\|U_2\| = (1 - \|H\|)^{-1} = \|G\|^{-1}$ in the last.

If $m > 1$, define $U^{(k)} = F^{*k} * \sum_0^\infty F^{*nm}$. Then from above, $U^{(0)} = U_1^{(0)} + U_2^{(0)}$ with $\|U_2^{(0)}\| < \infty$ and $U_1^{(0)}$ having a bounded continuous density $u_1^{(0)}$ with $u_1^{(0)}(x) \to 1/\mathbb{E}S_m = (m\mu)^{-1}$. This is readily seen to imply a similar decomposition of $U^{(k)} = F^{*k} * U^{(0)}$ and since $U = \sum_0^{m-1} U^{(k)}$, $U_1 = \sum_0^{m-1} U_1^{(k)}$ and $U_2 = \sum_0^{m-1} U_2^{(k)}$ have the desired properties. □

We proceed to give some main consequences of Stone's decomposition. The first is a version of the key renewal theorem $U * z(x) \to \mu^{-1} \int z$, where the strengthened assumption on F permits a weakening of the conditions on z, in particular to avoid reference to direct Riemann integrability.

Corollary 1.3 *Let z be bounded and Lebesgue integrable with $z(x) \to 0$, $x \to \infty$. Then $U * z(x) \to \mu^{-1} \int_0^\infty z(y)\,\mathrm{d}y$ provided F is spread out.*

Proof. By dominated convergence,

$$\begin{aligned} Z(x) &= U * z(x) \;=\; U_1 * z(x) + U_2 * z(x) \\ &= \int_0^x z(y) u_1(x-y)\,\mathrm{d}y \,+\, \int_0^x z(x-y)\,U_2(\mathrm{d}y) \\ &\to \int_0^\infty z(y)\mu^{-1}\,\mathrm{d}y \,+\, \int_0^\infty 0 \cdot U_2(\mathrm{d}y). \end{aligned}$$

□

Corollary 1.4 *Consider a regenerative process $\{X_t\}_{t \geq 0}$ with the cycle length distribution F being spread out with finite mean μ. Suppose as the only path regularity condition that $X_t(\omega)$ is measurable jointly in (t, ω). Then, no matter the initial conditions, the limiting distribution $\mathbb{P}_e$ of X_t exists in the sense of total variation convergence and is given by*

$$\mathbb{E}_e f(X_t) \;=\; \frac{1}{\mu} \int_0^\infty \mathbb{E}_0[f(X_s);\, Y > s]\,\mathrm{d}s.$$

Proof. It is easily seen that it is sufficient to consider the zero–delayed case. Define

$$Z(t) \;=\; \mathbb{P}_0(X_t \in A), \quad z(t) \;=\; \mathbb{P}_0(X_t \in A, Y > t).$$

Then $z(t)$ is Lebesgue measurable, and, being bounded by $\mathbb{P}_0(Y > t)$, also integrable with limit 0 at ∞. As in VI.1, $Z = z + F * Z = U * z$. Here

$$U_2 * z(t) \;\leq\; \int_0^t \mathbb{P}_0(Y > x - y)\,U_2(\mathrm{d}y),$$

$$\begin{aligned} |U_1 * z(t) - \mathbb{P}_e(X_t \in A)| \;&=\; \left| \int_0^t z(y) u_1(t-y)\,\mathrm{d}y \,-\, \frac{1}{\mu}\int_0^\infty z(y)\,\mathrm{d}y \right| \\ \leq\; \frac{1}{\mu}\int_t^\infty \mathbb{P}_0(Y > y)\,\mathrm{d}y \,&+\, \int_0^t \mathbb{P}_0(Y > y) \left| u_1(t-y) - \frac{1}{\mu} \right| \mathrm{d}y, \end{aligned}$$

and both these bounds are uniform in A and tend to zero as $t \to \infty$ (using dominated convergence). This proves t.v. convergence. □

A somewhat easier proof can be obtained using coupling; see the next section.

In many cases, it is also necessary for total variation convergence that F is spread out. For example (recall that $\mathrm{d}F_0/\mathrm{d}x = \overline{F}(x)/\mu$):

Corollary 1.5 *Let $\{B_t\}$ be the forward recurrence time process of a renewal process with interarrival distribution F with finite mean μ, and define $G_t(x) = \mathbb{P}(B_t \leq x)$. Then $G_t \to F_0$ in t.v. for any distribution of the initial delay if and only if F is spread out.*

Proof. Sufficiency follows from Corollary 1.4. Suppose F is not spread out so that for each n, F^{*n} is concentrated on a Lebesgue null set N_n, and consider the zero–delayed case. On $\{S_n \le t < S_{n+1}\}$, we have $t+B_t = S_{n+1} \in N_{n+1}$. Hence G_t is concentrated on the null set $\{y : y-t \in N_0 \cup N_1 \cup \cdots\}$ and the absolute continuity of F_0 yields $\|G_t - F_0\| = 1$. □

Finally we mention that instead of spread–out distributions one frequently works with distributions that are *strongly nonlattice*, i.e. satisfy *Cramér's condition* (C) $\overline{\lim}_{|s|\to\infty}\bigl|\widehat{F}[s]\bigr| < 1$, where $\widehat{F}$ is the characteristic function of F. We have:

Proposition 1.6 *F is spread out $\Rightarrow$ F is strongly nonlattice $\Rightarrow$ F is nonlattice.*

Proof. If F is itself absolutely continuous, then $\widehat{F}[s] \to 0$, $|s| \to \infty$, according to the Riemann–Lebesgue lemma. Thus if $F^{*n} \ge \epsilon G$ with G an absolutely continuous probability measure, we have

$$\overline{\lim_{|s|\to\infty}}\bigl|\widehat{F}[s]\bigr| \;=\; \overline{\lim_{|s|\to\infty}}\bigl|\widehat{F^{*n}}[s]\bigr|^{1/n} \;=\; \overline{\lim_{|s|\to\infty}}\bigl|\widehat{F^{*n}}[s]-\epsilon\widehat{G}[s]\bigr|^{1/n} \;\le\; (1-\epsilon)^{1/n}$$

and (C) holds. Finally, if F is lattice, say concentrated on $\{0, \pm\delta, \pm 2\delta, \ldots\}$, then $\widehat{F}[s] = 1$ for $s = 2k\pi/\delta$ and (C) cannot hold. □

In fact some results in renewal theory and regenerative processes require distributions that are only strongly nonlattice rather than spread out. However, the disadvantage of (C) is that the probabilistic significance is not clear and thus one has to rely on analytical methods.

Notes The theory was initiated by Stone (1966). An example where F is singular but F^{*2} not is given in Feller (1971, p. 146). It can be shown in continuation of Proposition 1.6 that discrete distributions cannot satisfy (C). See e.g. Bhattacharya and Rao (1976, p. 207).

2 The Coupling Method

2a. The Coupling Inequality
2b. The Classical Coupling of Markov Chains
2c. Coupling Proof of the Renewal Theorem
2d. Spread–Out Distributions and Exponential Rates

In the literature, the term *coupling* is used both in a broad and in a narrow sense. The broad sense is as follows: a coupling of two probability distributions $\mathbb{P}', \mathbb{P}''$ on $(\Omega', \mathscr{F}')$, resp. $(\Omega'', \mathscr{F}'')$, is defined as a probability distribution $\mathbb{P}$ on $(\Omega, \mathscr{F}) = (\Omega' \times \Omega'', \mathscr{F}' \otimes \mathscr{F}'')$ having marginals $\mathbb{P}'$ and $\mathbb{P}''$, i.e.

$$\mathbb{P}(A' \times \Omega'') = \mathbb{P}'(A'), \quad \mathbb{P}(\Omega' \times A'') = \mathbb{P}''(A'').$$

We shall use language such as "a coupling of X', X''," where X', X'' are r.v.'s, to denote a pair $(\widetilde{X}', \widetilde{X}'')$ of r.v.'s defined on a common probability space (a priori, X', X'' are not necessarily so) such that $\widetilde{X}' \stackrel{\mathscr{D}}{=} X'$, $\widetilde{X}'' \stackrel{\mathscr{D}}{=} X''$. For ease of notation, we omit the tilde, and broad sense coupling thus just means that the r.v.'s have been redefined on a common probability space without changing the marginal distributions.

A trivial case of a (broad sense) coupling is $\mathbb{P} = \mathbb{P}' \otimes \mathbb{P}''$; in terms of r.v.'s this just means that we take X', X'' independent with the prescribed distributions. However, the interesting examples involve dependence. One example is real–valued r.v.'s X', X'' such that $X' \leq_{\mathrm{so}} X''$ (stochastical ordering; see A4), where a classical result, stated in A4, on a.s. realization of stochastic order can be paraphrased as the existence of a broad sense coupling such that $X' \leq_{\mathrm{so}} X''$ a.s.

The set–up in the narrow (and more traditional) sense is that of two stochastic processes $\{X'_t\}_{t\in\mathbb{T}}$, $\{X''_t\}_{t\in\mathbb{T}}$, in discrete or continuous time and with the same state space E, and an associated random time $T \in \mathbb{T}$ (the *coupling time*) such that

$$X'_t = X''_t \quad \text{for all } t \geq T \tag{2.1}$$

(we shall encounter weakening of also this relation in connection with ϵ–coupling, requiring only that the processes are close rather than equal after T). A priori the two processes may be defined on different probability spaces, but the first step in the construction is to make them defined on the same $(\Omega, \mathscr{F}, \mathbb{P})$ (without changing the distributions).

2a *The Coupling Inequality*

We start by an inequality related to coupling in the broad sense ($\|\cdot\|$ is the total variation distance; cf. A8):

Proposition 2.1 *Let X', X'' r.v.'s taking values in the same space E and defined on a common probability space. Then*

$$\big\|\mathbb{P}(X' \in \cdot) - \mathbb{P}(X'' \in \cdot)\big\| \;\leq\; \mathbb{P}(X' \neq X''). \tag{2.2}$$

Proof. Write

$$\begin{aligned}
\mathbb{P}(X' \in A) &= \mathbb{P}(X' \in A, X' = X'') + \mathbb{P}(X' \in A, X' \neq X''),\\
\mathbb{P}(X'' \in A) &= \mathbb{P}(X'' \in A, X' = X'') + \mathbb{P}(X'' \in A, X' \neq X'')\\
&= \mathbb{P}(X' \in A, X' = X'') + \mathbb{P}(X'' \in A, X' \neq X'').
\end{aligned}$$

Subtracting, we get

$$\begin{aligned}
&\big|\mathbb{P}(X' \in A) - \mathbb{P}(X'' \in A)\big|\\
&\quad = \;\big|\mathbb{P}(X' \in A, X' \neq X'') - \mathbb{P}(X'' \in A, X' \neq X'')\big| \;\leq\; \mathbb{P}(X' \neq X'').
\end{aligned}$$

Taking supremum over A, the result follows. □

Corollary 2.2 *Let $\{X'_t\}_{t\in\mathbb{T}}$, $\{X''_t\}_{t\in\mathbb{T}}$ be stochastic processes defined on a common probability space, let θ_t be the shift, i.e. $(\theta_t X')_s = X'_{t+s}$, and let*

$T \le \infty$ *be a random time such that* (2.1) *holds. Then*

$$\big\|\mathbb{P}(\theta_t X' \in \cdot) - \mathbb{P}(\theta_t X'' \in \cdot)\big\| \ \le \ \mathbb{P}(T > t). \tag{2.3}$$

Proof. Replace X', X'' by $\theta_t X'$, $\theta_t X''$ in Proposition 2.1 and note that $\mathbb{P}(\theta_t X' \ne \theta_t X'') \le \mathbb{P}(T > t)$. □

The inequality (2.3) is commonly referred to as *the coupling inequality.* Its main applications are:

1. To show convergence in distribution of X_t as $t \to \infty$. In a Markovian setting, one takes $\{X'_t\}$ stationary, i.e. started by letting X'_0 have the stationary distribution π and $\{X''_t\}$ the given initial distribution ν. If a coupling with $T < \infty$ can be constructed, one has $\mathbb{P}(T > t) \to 0$ and
$$\big\|\mathbb{P}_\nu(X_t \in \cdot) - \pi(\cdot)\big\| \ \le \ \big\|\mathbb{P}(\theta_t X' \in \cdot) - \mathbb{P}(\theta_t X'' \in \cdot)\big\| \ \le \ \mathbb{P}(T > t) \ \to \ 0. \tag{2.4}$$
We will say in the following that the Markov chain (or process) *admits coupling* if a coupling with $T < \infty$ exists for any pair of initial distributions ν', ν'', i.e. if there exist stochastic processes $\{X'_n\}$, $\{X''_n\}$ and a random time $T < \infty$ defined on a common a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, such that $\{X'_n\}$, $\{X''_n\}$ are Markov chains with transition probabilities $(p_{ij})_{i,j\in E}$ and initial distributions ν', resp. ν'', and (2.1) holds for some random time $T < \infty$. If in addition a stationary distribution π exists, (2.4) shows that $X_n \overset{\mathscr{D}}{\to} \pi$ no matter the initial conditions.

2. To obtain estimates of the rate of convergence. One then shows that T can be chosen with $\mathbb{E}\varphi(T) < \infty$ for some φ increasing to ∞ (often, $\varphi(t) = t^p$ or $\mathrm{e}^{\epsilon t}$). Then
$$\big\|\mathbb{P}(X_t \in \cdot) - \pi(\cdot)\big\| \ \le \ \mathbb{P}(T > t) \ \le \ \frac{1}{\varphi(t)}\mathbb{E}\varphi(T) \ = \ \mathrm{O}\Big(\frac{1}{\varphi(t)}\Big)$$
[it should be noted as a limitation of the method that typically the convergence rates obtained in this way are not the best possible].

2b The Classical Coupling of Discrete Markov Chains

Let $\{X_n\}_{n=0,1,\ldots}$ be a Markov chain on a discrete state space E with transition probabilities $(p_{ij})_{i,j\in E}$.

Proposition 2.3 *A positive recurrent and aperiodic Markov chain on a discrete state space admits coupling.*

Proof. This is one of the relatively few examples where independent coupling works. Let $\Omega = E^{\mathbb{N}} \times E^{\mathbb{N}}$ (with $\mathscr{F}$ the obvious σ–field) and let $\mathbb{P}$ be such that the coordinate processes $\{X'_n\}$, $\{\widetilde{X}''_n\}$ (say) are Markov chains with transition probabilities $(p_{ij})_{i,j\in E}$ and initial distributions ν', resp. ν''. Obviously, $\{(X'_n, \widetilde{X}''_n)\}$ is Markov on $E \times E$ with n–step transition probabilities $q^n_{ij,k\ell} = p^n_{ik}p^n_{j\ell}$. For given i, j, k, ℓ, it follows from I.1.4 that $p^n_{ik} > 0$, $p^n_{j\ell} > 0$ for all sufficiently large n and hence $q^n_{ij,k\ell} > 0$. Thus $\{(X'_n, \widetilde{X}''_n)\}$ is irreducible. Further, if π is the stationary distribution for $\{X_n\}$, then $\pi\otimes\pi$ is stationary for $\{(X'_n, \widetilde{X}''_n)\}$. These facts imply that $\{(X'_n, \widetilde{X}''_n)\}$ is positive recurrent and aperiodic. By recurrence, $T_{ij} = \inf\{n : (X'_n, \widetilde{X}''_n) = (i, j)\}$ is finite for all i, j, and we define T either as T_{ii} for some fixed i or as $\min\{T_{ii} : i \in E\}$. Then T is a stopping time with the property $\widetilde{X}''_T = X'_T$. Now just define

$$X''_n \;=\; \begin{cases} \widetilde{X}''_n & n \le T \\ X'_n & n \ge T \end{cases} .$$

Then by the strong Markov property, $\{X''_n\}$ has the desired marginal distribution. So has $\{X'_n\}$ by construction, and clearly (2.1) holds. □

2c Coupling Proof of the Renewal Theorem

We say that two processes $\{X'_t\}$, $\{X''_t\}$ with continuous time parameter are *ϵ–coupled* if there exist (a.s. finite) random times T', T'' such that $|T' - T''| < \epsilon$ and

$$\theta_{T'}X' \;=\; \theta_{T''}X''. \tag{2.5}$$

A Markov process *admits ϵ–coupling* if for any ϵ and any two initial distributions ν', ν'' there exists ϵ–coupled versions $\{X'_t\}$, $\{X''_t\}$ with initial distributions ν', resp. ν''. Note that the existence of an ϵ–coupling can be rephrased as the existence of random times $T_\epsilon > 0, S_\epsilon$ such that $|S_\epsilon| < \epsilon$ and

$$X''_t = X'_{t+S_\epsilon}, \quad t \ge T_\epsilon. \tag{2.6}$$

The role of the concept is to provide convergence in distribution in some situations where t.v. convergence does not necessarily hold:

Proposition 2.4 *Consider a continuous–time Markov process $\{X_t\}$ having a stationary distribution π and right–continuous paths. If the process admits ϵ–coupling, then $X_t \xrightarrow{\mathscr{D}} \pi$ for any initial distribution.*

Proof. Let λ be an arbitrary initial distribution. We must show that $\mathbb{E}_\lambda f(X_t) \to \pi(f)$ for any continuous $f : E \to [0, 1]$.

Let $\nu' = \pi$, $\nu'' = \lambda$ and assume that (2.6) holds for a given $\epsilon \in (0, 1)$. Then for $t \ge T_\epsilon$,

$$\big|f(X'_t) - f(X''_t)\big| \;\le\; M'_t \;=\; \sup_{t-\epsilon\le s\le t+\epsilon} \big|f(X'_s) - f(X'_t)\big|.$$

Hence

$$\left|\mathbb{E}f(X_t') - \mathbb{E}f(X_t'')\right| \;\le\; \mathbb{E}M_t' + \mathbb{P}(T_\epsilon > t).$$

Since $\mathbb{E}f(X_t') = \pi(f)$, $\mathbb{E}M_t' = \mathbb{E}M_1'$ by stationarity, it follows that

$$\limsup_{t\to\infty}\left|\pi(f) - \mathbb{E}_\lambda f(X_t)\right| \;\le\; \mathbb{E}M_1'. \tag{2.7}$$

However, by stationarity $\{X_t\}$ has probability 0 to have a jump at $t = 1$ (cf. A2.3) and since f is continuous, we therefore have $M_1' \downarrow 0$ as $\epsilon \downarrow 0$. Thus the desired conclusion follows by letting $\epsilon \downarrow 0$ in (2.7). □

Now consider a renewal process in the notation of Ch. V with $\mu < \infty$ and recall that the distribution F_0 with density $\overline{F}(x)/\mu$ is stationary for the forward recurrence time process $\{B_t\}$. We let $\{B_t\}$, $\{B_t'\}$ be independent versions defined on a common probability space, such that $\{B_t'\}$ is stationary and $\{B_t\}$ has some arbitrary delay distribution.

Lemma 2.5 *For all large enough A, there exists a sequence $\tau_k \uparrow \infty$ of finite stopping times such that $B_{\tau_k} \le A$, $B_{\tau_k}' \le A$ for all k.*

Proof. Choose A with $F_0(A) > 1/2$. By the LLN VI.3.1 for cumulative processes (note that the proof only uses the elementary renewal theorem and not the renewal theorem itself!), we have

$$\frac{1}{T}\int_0^T I(B_t \le A)\,\mathrm{d}t \;\overset{\text{a.s.}}{\to}\; F_0(A), \qquad \frac{1}{T}\int_0^T I(B_t' \le A)\,\mathrm{d}t \;\overset{\text{a.s.}}{\to}\; F_0(A)$$

which in view of $F_0(A) > 1/2$ is only possible if the set $\{t:\; B_t \le A, B_t' \le A\}$ is unbounded w.p. 1. □

Lemma 2.6 *Assume that F is nonlattice. Then given A and $\epsilon > 0$, it is possible to choose a such that $\delta = \inf_{a\le x\le a+2A} h(x) > 0$, where $I(x) = [x-\epsilon, x+\epsilon]$, $h(x) = \mathbb{P}\bigl(Y_1 + \cdots + Y_n \in I(x) \text{ for some } n\bigr)$.*

Proof. Define $I'(x) = [x-\epsilon/2, x+\epsilon/2]$, $h'(x) = \mathbb{P}\bigl(Y_1 + \cdots + Y_n \in I'(x) \text{ for some } n\bigr)$. By V.5.1, we can choose a such that $I'(x)$ contains a point of the support of the renewal measure $\sum_0^\infty F^{*n}$ for each $x \ge a - \epsilon/2$ and hence $h'(x) > 0$. Choose $m, x_1, \ldots, x_m$ such that the $I'(x_j)$ cover $[a, a+2A]$ and let $\delta = \min_{j=1,\ldots,m} h'(x_j)$. □

Proof of the renewal theorem V.4.6. In the notation of Lemmas 2.5, 2.6, we have

$$B_{\tau_k}+[a, a+2A] \;\supseteq\; [a+A, a+2A], \quad B_{\tau_k}'+[a, a+2A] \;\supseteq\; [a+A, a+2A], \tag{2.8}$$

Therefore δ is a (pessimistic!) lower bound for the probability that $\{B_t\}$ has a renewal in $[\tau_k + a + A, \tau_k + a + 2A]$. Given that this occurs, let x be the position of the first such renewal. The probability that $\{B_t'\}$ has a renewal in $I(x)$ is then at least δ. In other words, the probability of $\{B_t\}$, $\{B_t'\}$ to have renewals at most ϵ apart in $[\tau_k + a + A, \tau_k + a + 2A]$

is at least δ^2. W.l.o.g., we may assume $\tau_{k+1} - \tau_k > a + 2A$ so that the adaptedness assumption of the geometric trials lemma A6.1 are satisfied. Thus, $\{B_t\}$ will eventually have a renewal at most ϵ apart from a renewal of $\{B_t'\}$. Replacing the following interarrival times for $\{B_t\}$ by those of $\{B_t'\}$ and denoting the resulting process by $\{B_t''\}$ provides an ϵ–coupling, and $B_t \xrightarrow{\mathscr{D}} F_0$ then follows by Proposition 2.4. □

2d Spread–Out Distributions and Exponential Rates

It is easy to adapt the above proof of the renewal theorem for the nonlattice case to give also say t.v. convergence of the recurrence times processes $\{A_t\}$, $\{B_t\}$ when F is spread out; cf. Corollary 1.5. In fact (with the convention that a renewal process admits coupling if $\{B_t\}$ does so):

Theorem 2.7 *A nonlattice renewal process with $\mu < \infty$ admits coupling if and only if F is spread out.*

The necessity follows from Corollary 1.5 since the coupling inequality applied to $\{B_t\}$ shows that the existence of a coupling implies t.v. convergence. For sufficiency, it is easy to either adapt the above proof of the renewal theorem for the nonlattice case to give t.v. convergence or apply Corollary 1.5; cf. Proposition 3.13 of the next section. We shall use a third variant that will also provide exponential rates with a small additional effort.

Lemma 2.8 *For a zero–delayed spread–out renewal process, there exists A, b such that the distributions of the B_t with $t \geq A$ have a common uniform component on $(0, b)$. That is, for some $\delta \in (0,1)$ and all $t \geq A$,*

$$\mathbb{P}(u < B_t \leq b) \ \geq \ \delta \frac{v-u}{b}, \quad 0 < u < v < b. \tag{2.9}$$

Proof. By Lemma 1.2, there exist m, $0 < c < d$, $\eta > 0$ such that $F^{*m}(v) - F^{*m}(u) \geq \eta(v-u)$ for $0 < c < u < v < d$. Let $b = (d-c)/2$, $a = (c+d)/2$. When $c < z < a$, $0 < u < v < b$, we then have $(u+z, v+z) \subseteq (c,d)$, and hence

$$\begin{aligned}\mathbb{P}(u < B_t \leq b) \ &\geq \ \int_{t-a}^{t-c} F^{*m}(t+u-y, t+v-y]\, U(\mathrm{d}y) \\ = \ &\int_c^a F^{*m}(u+z, v+z]\, U(t-\mathrm{d}z) \ \geq \ \eta(u-v)U(t-a, t-c).\end{aligned}$$

By Blackwell's renewal theorem, $U(t-a, t-c) \geq (a-c)/2\mu > 0$ for all large t. □

Proof of "if" in Theorem 2.7. We construct a zero–delayed and a stationary renewal process, $\{S_n\}$ and $\{S_n'\}$, on a common probability space in steps $k = 0, 1, 2, \ldots$. After step k, the renewal processes have been constructed in a certain random interval $[0, t_k]$, as have the overshoots B_{t_k}, B_{t_k}'. To get started, let $t_0 = 0$, $B_{t_0} = 0$ and choose B'_{t_0} according to F_0. Given step k has been completed, let

$$\begin{aligned} L_k &= \max\left[B_{t_k}, B'_{t_k}\right], \quad t_{k+1} \;=\; t_k + L_k + A, \\ s_k &= L_k + A - B_{t_k}, \quad s'_k \;=\; L_k + A - B'_{t_k} \end{aligned}$$

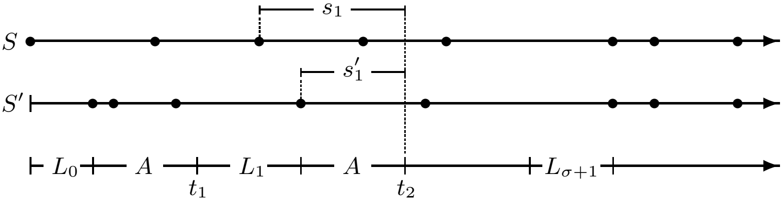

Figure 2.1

with A from Lemma 2.8; cf. Fig. 2.1. Then $s_k, s'_k \ge A$, and by Lemma 2.8 we can choose U_k, V_k, R_k, R'_k such that $\mathbb{P}(U_k = 1) = 1 - \mathbb{P}(U_k = 0) = \delta$, that V_k is uniform on $(0, b)$ and that

$$B_{t_{k+1}} \;=\; U_k V_k + (1 - U_k)R_k, \quad B'_{t_{k+1}} \;=\; U_k V_k + (1 - U_k)R'_k$$

have the overshoot distributions corresponding to s_k, resp. s'_k (U_k, V_k, R_k, R'_k are taken independent of all preceding $U_\ell, V_\ell, R_\ell, R'_\ell$). The renewals for $\{S_n\}$ in $[t_{k+1} - s_k, t_{k+1}]$ are then just taken according to the conditional distribution of the renewal process given that its overshoot at time s_k has the value of the constructed $B_{t_{k+1}}$, and similarly for $\{S'_n\}$. The procedure is stopped at step $\sigma = \inf\{k : U_k = 1\}$. Then the two renewal processes have a common renewal at time $T = t_\sigma + L_{\sigma+1}$, and defining $\{S''_n\}$ as the renewal process with the same renewals as $\{S_n\}$ before T and with the same renewals as $\{S'_n\}$ after provides the desired coupling. For the proof of Lemma 2.9 below, note that $\mathbb{P}(\sigma = n) = \delta(1-\delta)^n$. □

We now turn to the rate results and first note:

Lemma 2.9 *If* $\int_0^\infty e^{\eta x} F(dx) < \infty$ *for some* $\eta > 0$, *then also* $\mathbb{E}e^{\epsilon T} < \infty$ *for some* $\epsilon > 0$.

Proof. Since $z(t) = \mathbb{E}[e^{\eta B_t}; t < Y] \le e^{-\eta t}\mathbb{E}e^{\eta Y}$ is d.R.i., the usual renewal argument yields convergence of $Z(t) = \mathbb{E}e^{\eta B_t}$ to a finite limit. In particular, $Z(t) \le c_1 < \infty$ for all t. Now

$$\begin{aligned} \mathbb{E}\left[e^{\eta(A+L_{k+1})} \,\middle|\, B_{t_k}, B'_{t_k}\right] &\le \mathbb{E}\left[e^{\eta(A+B_{t_{k+1}})} + e^{\eta(A+B'_{t_{k+1}})} \,\middle|\, B_{t_k}, B'_{t_k}\right] \\ &= \mathbb{E}\left[e^{\eta(A+B_{s_k})} + e^{\eta(A+B'_{s_k})}\right] \;\le\; c \end{aligned}$$

where $c = 2c_1 e^{\eta A}$. Similarly $c_2 = \mathbb{E}e^{\eta(A+L_0)} < \infty$, and letting $T_n = \sum_0^n (A + L_k)$, it follows easily by induction that $\mathbb{E}e^{\eta T_n} \le c_2 c^n$. Now for some (large) p and some q (close to 1) with $1/p + 1/q = 1$, it holds that $c^{1/p}(1-\delta)^{1/q} < 1$

With $\epsilon = \eta/p$, Hölder's inequality then yields

$$\begin{aligned}\mathbb{E}\mathrm{e}^{\epsilon T} \;&\le\; \mathbb{E}\mathrm{e}^{\epsilon T_{\sigma+1}} \;=\; \sum_{n=1}^{\infty} \mathbb{E}\left[\mathrm{e}^{\epsilon T_{n+1}};\, \sigma = n\right] \\ &\le\; \sum_{n=1}^{\infty} \left[\mathbb{E}\mathrm{e}^{\eta T_{n+1}}\right]^{1/p} \mathbb{P}(\sigma = n)^{1/q} \;\le\; c_3 \sum_{n=1}^{\infty} c^{(n+1)/p}(1-\delta)^{n/q} \;<\; \infty\end{aligned}$$

where $c_3 = c_2^{1/p}\delta^{1/q}$. □

Theorem 2.10 *Assume* $\int_0^\infty \mathrm{e}^{\eta x} F(\mathrm{d}x) < \infty$ *for some* $\eta > 0$. *Then for some* $\epsilon > 0$:
(i) $\big\|\mathbb{P}_0(B_t \in \cdot) - F_0\big\| \;=\; \mathrm{O}(\mathrm{e}^{-\epsilon t})$, *cf. Corollary 1.5.*
(ii) *In Stone's decomposition,* $U_2[x,\infty) \;=\; \mathrm{O}(\mathrm{e}^{-\epsilon x})$, $u_1(x) \;=\; \mu^{-1} + \mathrm{O}(\mathrm{e}^{-\epsilon x})$.
(iii) *If* z *is measurable with* $z(x) \;=\; \mathrm{O}(\mathrm{e}^{-\delta x})$ *for some* $\delta > \epsilon$, *then*

$$U * z(t) \;=\; \frac{1}{\mu}\int_0^\infty z(x)\,\mathrm{d}x \;+\; \mathrm{O}(\mathrm{e}^{-\epsilon t}).$$

Proof. (i) is clear from (2.4) and Lemma 2.9. The proof of (ii) proceeds by reinspecting the derivation of Stone's decomposition. In the notation there, $H < F$ implies $\int \mathrm{e}^{\eta x}\, H(\mathrm{d}x) < \infty$ and hence $\int \mathrm{e}^{\epsilon x}\, H(\mathrm{d}x) < 1$ for some possibly smaller $\epsilon > 0$. This implies that $\int \mathrm{e}^{\epsilon x}\, U_2(\mathrm{d}x) < \infty$ and hence that $U_2[x,\infty) = \mathrm{O}(\mathrm{e}^{-\epsilon x})$. Using (i), the representation V.(2.8) of $U(x-a-b, x-a]$ and V.3.4, we have $U(x-a-b,x-a] = 1/\mu + \mathrm{O}(\mathrm{e}^{-\epsilon x})$ and hence (cf. (1.1), (1.2))

$$\begin{aligned}u_1(x) \;&=\; \int_0^x U * g(x-y)\, U_2(\mathrm{d}y) \;=\; \int_0^x \big[\|G\|/\mu + \mathrm{O}\big(\mathrm{e}^{-\epsilon(x-y)}\big)\big]\, U_2(\mathrm{d}y) \\ &=\; \|G\|\|U_2\|/\mu - \mathrm{O}\big(U_2[x,\infty)\big) + \mathrm{e}^{-\epsilon x}\!\int_0^x \mathrm{O}(\mathrm{e}^{\epsilon y})\, U_2(\mathrm{d}y) \;=\; \frac{1}{\mu} + \mathrm{O}(\mathrm{e}^{-\epsilon x}).\end{aligned}$$

Finally in (iii),

$$\begin{aligned}U * z(x) \;&=\; \int_0^x z(x-y)\, U_2(\mathrm{d}y) \;+\; \int_0^x z(x-y)u_1(y)\,\mathrm{d}y \\ &=\; \mathrm{e}^{-\epsilon x}\int_0^x \mathrm{O}(\mathrm{e}^{\epsilon y})\, U_2(\mathrm{d}y) \;+\; \int_0^x z(y)\Big\{\frac{1}{\mu} + \mathrm{O}\big(\mathrm{e}^{-\epsilon(x-y)}\big)\Big\}\,\mathrm{d}y \\ &=\; \mathrm{O}(\mathrm{e}^{-\epsilon x}) + \int_0^\infty \frac{z(y)}{\mu}\,\mathrm{d}y - \int_x^\infty \frac{z(y)}{\mu}\,\mathrm{d}y + \mathrm{e}^{-\epsilon x}\int_0^x \mathrm{O}\big(\mathrm{e}^{-(\delta-\epsilon)y}\big)\,\mathrm{d}y \\ &=\; \mathrm{O}(\mathrm{e}^{-\epsilon x}) + \int_0^\infty \frac{z(y)}{\mu}\,\mathrm{d}y.\end{aligned}$$

□

Corollary 2.11 *If, in addition to the conditions of Corollary 1.4, a regenerative process has* $\int_0^\infty \mathrm{e}^{\eta x} F(\mathrm{d}x) < \infty$ *and, in the delayed case,* $\mathbb{E}\mathrm{e}^{\eta Y_0} < \infty$

for some $\eta > 0$ *, then the t.v. convergence is exponentially fast. That is,* $\big\|\mathbb{P}(X_t \in \cdot) - \mathbb{P}_e(X_t \in \cdot)\big\| = \mathrm{O}(\mathrm{e}^{-\epsilon t})$ *for some* $\epsilon > 0$.

Proof. If $z_A(t) = \mathbb{P}_0(X_t \in A, t < Y)$, then $z_A(t) \le \mathbb{P}_0(Y > t) = \mathrm{O}(\mathrm{e}^{-\eta t})$, and a check of the above proof shows that

$$\mathbb{P}_0(X_t \in A) \;=\; U{*}z_A(t) \;=\; \mathrm{O}(\mathrm{e}^{-\epsilon t}) + \int_0^\infty \frac{z_A(y)}{\mu}\,\mathrm{d}y \;=\; \mathrm{O}(\mathrm{e}^{-\epsilon t}) + \mathbb{P}_e(X_t \in A)$$

uniformly in A. The delayed case is easily reduced to the zero–delayed one by conditioning upon Y_0. □

Problems

2.1 The purpose is to give a coupling proof of $X_n \xrightarrow{\mathscr{D}} \infty$ for an irreducible aperiodic null recurrent Markov chain. Let $\boldsymbol{P}$ be the transition matrix and $\boldsymbol{\nu}$ the stationary measure. Consider independent versions $\{X_n\}$, $\{X'_n\}$ with initial distributions $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$. Show that $\{Z_n\} = \{(X_n, X'_n)\}$ is irreducible and aperiodic. Define now $T = \inf\{n : X_n = X'_n = i\}$ for some fixed state i. Show that $\{Z_n\}$ is recurrent if and only $\mathbb{P}(T < \infty)$ for any choice of $i, \boldsymbol{\lambda}, \boldsymbol{\lambda}'$. If $\{Z_n\}$ is transient, let $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$ and show hereby $\mathbb{P}_{\boldsymbol{\lambda}}(X_n = i) \le \mathbb{P}(Z_n = (i,i))^{1/2} \to 0$. If $\{Z_n\}$ is recurrent, let $B \subseteq E$ be finite with $i \in B$, and define $\lambda'_j = \nu_j/\nu(B)$ for $j \in B$ and 0 otherwise. Show that

$$\overline{\lim}\,\mathbb{P}_{\boldsymbol{\lambda}}(X_n = i) \;=\; \overline{\lim}\,\mathbb{P}_{\boldsymbol{\lambda}'}(X'_n = i) \;\le\; \overline{\lim}\,\frac{(\boldsymbol{\nu}\boldsymbol{P}^n)_i}{\nu(B)} \;=\; \frac{\nu_i}{\nu(B)}$$

and hereby that $\mathbb{P}_{\boldsymbol{\lambda}}(X_n = i) \to 0$.

2.2 Show that $U(t) = t/\mu + \mathbb{E}Y^2/2\mu^2 + \mathrm{O}(\mathrm{e}^{-\alpha t})$ provided that F is spread out and $\int_0^\infty \mathrm{e}^{\delta x}\, F(\mathrm{d}x) < \infty$ for some $\delta > 0$.

Notes The idea of a coupling goes back to Doeblin in a 1938 paper, but today's interest in the subject was largely initiated by Pitman (1974). Two main monographs are Lindvall (1992) and Thorisson (2000).

Further interesting aspects not discussed here include the concept of a *maximal coupling*, the relations to the tail and invariant σ–fields and a *shift–coupling*. In the broad sense, a maximal coupling is one for which $\mathbb{P}(\widetilde{X}' \ne \widetilde{X}'')$ achieves the minimal value $\|\mathbb{P}(X' \in \cdot) - \mathbb{P}(X'' \in \cdot)\|$. In the narrow sense, a maximal coupling time achieves the exact rate of convergence. A shift–coupling is defined as an ϵ–coupling without the requirement $|T' - T''| < \epsilon$. A main theorem states that the existence of a shift–coupling is equivalent to $\mathbb{P}(X' \in F) = \mathbb{P}(X'' \in F)$ for all F in the invariant σ–field (there are similar characterizations of other sorts of couplings; see Thorisson, 2000).

For the history of the coupling proof of the renewal theorem, see Thorisson (2000) pp. 480–481. The present proof is basically a variant of Lindvall's (1977) argument.

For more on exponential convergence rates as in Theorem 2.10, see Lund *et al.* (1996) and references therein.

3 Markov Processes: Regeneration and Harris Recurrence

From the point of view of regenerative processes, the Markov case is a rather special one. Without doubt, a major force of the concept of a regenerative process is precisely that neither the Markov property nor other restrictions need to be put on the evolution in between regeneration points. Conversely, from the point of view of Markov processes on a general state space E, regeneration appears at first sight as a severe restriction. There is no apparent choice of regeneration points since e.g. the renewal processes of entrances to a fixed state x, so important in the discrete case, will only be nonterminating in quite special cases.

Nevertheless, the connection between Markov processes and regenerative processes has turned out to be of basic importance, and in fact ergodic theory for Markov processes in a simple and satisfying form is hardly known beyond the set–up to be developed below.

We consider as in I.8 a Markov process $\{X_t\}_{t\in\mathbb{T}}$ on E with $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = [0,\infty)$ and let $\mathscr{F}_t = \sigma(X_s : s \le t)$. If $\mathbb{T} = [0,\infty)$, it is assumed that E is Polish, $\mathscr{E}$ the Borel σ–algebra, that $\{X_t\}$ has D–paths and the strong Markov property holds. Write $P^t(x,A) = \mathbb{P}_x(X_t \in A)$.

Letting $\tau(R) = \inf\{t \ge 1 : X_t \in R\}$, we call a set $R \in \mathscr{E}$ *recurrent* if $\mathbb{P}_x(\tau(R) < \infty) = 1$ for all $x \in E$ (if $\mathbb{T} = [0,\infty)$, we need in addition to assume that $\tau(R)$ is measurable and that $X_{\tau(R)} \in R$). By the strong Markov property, this is equivalent to $\{t : X_t \in R\}$ being unbounded with probability 1, irrespective of initial conditions. We call R a *regeneration set* if R is recurrent and for some $r > 0$ the $P^r(x,\cdot)$, $x \in R$, contain a common component, i.e. for some $\epsilon \in (0,1)$ and some probability measure λ on E,

$$P^r(x,B) \ \ge\ \epsilon\lambda(B), \quad x \in R, \tag{3.1}$$

for all $B \in \mathscr{E}$. For example, this holds for a one–point set $R = \{x\}$ if and only if x is a recurrent state since then we may just take an arbitrary $r > 0$, $\epsilon = 1/2$ and $\lambda(B) = P^r(x,B)$. The following example is typical of applications and shows that regeneration sets exist in far more general situations:

Example 3.1 Assume that the transition functions contain components with smooth densities, i.e. for some μ, r and f^r we have

$$\begin{aligned} P^r(x,B) &\ \ge\ \int_B f^r(x,y)\,\mu(\mathrm{d}y), \\ E_0 &\ =\ \Bigl\{x \in E : \int_E f^r(x,y)\,\mu(\mathrm{d}y) \,>\, 0\Bigr\} \ne \emptyset \end{aligned} \tag{3.2}$$

where $f^r(x,y)$ is jointly continuous in x,y in a suitable topology on E. Then a regeneration set exists, provided that for some $x_0 \in E_0$ every neighbourhood of x_0 is recurrent. Indeed, choose $y_0 \in \mathrm{supp}(\mu)$ with $\delta = f^r(x_0,y_0) > 0$

and let R, S be neighbourhoods of x_0, resp. y_0, with $f^r(x,y) \geq \delta/2$, $x \in R$, $y \in S$. Then if $\lambda(B) = \mu(BS)/\mu(S)$, we have for $x \in R$ that

$$P^r(x,B) \;\geq\; \int_{BS} f^r(x,y)\,\mu(\mathrm{d}y) \;\geq\; \frac{\delta}{2}\mu(BS) \;=\; \frac{\delta\mu(S)}{2}\lambda(B). \qquad \Box$$

We call a Markov chain $\{X_n\}$ with a regeneration set *Harris recurrent* or just a *Harris chain* (the traditional equivalent definition looks somewhat different; see the end of the section; in continuous time the terminology is less well established). We shall justify the term "regeneration set" by showing that it is possible to construct $\{X_n\}$ simultaneously with a renewal process $S_0, S_1, \ldots$ w.r.t. which the Markov chain becomes regenerative. The idea is to randomize by w.p. ϵ letting a regeneration occur r time units after a visit to R and then restart according to λ. Choose an initial value $X_0 = x$ and just take the usual version of the process up to the time $\tau(R)$ where R is hit. Then realize $X_{\tau(R)+r}$ by w.p. ϵ letting the distribution be λ and a renewal epoch occur at $\tau(R)+r$, and w.p. $1-\epsilon$ letting the distribution be

$$\frac{1}{1-\epsilon}\left[P^r(X_{\tau(R)},\cdot) - \epsilon\lambda(\cdot)\right].$$

After that, realize the whole segment $\left\{X_{s+\tau(R)}\right\}_{0<s<r}$ by choosing it according to the conditional distribution of $\{X_s\}_{0<s<r}$ given that the boundary values X_0, X_r are the constructed $X_{\tau(R)}, X_{r+\tau(R)}$. Now repeat the procedure with the new initial value $X_{r+\tau(R)}$ and so on. That we get a Markov process with the given transition probabilities is intuitively obvious and easily verified. Also the distribution of $X_t = X_{S_n}$ at a renewal epoch $t = S_n$ is λ for all n and independent of $S_1, \ldots, S_n$. Hence the post–S_n process evolves in the same way for all n and is independent of $S_1, \ldots, S_n$. Thus we indeed have a regenerative process in the general sense of VI.1 where we do not require independent cycles. In fact, $X_{s+\tau(R)}$ needs not be independent of $X_{r+\tau(R)}$ if $0 < s < r$, and hence the last $r-1$ values in a cycle need not be independent of the next cycle. At least, the construction ensures that cycles are one–dependent (cycles $n+1, n+2, \ldots$ are independent of cycles $1, \ldots, n-1$).

We denote by $\mathbb{P}_\lambda$ the zero–delayed case where X_0 is chosen according to λ, by Y the length of the first cycle of the $\mathbb{P}_\lambda$–process.

The regeneration points obviously behave rather like stopping times, but are not so in the strict sense since in addition to $\mathscr{F}_\infty = \sigma(X_t;\, t \in \mathbb{T})$ they also depend on the 0–1 variables determining the randomizations. However, they fall into the framework of so–called *randomized stopping times.* We shall not go into a discussion of this subject but mention only that in the Markov chain case one of the possible definitions of τ being a randomized stopping time is

$$\mathbb{E}_x\big[g(X_n, X_{n+1}, \ldots);\, \tau > n\big] \;=\; \mathbb{E}_x\big[\mathbb{E}_{X_n} g(X_0, X_1, \ldots);\, \tau > n\big], \quad (3.3)$$

$$\mathbb{E}_x\big[g(X_s; s \geq t);\, \tau > t\big] \;=\; \mathbb{E}_x\big[\mathbb{E}_{X_t} g(X_s; s \geq 0);\, \tau > t\big] \quad (3.4)$$

for $\mathbb{T} = \mathbb{N}$, resp. $\mathbb{T} = [0, \infty)$. This relation will be needed below so we shall give a proof. Let $\mathbb{T} = \mathbb{N}$ (the continuous case is entirely similar) and let τ be any of $S_0, S_1, \ldots$. By the Markov property,

$$\mathbb{E}_x g(X_n, X_{n+1}, \ldots) \;=\; \mathbb{E}_x \mathbb{E}_{X_n} g(X_0, X_1, \ldots)$$

and therefore it is sufficient to prove (3.3) with $\tau > n$ replaced by $\tau \le n$. But conditionally upon τ, it holds on $\{\tau \le n\}$ that the Markov process is restarted according to λ at time τ. Thus on $\{\tau \le n\}$

$$\mathbb{E}_x \bigl[g(X_n, X_{n+1}, \ldots) \bigm| \tau; X_n \bigr] \;=\; \mathbb{E}_{X_n} g(X_0, X_1, \ldots)$$

and (3.3) follows easily.

A measure ν on $(E, \mathscr{E})$ is called *stationary* if $\nu \ge 0$, $\nu \ne 0$, ν is σ–finite and $\nu P^s = \nu$ for all $s \in \mathbb{T}$.

Theorem 3.2 *For a Markov process with a regeneration set, a stationary measure ν can be defined by*

$$\nu(f) \;=\; \mathbb{E}_\lambda \sum_{n=0}^{Y-1} f(X_n), \;\; \mathbb{T} = \mathbb{N}, \quad \nu(f) \;=\; \mathbb{E}_\lambda \int_0^Y f(X_t)\,\mathrm{d}t, \;\; \mathbb{T} = [0, \infty).$$

Proof. It is clear that $\nu \ge 0$ and $\nu \ne 0$. Also, a geometrical trial argument easily shows that if $E_{n,m} = \{x \in E : \mathbb{P}_x(\tau(R) \le n) \ge 1/m\}$ then $\nu(E_{n,m}) < \infty$. Since $E = \cup_{n,m} E_{n,m}$, ν is σ–finite. To show $\nu P^s = \nu$, let $\mathbb{T} = [0, \infty)$ (the discrete time case differs only in notation). Let f be fixed and define $g(x) = \mathbb{E}_x f(X_s)$. Then

$$\begin{aligned}
\nu P^s(f) \;&=\; \int \mathbb{E}_x f(X_s)\,\nu(\mathrm{d}x) \;=\; \int g(x)\,\nu(\mathrm{d}x) \;=\; \nu(g) \\
&=\; \mathbb{E}_\lambda \int_0^\infty g(X_t) I(Y > t)\,\mathrm{d}t \;=\; \int_0^\infty \mathbb{E}_\lambda \bigl[\mathbb{E}_{X_t} f(X_s);\, Y > t)\bigr]\,\mathrm{d}t
\end{aligned}$$

But according to (3.4) with $\tau = Y$, this is the same as

$$\begin{aligned}
&\int_0^\infty \mathbb{E}_\lambda \bigl[f(X_{s+t});\, Y > t)\bigr]\,\mathrm{d}t \\
&\;=\; \mathbb{E}_\lambda \int_s^{Y+s} f(X_u)\,\mathrm{d}u \;=\; \mathbb{E}_\lambda \Bigl\{ \int_s^Y + \int_Y^{Y+s} f(X_u)\,\mathrm{d}u \Bigr\} \\
&\;=\; \mathbb{E}_\lambda \Bigl\{ \int_s^Y + \int_0^s f(X_u)\,\mathrm{d}u \Bigr\} \;=\; \mathbb{E}_\lambda \int_0^Y f(X_u)\,\mathrm{d}u \;=\; \nu(f),
\end{aligned}$$

using the regeneration at Y in the third step. Since this holds for all f, the proof is complete. □

Corollary 3.3 *Let ν be as in Theorem 3.2 and $\mathbb{T} = \mathbb{N}$. Then $A \in \mathscr{E}$ is recurrent if and only if $\nu(A) > 0$, or equivalently if and only if $\mathbb{P}_\lambda(X_n \in A) > 0$ for some n.*

Proof. Starting from $X_0 = x$, we eventually end up with a regeneration, and thus A is recurrent if and only if $\mathbb{P}_\lambda(X_n \in A \text{ i.o.}) = 1$. Since cycles $1, 3, 5, \ldots$ are i.i.d. and $2, 4, 6, \ldots$ are so too, this is in turn equivalent to $\mathbb{P}_\lambda(X_n \in A \text{ for some } n < Y) > 0$ which again holds if and only if the expected number $\nu(A)$ of visits to A before Y is > 0. The last characterization now follows easily by a renewal argument. □

To investigate whether the stationary measure is unique, we first look at the case $\mathbb{T} = \mathbb{N}$. Let F be recurrent, $\{X_n^F\}$ the Markov chain restricted to F, cf. I.3, and let ν^F denote the restriction of ν to F (i.e. $\nu^F(A) = \nu(AF)$). Then:

Proposition 3.4 *If ν is stationary for $\{X_n\}$ and $0 < \nu(F) < \infty$, then ν^F is stationary for $\{X_n^F\}$.*

Proof. We may assume that $\nu^F(F) = \nu(F) = 1$. Letting $\mathbb{P}_\nu$ denote the measure defined for finite segments by

$$\mathbb{P}_\nu\bigl(X_0 \in A_0, \ldots, X_n \in A_n\bigr) \;=\; \int \mathbb{P}_x\bigl(X_0 \in A_0, \ldots, X_n \in A_n\bigr)\,\nu(\mathrm{d}x),$$

it is easily seen that $\mathbb{P}_\nu$ can be handled by the same formal rules as if ν was a proper probability (e.g. we have by stationarity that $\mathbb{P}_\nu(X_k \in A) = \nu(A)$). Let $A \subseteq F$ and define $c_n(A) = \mathbb{P}_\nu\bigl(X_0 \notin F, \ldots, X_{n-1} \notin F, X_n \in A\bigr)$. Then

$$\begin{aligned}
c_n(A) &= \mathbb{P}_\nu\bigl(X_1 \notin F, \ldots, X_n \notin F, X_{n+1} \in A\bigr)\\
&= c_{n+1}(A) + \mathbb{P}_\nu\bigl(X_0 \in F, X_1 \notin F, \ldots, X_n \notin F, X_{n+1} \in A\bigr),\\
\nu(A) &= \mathbb{P}_\nu(X_1 \in A) \;=\; \mathbb{P}_\nu(X_0 \in F, X_1 \in A) + c_1(A)\\
&\;\vdots\\
&= \sum_{k=1}^{n} \mathbb{P}_\nu\bigl(X_0 \in F, X_1 \notin F, \ldots, X_{k-1} \notin F, X_k \in A\bigr) \;+\; c_n(A).
\end{aligned}$$

Letting $n \to \infty$ yields

$$c_n(A) \;\to\; \nu(A) - \mathbb{P}_\nu\bigl(X_0 \in F, X_{\tau(F)} \in A\bigr) \;=\; \nu(A) - \mathbb{P}_{\nu^F}(X_1^F \in A).$$

But for $A = F$ the r.h.s. is just $1 - 1 = 0$. Thus $c_n(A) \le c_n(F) \to 0$, and $\nu(A) = \mathbb{P}_{\nu^F}(X_1^F \in A)$ and stationarity follows. □

Theorem 3.5 *For a* (discrete–time) *Harris chain, the stationary measure is unique up to a multiplicative constant.*

Proof. Existence was shown in in Theorem 3.2. For uniqueness, suppose first that $\mathbb{E}_\lambda Y = \nu(E) < \infty$ and let $\widetilde{\nu}$ be a different stationary measure with $\widetilde{\nu}(E) < \infty$. Then $\pi = \nu/\nu(E)$, $\widetilde{\pi} = \widetilde{\nu}/\widetilde{\nu}(E)$ are stationary distributions and by VI.1.5(ii) we have that

$$\frac{1}{d}\sum_{j=1}^{d} \mathbb{P}_{\widetilde{\pi}}(X_{nd+j} \in A) \;\to\; \frac{1}{\mathbb{E}_\lambda Y}\mathbb{E}_\lambda \sum_{n=0}^{Y-1} I(X_n \in A) \;=\; \pi(A),$$

where d is the period of Y. But by stationarity of $\widetilde{\pi}$, the l.h.s. is just $\widetilde{\pi}(A)$ for all n. Hence $\pi = \widetilde{\pi}$ so that $\nu, \widetilde{\nu}$ are proportional.

In the general case where not necessarily $\nu(E) < \infty$, $\widetilde{\nu}(E) < \infty$, we first assume $r = 1$ in (3.1). Proportionality follows if we can show that $\nu(A)/\nu(F) = \widetilde{\nu}(A)/\widetilde{\nu}(F)$ whenever $A \subseteq F$ and $\nu(F), \widetilde{\nu}(F) \in (0,\infty)$. Here F is recurrent according to Corollary 3.3 since $\nu(F) > 0$ and we can consider the chain $\{X_n^F\}$ for which $\nu^F, \widetilde{\nu}^F$ are both stationary by Proposition 3.4. Thus if we can prove that $\{X_n^F\}$ has a regeneration set, we have from above that $\nu^F = c\widetilde{\nu}^F$ and the desired conclusion follows. To this end, choose m, k, δ such that

$$R^F = \Bigl\{x \in F : \mathbb{P}_x\Bigl(X_m \in R, \sum_{n=1}^{m} I(X_n \in F) = k\Bigr) \geq \delta\Bigr\}$$

has positive ν–measure. Then R^F is recurrent for $\{X_n\}$, hence for $\{X_n^F\}$ and for $x \in R^F$ we have $\mathbb{P}_x(X_{k+1}^F \in A) \geq \delta\, \mathbb{P}_\lambda(X_{\tau(R)} \in A)$.

For a general r, consider a discrete time renewal process $\{S_n^*\}$, independent of $\{X_n\}$ and having interarrival distribution $\{f_n^*\}$ with finite mean and support on all of $\mathbb{N}$. Then the probability of a renewal at n is bounded away from 0, and it is easy to see by a geometric trial argument that R is recurrent for the Markov chain $\{X_n^*\} = X_{S_n^*}$. Further, $f_r^* > 0$ and $\mathbb{P}(X_1^* \in A \,|\, X_0^* = x) \geq \epsilon f_r^* \lambda(A)$. Thus a condition of type (3.1) holds with $r^* = 1$ so that from above the stationary measure of $\{X_n^*\}$ is unique up to a constant. But the transition kernel is $P^* = \sum_0^\infty f_n^* P^n$, and since $\nu P^n = \nu$, $\widetilde{\nu} P^n = \nu$, the measures $\nu, \widetilde{\nu}$ are both stationary for P^*, hence differ only by a constant. □

For $\mathbb{T} = \mathbb{N}$, we call the chain *aperiodic* if the $\mathbb{P}_\lambda$–distribution of Y is aperiodic (it follows from Proposition 3.10 below that this property does not depend on the choice of R, λ, ϵ). For $\mathbb{T} = [0,\infty)$, terminology such as "nonlattice cycles" or "spread–out cycles" refers to the $\mathbb{P}_\lambda$–distribution of Y in a similar manner. We call $\{X_t\}$ *positive recurrent* if $\mathbb{E}_\lambda Y = \|\nu\| < \infty$ and *null recurrent* if $\|\nu\| = \infty$ (an aperiodic positively recurrent Harris chain is simply called *Harris ergodic*). With $\pi = \nu/\|\nu\|$, the basic limit theorems for regenerative processes then give:

Theorem 3.6 *For a Harris ergodic chain, the $\mathbb{P}_x$–distribution of X_n converges to π in total variation. In particular, $P^n(x,A) \to \pi(A)$ for all $A \in \mathscr{E}$. For a continuous time positive recurrent Markov process with non–lattice cycles, the $\mathbb{P}_x$–distribution of X_t converges weakly to π.*

Theorem 3.6 is the main ergodic theorem for Harris chains, and we proceed to miscellaneous complements and extensions. First, since the LLN holds for identically distributed one–dependent variables $C_1, C_2, \ldots$ (consider $\{C_{2n}\}$ and $\{C_{2n+1}\}$ separately!), the same proof as in VI.3 yields:

Proposition 3.7 *In the positive recurrent discrete time case, the time–averages* $\sum_0^N f(X_n)/N$ *converge to* $\pi(f)$ *for any bounded measurable* f*. Similarly,* $\int_0^T f(X_t)\,\mathrm{d}t/T \to \pi(f)$ *in the positive recurrent continuous time case.*

Proposition 3.8 *Suppose that a continuous–time Markov process* $\{X_t\}_{t\geq 0}$ *either* (a) *has spread–out cycles or* (b) *that* (3.1) *holds for all* r *in an open interval. Then:*
(i) *Every discrete skeleton* $\{X_{n\delta}\}_{n\in\mathbb{N}}$ *is a Harris chain.*
(ii) *The stationary measure* ν *is unique up to a constant.*
(iii) *In the positive recurrent case* $\|\nu\| < \infty$*, the* $\mathbb{P}_x$*–distribution of* X_t *converges to* $\pi = \nu/\|\nu\|$ *in total variation.*
(iv) *In the null recurrent case,* $\mathbb{P}_x(X_t \in F) \to 0$ *for any set* $F \in \mathscr{E}$ *with* $\nu(F) < \infty$*.*

Proof. In case (b), we can impose an additional randomization by letting the regenerations occur at times after visits to R that are not fixed at r but uniformly distributed on say (a, b). Then it is immediately clear that the cycle length distribution is absolutely continuous, and we may proceed exactly as in the following argument for case (a). First (iii) follows by Corollary 1.4. For (i), we first show that R is recurrent for $\{X_{n\delta}\}$. Letting B_t be the forward recurrence time of the imbedded renewal process, it follows from $\{\delta\}$ being recurrent for $\{B_t\}$ that $[0, \delta]$ is recurrent for $\{B_{n\delta}\}$. Also when cycles are spread out, it is easy to see by a renewal argument that $g(t) = \mathbb{P}_\lambda(X_t \in R) \geq \epsilon > 0$ for t in an interval of length $> 2\delta$, therefore for $t \in [(m-1)\delta, m\delta]$ with m suitably chosen. Hence, if $\mathscr{G}_t = \sigma(X_s, B_s : s \leq t)$,

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbb{P}\big(X_{(n+m)\delta} \in R \,\big|\, \mathscr{G}_{n\delta}\big) &\geq \sum_{n=0}^{\infty} g(m\delta - B_{n\delta}) I(B_{n\delta} \leq \delta) \\ &\geq \epsilon \sum_{n=0}^{\infty} I(B_{n\delta} \leq \delta) \;=\; \infty\end{aligned}$$

and R being recurrent for $\{X_{n\delta}\}$ follows by the conditional Borel–Cantelli lemma. That R is a regeneration set for $\{X_{m\delta}\}$ is then easily proved: if $m_0\delta > r$, then for $x \in R$ we get from (3.1) that

$$\mathbb{P}^{m_0\delta}(x, A) \;\geq\; \epsilon \mathbb{P}_\lambda(X_{m_0\delta - r} \in A).$$

This proves (i), and (ii) is a consequence of (i) and Theorem 3.5. For (iv), check that $z(t) = \mathbb{P}_\lambda(X_t \in F, Y > t)$ satisfies the assumptions of Corollary 1.3. Hence $\mathbb{P}_\lambda(X_t \in F) = U * z(t)$ converges to $\int z/\mu = \nu(F)/\mu = 0$. That $\mathbb{P}_x(X_t \in F) \to 0$ then follows by conditioning upon Y_0. □.

Clearly, the proof of (iv) applies to the case $\mathbb{T} = \mathbb{N}$ as well, and thus:

Corollary 3.9 *In the discrete–time null recurrent case,* $\mathbb{P}_x(X_n \in F) \to 0$ *whenever* $\nu(F) < \infty$*.*

Turning to the periodicity problem, we have the following result concerning the existence of cyclic classes.

Proposition 3.10 *For $\mathbb{T} = \mathbb{N}$, there exists a $d = 1, 2, \ldots$ and a partitioning $E = E_1 \cup \cdots \cup E_d$ such that $P(x, E_{i+1}) = 1$ for all $x \in E_i \backslash N$, where N is a ν–null set* (here we identify E_{d+1} with E_1 and so on). *Furthermore, such a partitioning is unique in the sense that if a different one is given in terms of $\widetilde{d}, \widetilde{E}_1, \ldots, \widetilde{E}_d, \widetilde{N}$, then d is a multiple of $\widetilde{d}$, $d = c\widetilde{d}$, and after a cyclic permutation of the E_j one can achieve $\widetilde{E}_j = \cup_{k=0}^{c-1} E_{j+k\widetilde{d}}$ up to ν–null sets. Finally d can be characterized as the $\mathbb{P}_\lambda$–period of Y for some* [and therefore any] *choice of R, λ, ϵ in* (3.1).

Proof. We start from one representation (3.1) and define d to be the $\mathbb{P}_\lambda$–period of Y,

$$F_i \;=\; \left\{x \in E : P^{nd-r-i}(x, R) > 0 \text{ for some } n = 0, 1, 2, \ldots\right\}.$$

Since R is recurrent, $E = F_0 \cup \cdots \cup F_{d-1}$. The F_i need not be disjoint but, however, $\nu(F_i F_j) = 0$ for $i \neq j$. In fact, otherwise there is a m with $\mathbb{P}_\lambda(X_m \in F_i F_j) > 0$, implying that for some n_1, n_2 both $m + n_1 d - i$ and $m + n_2 d - j$ are in the support of Y, which is impossible.

A similar argument shows that $\mathbb{P}_\lambda(X_{nd+i} \in F_i) = 1$ for all n, i. Noting that if $\nu(A) = 0$, then $\mathbb{P}_\lambda(X_n \in A) = 0$ for all n, it follows that if we define $E_0 = F_0$, $E_i = F_i - E_0 - \cdots - E_{i-1}$ then E is the disjoint union of the E_i and $\mathbb{P}_\lambda(X_{nd+i} \in E_i) = 1$ for all n, i. To show that $E_{i,j} = \{x \in E_i : P(x, E_j) > 0\}$ is a ν–null set for $j \neq i + 1$, note similarly that otherwise $\mathbb{P}_\lambda(X_m \in E_{i,j}) > 0$ for some m. Here m must be of the form $nd + i$ and then $\mathbb{P}_\lambda(X_{nd+i+1} \in E_j) > 0$, which is only possible if $j = i + 1$.

Now let $\widetilde{E}_0, \ldots, \widetilde{E}_{\widetilde{d}-1}$ be a different set of cyclic classes, fix j and choose i with $\nu(E_i \widetilde{E}_j) > 0$. Let ψ be a probability measure that is equivalent to the restriction of ν to $E_i \widetilde{E}_j$. Then it is easy to see that if $A \subseteq E_i$, $\nu(A) > 0$, then $\mathbb{P}_\psi(X_{nd} \in A) > 0$ for all sufficiently large n. Letting first $A = E_i \widetilde{E}_j$, it follows that for some n both nd and $(n+1)d$ are multiples of $\widetilde{d}$. Hence $d = c\widetilde{d}$. Next with $A = E_i \backslash \widetilde{E}_j$, $\mathbb{P}_\psi(X_{nd} \in A) > 0$ would imply that nd is not a multiple of $\widetilde{d}$, which is impossible. Hence $\nu(E_i \backslash \widetilde{E}_j) = 0$. That is, if $\nu(E_i \backslash \widetilde{E}_j) > 0$, then $E_i \subseteq \widetilde{E}_j$ up to a ν–null set. Choose the numbering such that $E_0 \subseteq \widetilde{E}_0$. Then $E_i \subseteq \widetilde{E}_i$, $i = 0, \ldots, c-1$, $E_c \subseteq \widetilde{E}_0$, $E_{c+1} \subseteq \widetilde{E}_1$, …, $E_{2c} \subseteq \widetilde{E}_0$, …, and $\widetilde{E}_j = \cup_0^{c-1} E_{j+k\widetilde{d}}$ follows. □

Now let φ be a nontrivial σ–finite measure on $(E, \mathscr{E})$. We call $\{X_n\}$ φ–*recurrent* if any $F \in \mathscr{E}$ with $\varphi(F) > 0$ is recurrent, and φ*–irreducible* if to any $x \in E$ and $F \in \mathscr{E}$ with $\varphi(F) > 0$ we can find n with $\mathbb{P}_x(X_n \in F) > 0$ (obviously, φ–recurrence implies φ–irreducibility). Then:

Theorem 3.11 (OREY'S C–SET THEOREM) *Let $\{X_n\}$ be φ–irreducible and $\varphi(F) > 0$. Then we can find $C \subseteq F$, r and $\epsilon > 0$ such that $\varphi(C) > 0$ and $\mathbb{P}_x(X_r \in A) \geq \epsilon\varphi(AC)$ for all $x \in C$.*

The proof is a highly technical application of differentiation results for set functions and will not be given here (see the Notes for textbook references). However, the result permits us to characterize Harris chains (as defined here by existence of regeneration sets) in the following more traditional way:

Corollary 3.12 *A Markov chain is a Harris chain if and only if for some φ it is φ–recurrent. In that case, any set F with $\nu(F) > 0$ contains a regeneration set.*

Proof. If a regeneration set exists, the construction of the imbedded renewal process immediately shows that $\{X_n\}$ is λ–recurrent. Suppose, conversely, that $\{X_n\}$ is φ–recurrent, in particular φ–irreducible, and let $\varphi(F) > 0$. Choosing C as in the C–set theorem, we see that C is a regeneration set. Thus the stationary measure ν exists, the chain is ν–recurrent (Corollary 3.3) and we may repeat the argument to see that any F with $\nu(F) > 0$ contains a regeneration set. □

We remark that in practical cases the existence of regeneration sets seems far more easy to check than φ–recurrence. For example, for $E = \mathbb{R}$, the obvious choice of φ is frequently Lebesgue measure (possibly restricted to some interval) and it may be fairly easy to check that every interval is recurrent. But one needs to show recurrence of every Borel set of positive Lebesgue measure, and since such a set A can have a very complicated structure (e.g. A need not have interior points), this is a considerable task.

Finally:

Proposition 3.13 *A Markov chain $\{X_n\}$ with a stationary distribution π admits coupling* (cf. Section 2) *if and only if it is Harris ergodic. In that case, for any set F with $\pi(F) > 0$ and any two initial distributions μ, μ' it is possible to construct coupled versions $\{X_n\}$, $\{X'_n\}$ with the property $X_T = X'_T \in F$, where T is the coupling epoch.*

Proof. Suppose first that $\{X_n\}$ is Harris ergodic, let μ, μ', F be given and choose a regeneration set $R \subseteq F$ as in Corollary 3.12. We construct $\{X_n\}$, $\{X'_n\}$ by first realizing coupled versions of the imbedded renewal process (this is possible according to the discussion of coupling of ergodic Markov chains on a discrete state space given in Section 2b). Let T be the epoch of the first common renewal. Then we may choose $X_T = X'_T$ distributed according to λ (so that in particular $X_T = X'_T \in R \subseteq F$) and independent of the renewal process up to T, and construct the remaining X_n, $n \neq T$, according to their conditional distribution given X_T in such a way that $X_n = X'_n$, $n > T$.

Suppose, conversely, that $\{X_n\}$ admits coupling so that we have total variation convergence to π. Let $A \subseteq E$, $\pi(A) = 2\epsilon > 0$ and define $\tau =$

$\inf\{n \geq 1 : X_n \in A\}$. For any ψ, we may find $m(\psi)$ such that $\mathbb{P}_\psi(X_{m(\psi)} \in A) > \epsilon$ and hence $\mathbb{P}_\psi(\tau \leq m(\psi)) > \epsilon$. For a fixed $x \in E$, we now successively define integers $n(1) < n(2) < \cdots$ by $n(1) = m(\delta_x)$ and $n(k+1) = n(k) + m(\psi_k)$, where ψ_k is the conditional $\mathbb{P}_x$–distribution of $X_{n(k)}$ given $\tau > n(k)$. Then $\mathbb{P}_x(\tau > n(k)) \leq (1-\epsilon)^k$ so that $\mathbb{P}_x(\tau < \infty) = 1$ and π–recurrence follows. Null recurrence is excluded by the existence of a stationary distribution, and periodicity by total variation convergence as is easily seen from Proposition 3.10. □

Problems

3.1 The *Ornstein–Uhlenbeck process* with parameter $\xi > 0$ may be described as a Markov process with state space $\mathbb{R}$ and the $\mathbb{P}_x$–distribution of X_t being normal with mean $\mathrm{e}^{-\xi t}x$ and variance $(1-\mathrm{e}^{-2\xi t})/2\xi$. Show that (a) the normal distribution π with mean zero and variance $1/2\xi$ is stationary: (b) any discrete skeleton $\{X_{n\delta}\}_{n\in\mathbb{N}}$ is Harris recurrent [*Hint:* Test functions]; (c) X_t converges in total variation to π.

3.2 Show that if (3.1) holds for $r+1$ (with the same R, λ!) as well as r, then the chain is aperiodic.

Notes The theory was initiated largely by Harris in the 1950s (though Doeblin had some early results) and further main work done by Orey and others in the 1960s. The role of regenerative processes and minorization conditions such as (3.1) was realized independently by Nummelin and Athreya and Ney around 1978. A main textbook treatment is Meyn and Tweedie (1993). See also Orey (1971), Nummelin (1984) and Revuz (1984).

A nontrivial queueing application is given in XII.2. Among many further examples, we mention in particular Sigman (1988) and Dai (1995a).

4 Markov Renewal Theory

By a *Markov renewal process* we understand a point process where the interarrival times $T_0, T_1, \ldots$ are not necessarily i.i.d. but governed by a Markov chain $\{J_n\}$ with (finite or countable) state space E. This Markov dependence of the T_n may be formulated in various equivalent ways. One formulation is that T_n is sampled according to the current values of J_n, J_{n+1}. With $\mathscr{H} = \sigma(J_0, J_1, \ldots)$ this means that $T_0, T_1, \ldots$ are conditionally independent given $\mathscr{H}$ with

$$\mathbb{P}(T_n \leq t \,|\, \mathscr{H}) \;=\; \mathbb{P}\big(T_n \leq t \,\big|\, J_n, J_{n+1}\big) \;=\; G_{ij}(t) \tag{4.1}$$

on $\{J_n = i, J_{n+1} = j\}$ for a suitable family $(G_{ij})_{i,j\in E}$ of distributions on $(0,\infty)$. Equivalently, one may think of J_{n+1}, T_n being sampled simultaneously according to the current value of J_n. That is, $\{(J_{n+1}, T_n)\}$ is a Markov chain on $E \times (0,\infty)$ with the transition function depending only on the first coordinate. In particular, letting $\boldsymbol{Q} = (q_{ij})$ denote the transition matrix of $\{J_n\}$ and $F_{ij} = q_{ij}G_{ij}$, we have

$$F_{ij}(t) \;=\; \mathbb{P}\bigl(J_{n+1} = j, T_n \le t \,\big|\, J_n = i\bigr) \;=\; \mathbb{P}_i(J_1 = j, T_0 \le t) \qquad (4.2)$$

where $\mathbb{P}_i, \mathbb{E}_i$ refer to the case $J_0 = i$. The matrix $\boldsymbol{F}$ whose elements are the measures F_{ij} is called the *semi–Markov kernel*, and we define the associated *semi–Markov process* $\{W_t\}_{t\ge 0}$ by $W_t = J_0$ for $t < T_0$, $W_t = J_1$ for $T_0 \le t < T_1$ and so on. Obviously, the semi–Markov process and the Markov renewal process are in one–to–one correspondence (at least subject to regularity conditions like $q_{ii} = 0$, $T_0 + T_1 + \cdots = \infty$), and we shall not keep a formal distinction between them.

We note that the process reduces to a renewal process if E consists of one point, to a Markov chain with state space E if all G_{ij} are degenerate at 1 and to a continuous–time Markov process with state space E if all G_{ij} are exponential with intensities depending only on i, $\overline{G}_{ij}(t) = \mathrm{e}^{-\lambda(i)t}$. Thus, the Markov renewal process may be said to extend the continuous–time Markov jump process in the same way as the renewal process extends the Poisson process. We take these remarks as sufficient motivation for developing the theory and give just one practical example.

Example 4.1 Suppose in the traffic theory example V.1.2 that two types of vehicles are possible, e.g. cars and trucks. Then clearly the distribution of the distance between two vehicles depends in an essential way on their types. One could also model clumping by letting the type of a vehicle be its number in a clump. Suppose that a clump consists of n cars w.p. q_n $(q_1 + q_2 + \cdots = 1)$ and that the sizes of the clumps are independent. Then the Markov chain goes from state n to state 1 w.p. $q_n/(q_n + q_{n+1} + \cdots)$ and to state $n+1$ otherwise, and one could take all $G_{n1} = H_1$, all $G_{n(n+1)} = H_2$ (with H_1 stochastically larger than H_2). □

The following observation is the key to Markov renewal theory:

Proposition 4.2 *The instants t of returns to i $(W_t = i, W_{t-} \ne i)$ form a renewal process that is nonterminating if $\{J_n\}$ is recurrent. In that case, the mean interarrival time is μ/ν_i where $\boldsymbol{\nu}$ is the stationary measure for $\{J_n\}$ and $\mu = \sum_{i,j\in E} \nu_i \mu_{ij}$ with $\mu_{ij} = \int_0^\infty t\, F_{ij}(\mathrm{d}t)$.*

Proof. The first statement is obvious and letting $\tau = \inf\{n \ge 1 : J_n = i\}$, it is seen that if $J_0 = i$, then $T_0 + \cdots + T_{\tau-1}$ is the first interarrival interval of the renewal process. Hence, letting $m(i,j) = \int_0^\infty t\, G_{ij}(\mathrm{d}t)$ (so that $\mu_{ij} = q_{ij}m(i,j)$), the mean interarrival time is

$$\begin{aligned}\mathbb{E}_i(T_0 + \cdots + T_{\tau-1}) \;&=\; \mathbb{E}_i \sum_{n=0}^{\infty} \mathbb{E}_i\bigl[T_n;\, \tau > n \,\big|\, \mathscr{H}\bigr] \\ =\;\; &\mathbb{E}_i \sum_{n=0}^{\infty} I(\tau > n) m(J_n, J_{n+1})\end{aligned}$$

$$= \sum_{j,k\in E} m(j,k)\mathbb{E}_i \sum_{n=0}^{\infty} I\bigl(J_n = j, J_{n+1} = k, \tau > n\bigr)$$

$$= \sum_{j,k\in E} m(j,k)q_{jk}\mathbb{E}_i \sum_{n=0}^{\infty} I(J_n = j, \tau > n) \;=\; \sum_{j,k\in E} \mu_{jk}\frac{\nu_j}{\nu_i} \;=\; \frac{\mu}{\nu_i}.$$

□

In a standard manner, one can now prove that if $\{J_n\}$ is irreducible recurrent, then the interarrival distributions corresponding to Proposition 4.2 are either nonlattice for all i, or all lattice with the same span. We call the Markov renewal process (or equivalently the semi–Markov kernel $\boldsymbol{F}$) *nonlattice* in the first case and *lattice* in the second.

For two kernels $\boldsymbol{F}, \boldsymbol{G}$ we define $\boldsymbol{H} = \boldsymbol{G} * \boldsymbol{F}$ as the kernel with elements

$$H_{ij}(t) \;=\; \sum_{k\in E} G_{ik} * F_{kj}(t) \;=\; \sum_{k\in E} \int_0^t F_{kj}(t-u)\, G_{ik}(\mathrm{d}u)$$

and the convolution powers $\boldsymbol{F}^{*k}$ the obvious way (with $F_{ij}^{*0}(t) = \delta_{ij} I(t \geq 0)$). Using (4.2) and induction, one can immediately check that the interpretation in terms of Markov renewal processes is

$$F_{ij}^{*k}(t) \;=\; \mathbb{P}_i\bigl(J_k = j, T_0 + \cdots + T_{k-1} \leq t\bigr). \tag{4.3}$$

Define finally the *Markov renewal kernel* $\boldsymbol{U}$ as $\boldsymbol{U} = \sum_0^\infty \boldsymbol{F}^{*n}$. We then have the following generalization of Blackwell's theorem:

Theorem 4.3 *Consider a Markov renewal process with semi–Markov kernel $\boldsymbol{F}$ and $J_0 = W_0 = i$. Then $U_{ij}(t)$ is the expected number of returns to j before t,*

$$U_{ij}(t) \;=\; \sum_{n=0}^{\infty} \mathbb{P}_i\bigl(J_n = j, T_0 + \cdots + T_{n-1} \leq t\bigr). \tag{4.4}$$

In particular, $U_{ij}(t) < \infty$ and in the nonlattice case it holds in the notation of Proposition 4.2 that $U_{ij}(t+a) - U_{ij}(t) \to a\nu_j/\mu$, $t \to \infty$.

Proof. Here (4.4) is clear from (4.3), and the rest of the theorem is an immediate consequence of Proposition 4.2 and one–dimensional renewal theory. □

The *Markov renewal equation* (or *multivariate renewal equation*, system of *coupled renewal equations*, etc.) has the form

$$Z_i(t) \;=\; z_i(t) + \sum_{j\in E} \int_0^t Z_j(t-u)\, F_{ij}(\mathrm{d}u), \quad i \in E, \tag{4.5}$$

where the Z_i are unknown functions on $[0, \infty)$, the z_i known functions on $[0, \infty)$ and the F_{ij} known bounded measures on $[0, \infty)$. Equation (4.5) can be rewritten in matrix form as $\boldsymbol{Z} = \boldsymbol{z} + \boldsymbol{F} * \boldsymbol{Z}$, with the convolution defined in the manner consistent with (4.5), and in a similar manner as in one dimension we have:

Proposition 4.4 *Suppose that* $\boldsymbol{F}$ *is a semi–Markov kernel* (i.e. $\boldsymbol{Q} = \bigl(\|F_{ij}\|\bigr)$ a transition matrix) *and that* $\{J_n\}$ *is irreducible recurrent. Then if* $z_i \geq 0$ *and the* z_i *are bounded on finite intervals uniformly in* i*, it holds that* $\boldsymbol{Z} = \boldsymbol{U} * \boldsymbol{z}$ *is the unique solution to* (4.5) *with the* Z_i *uniformly bounded on finite intervals.*

Proof. Since $\boldsymbol{U} = \boldsymbol{F}^{*0} + \boldsymbol{F} * \boldsymbol{U}$, $\boldsymbol{F}^{*0} * \boldsymbol{z} = \boldsymbol{z}$, it is clear that $\boldsymbol{Z} = \boldsymbol{U} * \boldsymbol{z}$ is well defined and solves (4.5). Given two solutions of the type considered, their difference $\boldsymbol{D} = (D_i)$ satisfies $\boldsymbol{D} = \boldsymbol{F} * \boldsymbol{D} = \cdots = \boldsymbol{F}^{*k} * \boldsymbol{D}$ so that with $D_i^* = |D_i|$ we get $|\boldsymbol{D}^*| \leq \boldsymbol{F}^{*k} * \boldsymbol{D}^*$ and $K_T < \infty$ where $K_T = \sup_{i\in E, t\leq T} D_i^*(t)$. To prove that $\boldsymbol{D} \equiv \boldsymbol{0}$, let $0 \leq t \leq T$ and assume w.l.o.g. that $K_T = 1$. Then $\boldsymbol{D}^* \leq \boldsymbol{F}^{*k} * \boldsymbol{1}$ on $[0,T]$ and hence by (4.4),

$$D_i^*(t) \;\leq\; \sum_{j\in E} F_{ij}^{*k}(t) \;=\; \mathbb{P}_i(T_0 + \cdots + T_{k-1} \leq t).$$

We claim that $T_0 + T_1 + \cdots = \infty$ a.s. so that indeed $D_i(t) = 0$ follows as $k \to \infty$. To see this, note simply that the T_n with $J_n = i, J_{n+1} = j$ are i.i.d. given $\mathscr{H}$ and not degenerate at 0, and that $\mathbb{P}(J_n = i, J_{n+1} = j \text{ i.o.}) = 1$ for some i, j by recurrence. □

Using Theorem 4.3, one can now deduce exactly as in one dimension that if z_j is directly Riemann integrable (d.R.i.), then

$$U_{ij} * z_j(t) \;\to\; \frac{\nu_j}{\mu}\int_0^\infty z_j(x)\,\mathrm{d}x. \tag{4.6}$$

From this one expects the generalization

$$Z_i(t) \;=\; \sum_{j\in E} U_{ij} * z_j(t) \;\to\; \frac{1}{\mu}\sum_{j\in E}\nu_j\int_0^\infty z_j(x)\,\mathrm{d}x. \tag{4.7}$$

of the key renewal theorem. However, if E is infinite one cannot deduce (4.7) from (4.6) without imposing some further conditions. We shall not go into this but will be satisfied by noting:

Corollary 4.5 *Suppose in addition to the assumptions of Proposition 4.4 that* $\boldsymbol{F}$ *is nonlattice, that* E *is finite or, more generally, that* $z_i \equiv 0$ *except for a finite number of* i*, and that the* z_i *are d.R.i. Then* (4.7) *holds.*

We shall also derive an analogue of the asymptotic estimates of V.7 for the case where the rows of $\boldsymbol{Q}$ not necessarily have sum 1:

Theorem 4.6 *Consider the Markov renewal equation* (4.5) *with* $E = \{1,\ldots,p\}$ *and* $\boldsymbol{Q} = \bigl(\|F_{ij}\|\bigr)$ *irreducible. Suppose that for some real* β *the matrix* $\boldsymbol{A} = (a_{ij})$ *where* $a_{ij} = \int_0^\infty \mathrm{e}^{\beta u}\,F_{ij}(\mathrm{d}u)$ *has spectral radius* 1 *and* *choose* $\boldsymbol{\nu}$, $\boldsymbol{h}$ *with* $\boldsymbol{\nu A} = \boldsymbol{\nu}$, $\boldsymbol{Ah} = \boldsymbol{h}$, $\nu_i > 0$, $h_i > 0$, $i = 1, \ldots, p$. *Then*

$$\widetilde{F}_{ij}(\mathrm{d}u) \;=\; \frac{h_j}{h_i}\mathrm{e}^{\beta u}\, F_{ij}(\mathrm{d}u)$$

defines a semi–Markov kernel $\widetilde{\boldsymbol{F}}$ *with* $\widetilde{\boldsymbol{Q}} = \big(\|\widetilde{F}_{ij}\|\big)$ *irreducible recurrent. Let further* $\widetilde{Z}_i(t) = \mathrm{e}^{\beta t}Z_i(t)/h_i$, $\widetilde{z}_i(t) = \mathrm{e}^{\beta t}z_i(t)/h_i$. *Then* $\boldsymbol{Z} = \boldsymbol{z} + \boldsymbol{F} * \boldsymbol{Z}$ *implies that* $\widetilde{\boldsymbol{Z}} = \widetilde{\boldsymbol{z}} + \widetilde{\boldsymbol{F}} * \widetilde{\boldsymbol{Z}}$. *Hence if* $\widetilde{\boldsymbol{F}}$ *is nonlattice and the* $\widetilde{z}_i$ *d.R.i., then*

$$\lim_{t\to\infty} \widetilde{Z}_i(t) \;=\; \frac{1}{\widetilde{\mu}}\sum_{j=1}^{p}\widetilde{\nu}_j \int_0^\infty \widetilde{z}_j(x)\,\mathrm{d}x,$$

where $\widetilde{\boldsymbol{\nu}}$ *is stationary for* $\widetilde{\boldsymbol{Q}}$, *hence*

$$\lim_{t\to\infty} \mathrm{e}^{\beta t}Z_i(t) \;=\; \frac{h_i\displaystyle\sum_{j=1}^{p}\nu_j\int_0^\infty \mathrm{e}^{\beta x}z_j(x)\,\mathrm{d}x}{\displaystyle\sum_{k,j=1}^{p}\nu_k h_j\int_0^\infty x\mathrm{e}^{\beta x}\,F_{kj}(\mathrm{d}x)}\,. \tag{4.8}$$

Proof. That $\widetilde{\boldsymbol{Q}} = (h_j a_{ij}/h_i)$ is a transition matrix is immediate and has in fact already been noted in I.6 where it was also found that $\widetilde{\nu}_j = \nu_j h_j$ (the existence of $\boldsymbol{\nu}$, $\boldsymbol{h}$, $\widetilde{\boldsymbol{\nu}}$ is ensured by I.6.5). The rest of the proof is trivial manipulation. □

For conditions for β to exist, see Problem 4.3.

Example 4.7 Consider the Lotka–Sharpe population model from V.2.2, but assume now that each woman has one of p types. The type can change during life and could, for example, be one of p social groups, one of p geographical regions in which the woman lives, or the parity of the woman, i.e. the number of children already born (then group p comprises all women having $p-1$ or more children). In such situations, it is highly relevant to assume that the birth rates and survival rates depend on types, say type i women aged a give birth to type j daughters at rate $\lambda_{ij}(a)$ and can be found t time units later in group j in an average proportion of ${}_tp_a(i,j)$. Then if $Z_i(t)$ is the rate of birth of type i girls at time t, $f_0^{(j)}(a)$ the density of type j women aged a in the initial population, we get exactly as for $p=1$

$$Z_i(t) \;=\; z_i(t) + \sum_{j,k=1}^{p}\int_0^t Z_j(t-s)\,{}_sp_0(j,k)\lambda_{ki}(s)\,\mathrm{d}s,$$

where

$$z_i(t) \;=\; \sum_{j,k=1}^{p}\int_0^\infty f_0^{(j)}(a)\,{}_tp_a(j,k)\lambda_{ki}(a+t)\,\mathrm{d}a.$$

This is of the form (4.5) with

$$\frac{\mathrm{d}F_{ij}(x)}{\mathrm{d}x} \;=\; \sum_{k=1}^{p} {}_x p_0(j,k)\lambda_{ki}(x),$$

and the program of Theorem 4.6 may be carried through to obtain asymptotic estimates of the form $Z_i(t) \sim \mathrm{e}^{-\beta t} c_i$. □

Problems

4.1 Suppose that the Markov chain $J_0, J_1, \ldots$ imbedded in a Markov renewal process is transient and define the lifetime as $L = T_0 + T_1 + \cdots$. Explain that, in contrast to the recurrent case, it is possible that $L < \infty$ a.s. Define $Z_i(t) = \mathbb{P}_i(L \le t)$. Show that $\boldsymbol{Z} = \boldsymbol{U} * \boldsymbol{Z}$ and that the solution of the Markov renewal equation needs not be unique in the transient case.

4.2 Consider a Markov renewal equation of the form

$$\begin{aligned} Z_1(t) &= z_1(t) + F_{11} * Z_1(t) + F_{12} * Z_2(t) \\ Z_2(t) &= z_2(t) + \qquad\qquad\qquad\quad F_{22} * Z_2(t) \end{aligned}$$

where $0 < \|F_{11}\| < 1$, $\|F_{11}\| + \|F_{12}\| = \|F_{22}\| = 1$. Show that the solution is unique and find its limiting behaviour.

4.3 Consider the set–up of Theorem 4.6 and suppose that $\boldsymbol{Q}$ is irreducible with $1 < \mathrm{spr}(\boldsymbol{Q}) < \infty$. Show that β always exists. [*Hint:* Let $\boldsymbol{A} = \boldsymbol{A}_\beta$ be as in Theorem 4.6, $\rho(\beta) = \mathrm{spr}(\boldsymbol{A}_\beta)$, $p(\boldsymbol{B}, \lambda) = \det(\boldsymbol{B} - \lambda \boldsymbol{I})$. Show that $\partial p/\partial\lambda$ evaluated at $\boldsymbol{B} = \boldsymbol{A}_\beta$, $\lambda = \rho(\beta)$ is nonzero and thereby, using the implicit function theorem, that $\rho(\beta)$ is continuous in β. Show finally $\rho(\beta) \to 0$, $\beta \to -\infty$.]

Notes See the next section.

5 Semi–Regenerative Processes

The concept of semi–regenerativity generalizes regenerative processes by allowing the regeneration points to be of several types, indexed by $i \in E$ where E is finite or countable. Thus instead of an imbedded renewal process we have an imbedded Markov renewal process specified say by $\{(J_n, T_n)\}_0^\infty$. Each time state i is entered, the semi–regenerative process $\{X_t\}$ is restarted subject to the ith set of initial conditions and independent of the Markov renewal process up to that time.

More formally, let $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = [0, \infty)$ and let $(\mathbb{P}_i)_{i \in E}$ be a governing set of probabilities for $\{X_t\}_{t \in \mathbb{T}}$. We then call $\{X_t\}$ *semi–regenerative* if we can find a Markov renewal process (possibly defined on an enlarged probability space), such that for any n the conditional distribution of $\{X_{t+T_0+\cdots+T_{n-1}}\}_{t \in \mathbb{T}}$ given $T_0, \ldots, T_{n-1}$, $J_0, \ldots, J_{n-1}$, $J_n = i$ is the same as the $\mathbb{P}_i$–distribution of $\{X_t\}$ itself. Thus if E consists of one point, the concept reduces to regenerative processes. Even in the general case, we have:

Proposition 5.1 *Any semi–regenerative process with* $\{J_n\}$ *recurrent is regenerative, with the instants of returns of the Markov renewal process to* $i \in E$ *as an imbedded renewal process.*

This is an immediate consequence of the definitions. From the point of view of proving the existence of limiting distributions, semi–regenerative processes are therefore not a much more powerful tool than regenerative processes. Rather, the formulas derived for the limits may be somewhat more convenient than the expression $(\mathbb{E}C)^{-1}\mathbb{E}\int_0^C f(X_t)\,\mathrm{d}t$ for regenerative processes, and at least serve the purpose of doing some reduction once and for all:

Proposition 5.2 *Consider a semi–regenerative process with* $\{J_n\}$ *irreducible recurrent, say with stationary measure* $\boldsymbol{\nu}$*. Define* $C = T_0$ *and suppose that* $\mu = \sum_{i\in E}\nu_i\mathbb{E}_iC$ *is finite. Then:*
(i) *if* $\mathbb{T} = [0,\infty)$*, the imbedded Markov renewal process is nonlattice and* $\{X_t\}$ *has metric state space and right–continuous paths, then the limiting distribution exists and is given by*

$$\mathbb{E}_e f(X_t) \;=\; \frac{1}{\mu}\sum_{j\in E}\nu_j\,\mathbb{E}_j\int_0^C f(X_t)\,\mathrm{d}t; \tag{5.1}$$

(ii) *if* $\mathbb{T} = \mathbb{N}$ *and the imbedded Markov renewal process is aperiodic on* $\mathbb{N}$*, then the limiting distribution exists and is given by* (5.1) *with* $\int_0^C$ *replaced by* $\sum_0^{C-1}$.

Proof. First check that the expression given for μ is the same as in Proposition 4.2. Hence if $\mu < \infty$, we may appeal to Proposition 5.1 and use VI.1.3 to get the existence of the limit as well as the expression

$$\mathbb{E}_e f(X_t) \;=\; \frac{1}{\mathbb{E}_i\widetilde{C}}\,\mathbb{E}_i\int_0^{\widetilde{C}} f(X_t)\,\mathrm{d}t \;=\; \frac{\nu_i}{\mu}\,\mathbb{E}_i\int_0^{\widetilde{C}} f(X_t)\,\mathrm{d}t \tag{5.2}$$

where $\widetilde{C} = T_0 + \cdots + T_{\tau-1}$, $\tau = \inf\{n \ge 1 : J_n = i\}$. Now the semi–regenerative property implies

$$\mathbb{E}_i\left[\int_{T_0+\cdots+T_{k-1}}^{T_0+\cdots+T_k} f(X_t)\,\mathrm{d}t\;\middle|\;T_0,\ldots,T_{k-1},J_0,\ldots,J_k\right] \;=\; \mathbb{E}_j\int_0^C f(X_t)\,\mathrm{d}t$$

on $\{J_k = j\}$. Hence (5.2) can be written as

$$\frac{\nu_i}{\mu}\,\mathbb{E}_i\sum_{n=0}^{\tau-1}\mathbb{E}_{J_n}\int_0^C f(X_t)\,\mathrm{d}t \;=\; \frac{\nu_i}{\mu}\sum_{j\in E}\mathbb{E}_j\int_0^C f(X_t)\,\mathrm{d}t\cdot\mathbb{E}_i\sum_{n=0}^{\tau-1}I(J_n = j),$$

and since the last factor is just ν_j/ν_i, (5.1) follows. The proof of (ii) is entirely similar. □

Corollary 5.3 *Let* $\{X_n\}$ *be an irreducible recurrent Markov chain with discrete state space* E *and let* X_k^F *be the value of* X_n *at the kth visit to*

$F \subset E$, $\tau(F) = \inf\{n \geq 1 : X_n \in F\}$. Then a measure $\boldsymbol{\nu}$ is stationary for $\{X_n\}$ if and only if $(\nu_i)_{i \in F}$ is stationary for $\{X_n^F\}$ and

$$\nu_j = \sum_{k \in F} \nu_k \, \mathbb{E}_k \sum_{n=0}^{\tau(F)-1} I(X_n = j) \, . \tag{5.3}$$

Proof. Let $i \in F$ be fixed and define $\tau = \inf\{n \geq 1 : X_n = i\}$, $\nu_j = \mathbb{E}_i \sum_0^{\tau-1} I(X_n = j)$. Then according to I.3.9 (cf. also Proposition 3.4), it only remains to show that (5.3) holds. But letting $J_n = X_n^F$, $C = \tau(F)$, $\tau^F = \inf\{n \geq 1 : X_n^F = i\}$, the expectation in the definition of ν_j may then be evaluated exactly as in the proof of Proposition 5.2 and we get

$$\begin{aligned} \nu_j &= \sum_{k \in F} \mathbb{E}_i \sum_{n=0}^{\tau^F - 1} I(X_n^F = k) \, \mathbb{E}_k \sum_{n=0}^{C-1} I(X_n = j) \\ &= \sum_{k \in F} \nu_k^{(F)} \, \mathbb{E}_k \sum_{n=0}^{C-1} I(X_n = j) \;=\; \sum_{k \in F} \nu_k \, \mathbb{E}_k \sum_{n=0}^{C-1} I(X_n = j) \end{aligned}$$

□

Notes A classical source for Markov renewal theory and semi–regenerative processes (in the case of a discrete E) is Çinlar (1975). As argued in the text, the extension from renewal theory and regenerative processes does not present intrinsically new mathematical difficulties in the discrete case, but nevertheless the versatility of the set–up makes it highly useful and popular in applications. A broad spectrum of topics in the area can be found in the volume edited by Janssen and Limnios (1999).

Markov renewal theory on a general state space E has received considerable attention. See Alsmeyer (1997) for a recent contribution and references.

6 Palm Theory, Rate Conservation and PASTA

Assume that a stochastic process $\{X_t\}$ has a stationary distribution π and that we sample it at a sequence of random time points $\{T_k\}$. Then the stationary distribution of the X_{T_k} (if it exists) is typically not π but some other distribution ν. What can be said about the relation between π and ν? In particular, in which situations is indeed $\pi = \nu$?

The prototype of this sort of question is the comparison of the stationary current life distribution in a renewal process and the interarrival distribution. Here the waiting time paradox means that sampling at an arbitrary point of time favours long interarrival intervals. The precise formulas for this biasing effect are given in V.3, and we will see that closely analogous results hold more generally.

Example 6.1 As a further motivating example, consider a Markov–modulated Poisson process generated by an ergodic Markov process $\{J_t\}$ with a finite state space E and intensities β_i, $i \in E$. Then arrivals occur as in a Poisson process with intensity β_i when $J_t = i$. Write π for the stationarity distribution for $\{J_t\}$ and ν for the stationarity distribution of the Markov chain $\{J_{T_k}\}$ where $\{T_k\}$ is the sequence of arrival epochs. To compute ν, we use a time–average argument. The fact that the average time spent in state i up to time T is approximately $\pi_i T$ implies that the number of arrivals up to time T occuring while $J_t = i$ is approximately

$$\beta_i \int_0^T I(J_t = i)\,\mathrm{d}t \;\approx\; T\beta_i\pi_i .$$

Summing over i, it follows that the total number of arrivals is approximately $T\sum_{i\in E}\beta_i\pi_i$. Identifying the fraction of the arrivals occuring in state i with ν_i, we finally get

$$\nu_i \;=\; \frac{\beta_i\pi_i}{\sum_{j\in E}\beta_j\pi_j}. \tag{6.1}$$

Again, we have an instance of biasing: sampling at arrival times favours states i with a large β_i. □

Palm theory is a general framework in which to carry out these types of calculations (and much more general ones!). The traditional setting is marked point processes, but to adhere more with the mainstream of this book, we choose a formally different but mathematically equivalent one involving a pair of a stochastic process and an associated sequence of random time points.

The object that we sample "at an arbitrary point of time" or at the event times is a stochastic process $\mathscr{X} = \{X_t\}$; most often we assume that $\mathscr{X}$ has doubly infinite time, $t \in (-\infty, \infty)$ (in the stationary case, this can always be obtained by a standard construction). The shifted process $\theta_s\mathscr{X}$ is defined by $(\theta_s\mathscr{X})_t = X_{t+s}$. The event times at which we sample are

$$\cdots < T_{-2} < T_{-1} < 0 \le T_0 < T_1 < \cdots$$

and the sequence $\{T_n\}_{n=0,\pm1,\pm2,\ldots}$ is denoted $\mathscr{T}$. The shift $\theta_s\mathscr{T}$ is defined by $(\theta_s\mathscr{T})_k = T_{N_s+k} - s$ where $N_t = \max\{\ell : T_\ell \le t\}$ (note that $(\theta_{T_j}\mathscr{T})_k = T_{j+k} - T_j$). Write further $N(a,b] = N(b) - N(a)$, $a < b$.

We write $\mathscr{Z} = (\mathscr{X}, \mathscr{T})$, $\theta_s\mathscr{Z} = (\theta_s\mathscr{X}, \theta_s\mathscr{T})$ and call $\mathscr{Z}$ *time–stationary* w.r.t. some probability measure $\mathbb{P}$ if the $\mathbb{P}$–distribution of $\theta_s\mathscr{Z}$ does not depend on s. Similarly, $\mathscr{Z}$ is *event–stationary* w.r.t. some $\mathbb{P}_0$ if the $\mathbb{P}_0$–distribution of $\theta_{T_k}\mathscr{Z}$ does not depend on k (we will adopt the convention that also $\mathbb{P}_0(T_0 = 0) = 1$ is required).

If $\mathscr{Z}$ is time–stationary, we define the intensity λ as the rate of events occuring per unit time,

$$\lambda \;=\; \frac{1}{h}\mathbb{E}N(t, t+h] \;=\; \frac{1}{h}\mathbb{E}\#\{i : T_i \in (t, t+h]\};$$

we shall assume in the following that $\lambda < \infty$. The interpretation is similar as for a Poisson process (except for certain independence properties!):

Proposition 6.2 *In the time-stationary case, λ does not depend on t and h, and one has $\lambda > 0$, $\mathbb{P}(T_0 = 0) = 0$ and*

$$\lim_{h\downarrow 0} \frac{\mathbb{P}(T_0 \le h)}{h} \;=\; \lim_{h\downarrow 0} \frac{\mathbb{P}(N_h \ge 1)}{h} \;=\; \lambda, \tag{6.2}$$

$$\lim_{h\downarrow 0} \frac{\mathbb{P}(T_1 \le h)}{h} \;=\; \lim_{h\downarrow 0} \frac{\mathbb{E}[N_h; N_h \ge 2]}{h} \;=\; 0. \tag{6.3}$$

Proof. For the first statement, define $\varphi(h) = \mathbb{E}N(t, t+h]$. By stationarity, $\varphi(h)$ does not depend on t and is additive, $\varphi(h_1 + h_2) = \varphi(h_1) + \varphi(h_2)$. Since φ is increasing, this implies $\varphi(h) = h\varphi(1)$ and the independence of λ on h. That $\lambda > 0$ follows by choosing h with $\mathbb{P}(0 < T_0 \le h) > 0$ (then $\mathbb{E}N(0, h] > 0$), and $\mathbb{P}(T_0 = 0) = 0$ follows since the probability of an arrival at t does not depend on t and therefore must be 0 (cf. also A2.3). The proof of (6.2), (6.3) is given below. □

Theorem 6.3 *Assume that $\mathscr{Z} = (\mathscr{X}, \mathscr{T})$ is time–stationary w.r.t. $\mathbb{P}$. Define a new probability measure by*

$$\mathbb{P}_0(\mathscr{Z} \in F) \;=\; \frac{1}{\lambda h}\, \mathbb{E} \sum_{i:\, t < T_i \le t+h} I(\theta_{T_i}\mathscr{Z} \in F). \tag{6.4}$$

Then $\mathbb{P}_0$ does not depend on the choice of t and $h > 0$. Furthermore, $\mathscr{Z}$ is event–stationary w.r.t. $\mathbb{P}_0$ and and $\mathbb{E}_0 T_1 = \lambda^{-1}$.

Proof. That $\mathbb{P}_0$ is a probability measure follows easily by checking that the r.h.s. of (6.4) has the relevant properties (e.g. σ–additivity) as a function of F, and the independence of t, h follows as in the proof of Proposition 6.2. Taking $F = \{T_0 = 0\}$ and using $(\theta_{T_i}\mathscr{T})_0 = T_i - T_i = 0$ yields $\mathbb{P}_0(T_0 = 0)$ $= 1$.

Now consider event–stationarity. Taking $t = 0$ in (6.4) yields

$$\begin{aligned}
&\mathbb{P}_0(\theta_{T_1}\mathscr{Z} \in F) \\
&= \; \mathbb{P}_0(\mathscr{Z} \in \theta_{T_1}^{-1}F) \;=\; \frac{1}{\lambda h}\,\mathbb{E}\sum_{i=1}^{N_h} I(\theta_{T_i}\mathscr{Z} \in \theta_{T_1}^{-1}F) \\
&= \; \frac{1}{\lambda h}\,\mathbb{E}\sum_{i=1}^{N_h} I(\theta_{T_{i+1}}\mathscr{Z} \in F) \;=\; \frac{1}{\lambda h}\,\mathbb{E}\sum_{i=2}^{N_h+1} I(\theta_{T_i}\mathscr{Z} \in F) \\
&\le \; \frac{1}{\lambda h}\Big(1 + \mathbb{E}\sum_{i=1}^{N_h} I(\theta_{T_i}\mathscr{Z} \in F)\Big) \;=\; \frac{1}{\lambda h} + \mathbb{P}_0(\mathscr{Z} \in F).
\end{aligned}$$

Letting $h \to \infty$ yields $\mathbb{P}_0(\theta_{T_1}\mathscr{Z} \in F) \le \mathbb{P}_0(\mathscr{Z} \in F)$. The proof of $\ge$ is similar (or follows by replacing F by F^c).

Finally, if $T_0(1) = \inf\{T_k : T_k > 1\}$, then $T_0(1) \overset{\mathscr{D}}{=} 1 + T_0$ and hence

$$\lambda \mathbb{E}_0 T_1 \;=\; \mathbb{E}\sum_{i=1}^{N_1}(T_{i+1} - T_i) \;=\; \mathbb{E}[T_0(1) - T_0;\, T_0 \le 1] \;=\; \mathbb{E}[T_0(1) - T_0] \;=\; 1.$$

□

The probability measure $\mathbb{P}_0$ in Theorem 6.3 is called the *Palm distribution* associated with $\mathbb{P}$. The traditional interpretation of $\mathbb{P}_0$ is as the conditional $\mathbb{P}$–distribution given an event at 0, i.e. $T_0 = 0$. In fact, it follows from (6.2), (6.3) that

$$\begin{aligned}\mathbb{P}_0(\mathscr{X} \in F) \;&=\; \lim_{h\downarrow 0}\frac{1}{\lambda h}\,\mathbb{E}\sum_{i:\,0<T_i\le h} I(\theta_{T_i}\mathscr{X} \in F)\\ =\;\lim_{h\downarrow 0}\frac{1}{\mathbb{P}(T_0\le h)}\,&\mathbb{E}I(\theta_{T_0}\mathscr{X}\in F, T_0 \le h) \;=\; \lim_{h\downarrow 0}\mathbb{P}(\theta_{T_0}\mathscr{X}\in F \,|\, T_0 \le h),\end{aligned}$$

which (at least from the intuitive/heuristical point of view) can be identified with $\mathbb{P}(\mathscr{X} \in F \,|\, T_0 = 0)$.

The following result is called the *Palm inversion formula* and shows how to retrieve $\mathbb{P}$ in terms of $\mathbb{P}_0$. Basically, the formula is just the same as the one defining the stationary distribution of a regenerative process.

Theorem 6.4 *Assume that $\mathscr{X} = (\mathcal{X}, \mathscr{T})$ is event–stationary w.r.t. $\mathbb{P}_0$. Define a new probability measure $\mathbb{P}'$ by*

$$\mathbb{P}'(\mathscr{X} \in F) \;=\; \frac{1}{k\mathbb{E}_0 T_1}\,\mathbb{E}_0\int_0^{T_k} I(\theta_t\mathscr{X} \in F)\,\mathrm{d}t\,. \tag{6.5}$$

Then $\mathbb{P}'$ does not depend on $k = 1, 2, \ldots$, and $\mathscr{X}$ is time–stationary w.r.t. $\mathbb{P}'$. If $\mathbb{P}_0$ is the Palm distribution of a time–stationary $\mathbb{P}$, then $\mathbb{P}' = \mathbb{P}$.

Proof. That $\mathbb{P}'$ is a probability measure follows just as before. Let

$$W = W(\mathscr{X}) = \int_0^{T_1} I(\theta_t\mathscr{X} \in F)\,\mathrm{d}t\,.$$

Recalling $\mathbb{P}_0(T_0 = 0) = 1$, we have

$$W(\theta_{T_k}\mathscr{X}) = \int_{T_k}^{T_{k+1}} I(\theta_t\mathscr{X} \in F)\,\mathrm{d}t$$

$\mathbb{P}_0$–a.s., which in conjunction with event–stationarity implies the desired independence of k.

For time–stationarity, we must prove that $\mathscr{X}$ and $\theta_s\mathscr{X}$ have the same $\mathbb{P}'$–distribution. But

$$\begin{aligned}\mathbb{P}'(\theta_s\mathscr{X} \in F) \;&=\; \frac{1}{k\mathbb{E}_0 T_1}\,\mathbb{E}_0\int_0^{T_k} I(\theta_{s+t}\mathscr{X} \in F)\,\mathrm{d}t\\ =\;\frac{1}{k\mathbb{E}_0 T_1}\,\mathbb{E}_0&\int_s^{T_k+s} I(\theta_t\mathscr{X} \in F)\,\mathrm{d}t\end{aligned}$$

$$\leq \frac{1}{k\mathbb{E}_0 T_1}\Bigl(s+\mathbb{E}_0\int_0^{T_k} I(\theta_t\mathscr{X}\in F)\,\mathrm{d}t\Bigr) = \frac{s}{k\mathbb{E}_0 T_1} + \mathbb{P}'(\mathscr{X}\in F).$$

Letting $k\to\infty$ yields $\mathbb{P}'(\theta_s\mathscr{X}\in F)\leq\mathbb{P}(\mathscr{X}\in F)$, and the proof of $\geq$ is similar.

Finally, assume that $\mathbb{P}_0$ is the Palm distribution of $\mathbb{P}$ and let $T_0(1)$ be as in the proof of Theorem 6.3. Then

$$\begin{aligned}\mathbb{P}'(\mathscr{X}\in F) &= \frac{1}{\mathbb{E}_0 T_1}\frac{1}{\lambda}\mathbb{E}\sum_{0<T_k\leq 1}\int_{T_k}^{T_{k+1}} I(\theta_t\mathscr{X}\in F)\,\mathrm{d}t\\ &= \mathbb{E}\int_{T_0}^{T_0(1)} I(\theta_t\mathscr{X}\in F)\,\mathrm{d}t = \mathbb{E}\int_0^1 I(\theta_t\mathscr{X}\in F)\,\mathrm{d}t = \mathbb{P}(\mathscr{X}\in F),\end{aligned}$$

where the two last steps both use the time–stationarity of $\mathbb{P}$. □

Proof of (6.2), (6.3). By the Palm inversion formula with $k=1$,

$$\begin{aligned}\mathbb{P}(0<T_0\leq h) &= \lambda\mathbb{E}_0\int_0^{T_1} I(0<T_1-t\leq h)\,\mathrm{d}t\\ &= \lambda\mathbb{E}_0\int_0^{T_1} I(0<v\leq h)\,\mathrm{d}v = \lambda\mathbb{E}_0(T_1\wedge h).\end{aligned}$$

But $(T_1\wedge h)/h$ is dominated by 1 and has limit 1 as $h\downarrow 0$ so that the expectation has limit 1, establishing the claimed asymptotics for $\mathbb{P}(0<T_0\leq h)$ in (6.2). Similarly,

$$\begin{aligned}\mathbb{P}(N(0,h]=1) &= \lambda\mathbb{E}_0\int_0^{T_1} I\bigl(0<T_1-t\leq h, T_2>t+h\bigr)\,\mathrm{d}t\\ &= \lambda\mathbb{E}_0\int_0^{T_1} I\bigl(T_1-T_2+h<v\leq h\bigr)\,\mathrm{d}v\\ &= \lambda\mathbb{E}_0\bigl[T_1\wedge h - T_1\wedge(T_1-T_2+h)\bigr]^+ = \lambda h+\mathrm{o}(h),\end{aligned}$$

where the last identity follows by the same dominated convergence argument (note that $(T_1-T_2+h)^+$ equals 0 for h small enough since $T_1<T_2$). Combining these estimates with the definition of λ yields $\mathbb{E}[N_h;N_h\geq 2]=\mathrm{o}(h)$, and the rest of the proof is then easy. □

We now turn to some first applications. The first result is to intensity–driven point processes (part (ii) will be used in XI.5; for Cox processes, see A3):

Proposition 6.5 *Consider a Cox process with time–stationary intensity function* $\{\beta(t)\}$ *with* $\lambda=\mathbb{E}\beta(t)<\infty$, *and write* $B(t)=\int_0^t\beta(s)\mathrm{d}s$. *Then:* (i) *the Palm distribution of* $\beta(0)$ *is given by* $\mathbb{E}_0 f(\beta(0))=\mathbb{E}\bigl[\beta(t)f(\beta(t))\bigr]/\lambda$; (ii) $\mathbb{E}_0 B^{-1}(a)=a/\lambda$ *for all* a.

Proof. Part (i) is straightforward from $\mathbb{P}(N(0,h]=1,\beta(0)\in A)=\mathbb{P}(\beta(0)\in A)\beta(0)h+\mathrm{o}(h)$ (use the a.s. continuity of $\beta(\cdot)$ at 0, cf. A2.3). For (ii), we use the representation $T_k = B^{-1}(T'_k)$ in terms of an independent Poisson process $\{N'_t\}$, cf. A3. Since T'_k is Erlang(k) with unit rate, this gives $\mathbb{E}_0 T_k = \int_0^\infty a^{k-1}/(k-1)!\, \mathrm{e}^{-a}\mathbb{E}_0 B^{-1}(a)\,\mathrm{d}a$. Multiplying by $(1-s)^{k-1}$ and summing over $k = 1, 2, \ldots$ gives

$$\int_0^\infty \mathrm{e}^{-sa}\mathbb{E}_0 B^{-1}(a)\,\mathrm{d}a \;=\; \sum_{k=1}^{\infty}(1-s)^{k-1}\mathbb{E}_0 T_k \;=\; \sum_{k=1}^{\infty}(1-s)^{k-1}\frac{k}{\lambda} \;=\; \frac{1}{\lambda s^2},$$

and $\mathbb{E}_0 B^{-1}(a) = a/\lambda$ follows by Laplace transform inversion. □

We next give a formal version of the rate conservation law, which was used at the intuitive level in V.3 to derive the stationary excess distribution F_0 of a renewal process and will be further applied to the $GI/G/1$ queue in X.3–4.

Theorem 6.6 *Let $\mathscr{Z} = (\mathscr{X}, \mathscr{T})$ be time–stationary, such that $\mathscr{X}$ is real–valued and a jump of $\mathscr{X}$ of size U_k occur at time T_k* [formally, $U_k = 0$ is not excluded], *and that $\{X_t\}$ is right differentiable and continuous on (T_k, T_{k+1}) with sample path derivative Y_t at t. Then $\lambda\mathbb{E}_0 U_0 + \mathbb{E}Y_0 = 0$ provided all expectations are finite.*

Proof. Just note that

$$0 \;=\; \mathbb{E}[X_1 - X_0] \;=\; \mathbb{E}\left[\sum_{i:\,0<T_i\le 1} U_i + \int_0^1 Y_t\,\mathrm{d}t\right] \;=\; \lambda\mathbb{E}_0 U_0 + \mathbb{E}Y_0. \quad □$$

We finally consider PASTA (Poisson Arrivals See Time Averages), which in the language of Palm theory means that the time– and event–stationary distributions coincide for systems driven by a Poisson input process $\{N_t\}$. By “driven by Poisson input” we mean that for any t the processes

$$\{N_s - N_t\}_{s\ge t} \quad \text{and} \quad \{(X_s, N_s)\}_{s<t} \tag{6.6}$$

are independent. In particular, X_{t-} is independent of the event A_t of a Poisson epoch at t. However, typically X_t and A_t are dependent: e.g. in classical examples such as queueing models with Poisson arrival of customers or shot–noise processes, an arrival at t typically triggers a jump of $\{X_t\}$.

Theorem 6.7 *Assume that $\mathscr{Z}$ is time–stationary w.r.t. $\mathbb{P}$, that $\mathscr{T}$ is a Poisson process with rate λ and that there is independence in* (6.6)*. Then $\mathbb{P}_0(X_{0-} \in F) = \mathbb{P}(X_0 \in F)$.*

Proof. Taking $h = 1$ in the definition of $\mathbb{P}_0$ and noting that an event in $[t, t+\mathrm{d}t)$ occurs w.p. $\lambda\,\mathrm{d}t$, we get

$$\mathbb{P}_0(X_{0-} \in F) \;=\; \frac{1}{\lambda}\int_0^1 \mathbb{P}\big(X_{t-} \in F \,\big|\, A_t\big)\lambda\,\mathrm{d}t \;=\; \int_0^1 \mathbb{P}(X_{t-} \in F)\,\mathrm{d}t$$

$$= \int_0^1 \mathbb{P}(X_t \in F)\,\mathrm{d}t \;=\; \mathbb{P}(X_0 \in F),$$

noting that by right–continuity and stationarity $X_{t-} = X_t$ $\mathbb{P}$–a.s. for any fixed t (cf. A2.3). □

Problems

6.1 Let $\mathscr{Z} = (\mathscr{X}, \mathscr{T})$ be time–stationary on $\mathbb{R}$ and $\Phi : \mathbb{R} \times (D \times \mathbb{R}^{\mathbb{Z}})$ be a bounded measurable functional. Show *Campbell's formula*

$$\mathbb{E} \sum_{k=-\infty}^{\infty} \Phi(T_k, \theta_{T_k} \mathscr{Z}) \;=\; \lambda \int_{-\infty}^{\infty} \mathbb{E}_0 \Phi(s, \mathscr{Z})\,\mathrm{d}s\,.$$

[*Hint:* Consider first the case where $\Phi(s, \mathscr{Z}) = I(t < s \le t+h)\Psi(\mathscr{Z})$.]

6.2 An event F is called time–shift invariant for $\mathscr{Z}$ if $\{\theta_t \mathscr{Z} \in F\}$ does not depend on t, and event-shift invariant if $\{\theta_{T_k} \mathscr{Z} \in F\}$ does not depend on k. Show that these two concepts are the same. [*Hint:* To see that event–shift invariance implies time-shift invariance, let $T_k \le t < T_{k+1}$ and use $\{\theta_t \mathscr{Z} \in F\} = \{\theta_{T_1}\theta_t \mathscr{Z} \in F\}$.] The set of events that are invariant in this sense is denoted by $\mathscr{I}$, the *invariant σ–field*, and a stationary process is ergodic if $\mathbb{P}(\mathscr{Z} \in I) = 0$ or 1 for $I \in \mathscr{I}$.

6.3 For a time–stationary process, define a new one $\mathscr{Z}^*$ by $\mathscr{Z}^* = \theta_U \mathscr{Z}$ where U is uniform on $(0, T)$ and independent. Show that $\mathscr{Z}^* \overset{\mathscr{D}}{=} \mathscr{Z}$ no matter T. Let next $\mathscr{Z}_0^* = \theta_V \mathscr{Z}$ where V is uniform on $\{T_0, \ldots, T_K\}$ and independent. Show using Birkhoff's ergodic theorem that $\mathscr{Z}_0^*$ has a limit $\mathbb{P}_0^*$ in distribution as $K \to \infty$, and that $\mathbb{P}_0^* = \mathbb{P}_0$ in the ergodic case but not in general.

6.4 Let $z(t)$ be the periodic extension of the function $x(t) = t$ for $0 \le t \le 1$, $= 1 - (t-1)/2$ for $1 \le t \le 3$, and $X_t = z(t+U)$ where U is uniform on $(0, 3)$. Let $\{T_k\}$ be the epochs of crossing of level $1/2$. Compute $\mathbb{P}_0(X_0' = a)$, $a = 1, -1/2$. Now define $\mathbb{P}_0^{\#}(\cdot) = \lim_{\epsilon \downarrow 0} \mathbb{P}(\cdot \mid X_0 \in (1/2 - \epsilon, 1/2 + \epsilon))$ and show that $\mathbb{P}_0^{\#} \ne \mathbb{P}_0$.

Notes Franken *et al.* (1982) is a classical reference for Palm theory. More recent expositions are in Daley and Vere–Jones (1988), Sigman (1995), Serfozo (1999), Rolski *et al.* (1999), Thorisson (2000) and Baccelli and Brémaud (2002).

Problem 6.3 shows that the evaluation of $\mathbb{P}_0$ as a limiting time average performed in Example 6.1 is only valid in the ergodic case. Sigman (1995) gives a careful study using $\mathbb{P}_0^*$ as the fundamental object, including an "empirical Palm inversion formula" showing how to retrieve $\mathbb{P}$ in terms of $\mathbb{P}_0^*$.

In Problem 6.4, $\mathbb{P}_0^{\#}$ is constructed from a conditioning based upon a vertical window (in the x–direction) rather than the horizontal one (in the t–direction) used in Palm theory. This is common in Gaussian processes (where also more sophisticated windows have been proposed); see e.g. the treatment of Rice's formula in Leadbetter *et al.* (1983) [Rice's formula gives the distribution of the sample path derivative X' at a level crossing under suitable sample path differentiability assumptions].

The rate conservation law is due largely to Miyazawa and surveyed in his (1994) paper; see also Sigman (1995). The literature on PASTA and its generalizations is extensive. See e.g. Wolff (1989) for the basic theory and Melamed and Yao (1995) for generalizations and similar results.

VIII
Random Walks

1 Basic Definitions

We consider a random walk $S_n = X_1 + \cdots + X_n$ $(S_0 = 0)$ where the X_n are i.i.d. with common distribution F. The case where F has support contained in a half–line $(-\infty, 0]$ or $[0, \infty)$ is to a large extent covered by renewal theory, and so we assume that supp(F) contains points of both positive and negative sign (in particular, F is nondegenerate). In statements concerning the mean $\mathbb{E}X$, it is understood that this is welldefined, i.e. that $\mathbb{E}X^+$ and $\mathbb{E}X^-$ are not both infinite (thus we may have $\mathbb{E}X = +\infty$ or $\mathbb{E}X = -\infty$). Also, in expressions such as $\mathbb{P}(S_\tau \in A)$ or $\mathbb{E}\mathrm{e}^{itS_\tau}$ with τ a stopping time it is understood that the integration is carried out on $\{\tau < \infty\}$ only (in contrast, $\sum_0^{\tau-1}$ may well be an infinite sum).

The relevance of random walks for queueing theory should already be clear from the discussion of Lindley processes in III.6. A main point was found there to be the study of the distribution of the maximum, but a number of further quantities are important, both as technical tools and because of other queueing interpretations (e.g. the so–called *ladder epochs* and *ladder heights*, a terminology arising from path decompositions to be discussed in Section 2, will be found in Ch. X to be closely related to busy periods and busy cycles). For the sake of easy reference, we start by a list of the functionals to play an important role in the following. For graphical illustrations, see Fig. 1.1 (path (b) has $M = 0$ and $\tau_+ = \infty$).

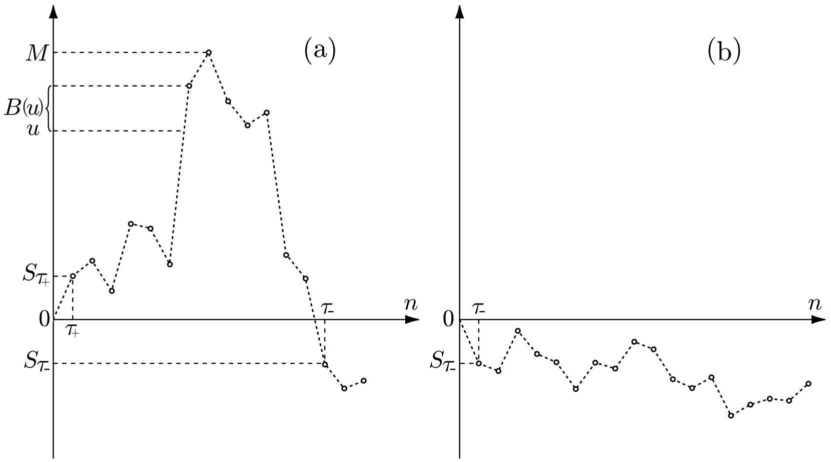

Figure 1.1

M_n the *partial maximum* $\max_{0\le k\le n} S_k$ of the first n partial sums.

M the (total) *maximum* $\sup_{0\le k<\infty} S_k$ (which may be infinite). Clearly, $M_n \uparrow M$ as $n \to \infty$.

τ_+ $= \tau_+^{\mathrm{s}} = \inf\{n \ge 1 : S_n > 0\}$, the first (strict) *ascending ladder epoch* or the *entrance time* to $(0,\infty)$. The distribution of τ_+ may be defective, i.e. $\mathbb{P}(\tau_+ = \infty) = \mathbb{P}(S_n \le 0 \text{ for all } n \ge 1) > 0$.

S_{τ_+} the first (strict) *ascending ladder height* (defined on $\{\tau_+ < \infty\}$ only).

G_+ the (strict) *ascending ladder height distribution* $G_+(x) = \mathbb{P}(S_{\tau_+} \le x)$. Here G_+ is concentrated on $(0,\infty)$ and may be defective, i.e. $\|G_+\| = \mathbb{P}(\tau_+ < \infty) < 1$.

τ_- $= \tau_-^{\mathrm{w}} = \inf\{n \ge 1 : S_n \le 0\}$ the first (weak) *descending ladder epoch* or the *entrance time* to $(-\infty, 0]$.

S_{τ_-} the first (weak) *descending ladder height* (defined on $\{\tau_- < \infty\}$ only).

G_- the (weak) *descending ladder height distribution* $G_-(x) = \mathbb{P}(S_{\tau_-} \le x)$. Here G_- is concentrated on $(-\infty, 0]$ and may be defective, i.e. $\|G_-\| = \mathbb{P}(\tau_- < \infty) < 1$.

$\tau(u)$ the time $\inf\{n \ge 1 : S_n > u\}$ of *first passage* to level $u \ge 0$ or the *entrance time* to (u, ∞). The distribution of $\tau(u)$ may be defective. Clearly, $\tau(0) = \tau_+$.

$B(u)$ the *overshoot* $S_{\tau(u)} - u$. Clearly, $B(0)$ is the ascending ladder height S_{τ_+}.

$B(\infty)$ a r.v. having the limiting distribution (if it exists) of $B(u)$.

It is seen that there is a slight asymmetry between positive and negative values, cf. the strict inequality $S_n > 0$ in the definition of τ_+ and the weak inequality $S_n \le 0$ in the definition of τ_-; this corresponds to different treatments of values $n > 0$ with $S_n = 0$ (if F has a density, this event has probability zero and the difference vanishes). Weak ascending and strict descending ladder epochs can be defined the obvious way by

$$\tau_+^{\mathrm{w}} = \inf\{n \ge 1 : S_n \ge 0\}, \quad \tau_-^{\mathrm{s}} = \inf\{n \ge 1 : S_n < 0\}.$$

The corresponding ladder heights are $S_{\tau_+^{\mathrm{w}}}$, $S_{\tau_-^{\mathrm{s}}}$ with distributions say G_+^{w}, G_-^{s}. These quantities may be needed in arguments involving sign changes, say if we want to study the minimum $\inf_{0\le k<\infty} S_k$ rather than the maximum M. Fortunately, a separate treatment can almost always be avoided by reference to the following result:

Proposition 1.1 *Define $\zeta = \mathbb{P}(S_{\tau_-} = 0) = \mathbb{P}(\tau_- \ne \tau_-^{\mathrm{s}})$. Then $\zeta < 1$ and $\zeta = \mathbb{P}(S_{\tau_+^{\mathrm{w}}} = 0) = \mathbb{P}(\tau_+ \ne \tau_+^{\mathrm{w}})$, and if δ_0 is the distribution degenerate at zero, then*

$$G_+^{\mathrm{w}} = \zeta\delta_0 + (1-\zeta)G_+, \quad G_- = \zeta\delta_0 + (1-\zeta)G_-^{\mathrm{s}}. \tag{1.1}$$

The proof is based upon a trivial but important observation:

Lemma 1.2 *Let n be fixed and define $S_k^* = S_n - S_{n-k} = X_{n-k+1} + \cdots + X_n$, $k = 0, \ldots, n$. Then $\{S_k^*\}_0^n \overset{\mathscr{D}}{=} \{S_k\}_0^n$.*

Proof of Proposition 1.1. For any n, we get by Lemma 1.2 that

$$\begin{aligned}\mathbb{P}(S_{\tau_+^{\mathrm{w}}} = 0, \tau_+^{\mathrm{w}} = n) &= \mathbb{P}(S_0^* = 0, S_k^* < 0, k = 1, \ldots, n-1, S_n^* = 0)\\ &= \mathbb{P}(S_0 = 0, S_n - S_{n-k} < 0, k = 1, \ldots, n-1, S_n = 0)\\ &= \mathbb{P}(S_0 = 0, S_\ell > 0, \ell = 1, \ldots, n-1, S_n = 0) = \mathbb{P}(S_{\tau_-} = 0, \tau_- = n)\end{aligned}$$

($\ell = n-k$) and $\zeta = \mathbb{P}(S_{\tau_+^{\mathrm{w}}} = 0)$ follows by summation over n. The relation (1.1) is obvious, and finally $1 - \zeta \ge \mathbb{P}(X_1 < 0) > 0$. □

We shall also need:

Lemma 1.3 *G_+ is lattice with span d if and only if F is so, and in particular nonlattice if and only if F is so. The same statement holds with G_+ replaced by any of G_+^{w}, G_-, G_-^{s}.*

Proof. It has to be shown that G_+ is concentrated on $\{0, \pm d, \pm 2d, \ldots\}$ if and only if F is so. The "if" part is obvious, and for the converse we may assume $d = 1$. That is, we have to show that if G_+ is concentrated on $\mathbb{N}$, then F is concentrated on $\mathbb{Z}$. Obviously, $\mathrm{supp}(F) \cap (0, \infty) \subseteq \mathrm{supp}(G_+) \subseteq \mathbb{N}$ so we have to show that $X_1 I(X_1 < 0) \in \mathbb{Z}$. But this follows since a path with $X_1 < 0$, $X_k > 0$, $k = 2, \ldots, \tau_+$ has positive probability and satisfies $n = S_{\tau_+} = X_1 + m$ for some $n, m \in \mathbb{N}$. The last assertion of the lemma follows by symmetry arguments and (1.1). □

Problems

1.1 Show that $\tau_+ \overset{\mathscr{D}}{=} T_0 + T_1 + \cdots + T_{\sigma-1}$ where $T_0, T_1, \ldots$ are independent, T_0 is distributed as τ_+^{w} given $S_{\tau_+^{\mathrm{w}}} > 0$, $T_1, T_2, \ldots$ are distributed as τ_+^{w} given $S_{\tau_+^{\mathrm{w}}} = 0$ and σ is independent of the T_k with $\mathbb{P}(\sigma = n) = \zeta^{n-1}(1-\zeta)$, $n = 1, 2, \ldots$.

Notes Random walks are of course one of the classical areas of probability theory and give rise to a broad spectrum of problems, of which the present treatment only covers a rather narrow range.

2 Ladder Processes and Classification

By iterating the definitions of τ_+, τ_-, we can define whole sequences $\{\tau_+(n)\}$, $\{\tau_-(n)\}$ of ladder epochs by $\tau_+(1) = \tau_+$, $\tau_-(1) = \tau_-$ and

$$\begin{aligned} \tau_+(n+1) &= \inf\left\{k > \tau_+(n) : S_k > S_{\tau_+(n)}\right\}, \\ \tau_-(n+1) &= \inf\left\{k > \tau_-(n) : S_k \le S_{\tau_-(n)}\right\}. \end{aligned}$$

The points in the plane of the form $\left(\tau_+(n), S_{\tau_+(n)}\right)$ are called the *ascending ladder points*. Similarly, the $\left(\tau_-(n), S_{\tau_-(n)}\right)$ are the *descending ladder points*, $\left\{S_{\tau_+(n)}\right\}$ is the *ascending ladder height process* and so on; see Fig. 2.1.

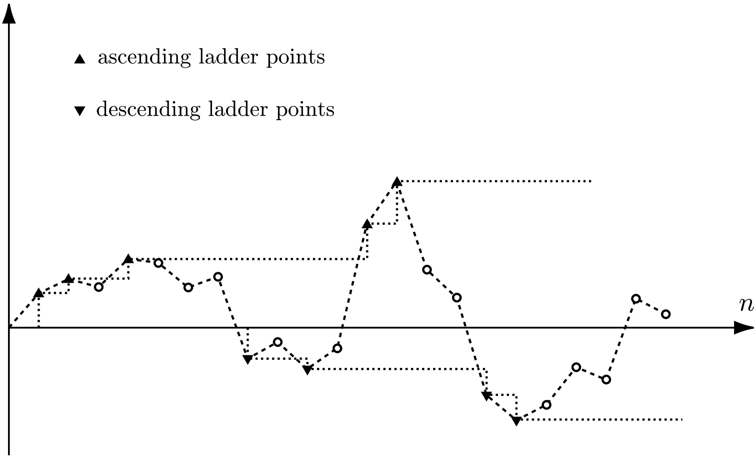

Figure 2.1

The importance of these concepts is due to the fact that the segments of the random walk between (say) ascending ladder points are just i.i.d. replicates. For example, the ascending ladder epoch process $\{\tau_+(n)\}$ is a discrete–time renewal process with governing probabilities $f_n = \mathbb{P}(\tau_+ = n)$, thus terminating if and only if $\|G_+\| < 1$. Similarly, the ascending ladder height process $\left\{S_{\tau_+(n)}\right\}$ is a renewal process governed by G_+, hence proper if and only if $\|G_+\| = 1$. Further, its forward recurrence time process is easily seen to coincide with the overshoot process $\{B(u)\}_{u\geq 0}$ of the random walk. Hence the renewal theorem V.4.6 and Lemma 1.3 yield the following result (the lattice version is obvious and omitted):

Theorem 2.1 *The overshoot $B(u)$ is proper if and only if the ascending ladder height process is nonterminating, i.e. $\|G_+\| = 1$. In that case, it holds as $u \to \infty$ that $B(u) \xrightarrow{\mathscr{D}} \infty$ if $\mathbb{E}S_{\tau_+} = \infty$, whereas if $\mathbb{E}S_{\tau_+} < \infty$ and F is nonlattice, then $B(u) \xrightarrow{\mathscr{D}} B(\infty)$ with $B(\infty)$ having density $(1 - G_+(x))/\mathbb{E}S_{\tau_+}$.*

Also, it is clear that the maximum M equals the lifetime $\sup\{S_{\tau_+(n)} : \tau_+(n) < \infty\}$ of the ascending ladder height process. Hence if we let $U_+ = \sum_0^\infty G_+^{*n}$ denote the corresponding renewal measure, V.2.9 yields the elementary but important:

Theorem 2.2 *The maximum M is finite if and only if $\|G_+\| < 1$. In that case, the distribution of M is the normalized ascending ladder height renewal measure $U_+/\|U_+\| = (1 - \|G_+\|)U_+$.*

The renewal measure U_+ is of basic importance in the following, and we proceed to give yet a third interpretation as the pre–τ_- occupation measure:

Theorem 2.3 (a)
$\mathbb{P}\big(S_n > S_k, k = 0,\ldots,n-1, S_n \in A\big) = \mathbb{P}\big(S_k > 0, k = 1,\ldots,n, S_n \in A\big)$*;*
(b) $U_+(A) = \mathbb{E}\sum_{n=0}^{\tau_- -1} I(S_n \in A)$, $U_-(A) = \mathbb{E}\sum_{n=0}^{\tau_+ -1} I(S_n \in A)$*;*
(c) $\mathbb{E}\tau_- = \|U_+\| = \big(1 - \|G_+\|\big)^{-1}$, $\mathbb{E}\tau_+ = \|U_-\| = \big(1 - \|G_-\|\big)^{-1}$.

Proof. Here (a) is an immediate consequence of Lemma 1.2, the first part of (b) is obtained by summing over n and using $\{S_k > 0, k = 1,\ldots,n\} = \{\tau_- > n\}$, and the first part of (c) follows by letting $A = [0,\infty)$ (or $\mathbb{R}$) in (b). The second parts of (b), (c) follow in a similar way by replacing $>$ by $\leq$ in (a). □

We next classify the random walk into several types. The first result is as follows:

Theorem 2.4 *For any random walk with F not degenerate at 0, one of the following possibilities occur:*
(i) (OSCILLATING CASE) *G_+ and G_- are both proper, $\|G_+\| = \|G_-\| = 1$ and $\overline{\lim}\, S_n = +\infty$, $\underline{\lim}\, S_n = -\infty$ a.s. Furthermore $\mathbb{E}\tau_+ = \mathbb{E}\tau_- = \infty$;*
(ii) (DRIFT TO $+\infty$) *G_+ is proper and G_- defective, and $S_n \to +\infty$ a.s. Furthermore $\mathbb{E}\tau_+ = \big(1 - \|G_-\|\big)^{-1} < \infty$;*
(iii) (DRIFT TO $-\infty$) *G_+ is defective and G_- proper, and $S_n \to -\infty$ a.s. Furthermore $\mathbb{E}\tau_- = \big(1 - \|G_+\|\big)^{-1} < \infty$.*

A sufficient condition for (i) *is* $\mathbb{E}X = 0$, *for* (ii) $\mathbb{E}X > 0$ *and for* (iii) $\mathbb{E}X < 0$.

Necessary and sufficient (though rather intractable) conditions covering also the case $\mathbb{E}X^+ = \mathbb{E}X^- = \infty$ are given in Corollary 4.4. Note that if say $-\infty < \mathbb{E}X < 0$, then Wald's identity applies and yields

$$1 - \|G_+\| \;=\; 1/\mathbb{E}\tau_- \;=\; \mathbb{E}X/\mathbb{E}S_{\tau_-}. \tag{2.1}$$

Proof. Since $\overline{\lim}\, S_n$ is exchangeable, we have by the Hewitt–Savage 0–1 law that $\overline{\lim}\, S_n = a$ for some constant $a \in [-\infty, \infty]$. If $|a| < \infty$, then $\overline{\lim}\,(S_n - X_1) = a$ would imply $a + X_1 \overset{\mathscr{D}}{=} a$, which is impossible. Similarly, $\mathbb{P}\bigl(|\underline{\lim}\, S_n| = \infty\bigr) = 1$, and hence indeed only the possibilites (i$'$) $\overline{\lim}\, S_n = \infty$, $\underline{\lim}\, S_n = -\infty$, (ii$'$) $S_n \to \infty$, (iii$'$) $S_n \to -\infty$ occur.

Since $S_{\tau_+} > 0$ and obviously $\overline{\lim}\, S_n = \overline{\lim}\, S_{\tau_+(n)}$, we see that $\overline{\lim}\, S_n = \infty$ if and only if the ascending ladder height renewal process is nonterminating, i.e. if and only if $\|G_+\| = 1$. Similarly, $\underline{\lim}\, S_n = -\infty$ if and only if $\|G^{\mathrm{s}}_-\| = 1$, i.e. if and only if $\|G_-\| = 1$ (Proposition 1.1). Noting that the expressions for $\mathbb{E}\tau_-$, $\mathbb{E}\tau_+$ are just Proposition 2.3(c), it is then seen that indeed (i) $\Longleftrightarrow$ (i$'$), (ii) $\Longleftrightarrow$ (ii$'$), (iii) $\Longleftrightarrow$ (iii$'$).

By the LLN $S_n/n \overset{\text{a.s.}}{\to} \mathbb{E}X$, it follows that $S_n > 0$ eventually if $\mathbb{E}X > 0$, and hence $\mathbb{E}X > 0 \Rightarrow$(ii); similarly $\mathbb{E}X < 0 \Rightarrow$(iii). To see that (i) holds if $\mathbb{E}X = 0$, suppose that we are, for example, in case (iii). Then $\mathbb{E}\tau_- < \infty$ and Wald's identity yields $\mathbb{E}S_{\tau_-} = 0$ which is impossible since $\mathbb{P}(S_{\tau_-} < 0) > 0$. □

Now define $U = \sum_0^\infty F^{*n}$ so that for any Borel set $A \subseteq \mathbb{R}$

$$U(A) \;=\; \sum_{n=0}^{\infty} \mathbb{P}(S_n \in A) \;=\; \mathbb{E}\sum_{n=0}^{\infty} I(S_n \in A) \tag{2.2}$$

is the expected number of visits of the random walk to A (which may of course be infinite). Sometimes the term "the renewal measure of the random walk" is used, but "occupation measure" seems more appropriate.

Lemma 2.5 *If F is nonlattice, then* $\mathrm{supp}(U) = \mathbb{R}$. *If F is aperiodic on $\mathbb{Z}$, then* $\mathrm{supp}(U) = \mathbb{Z}$.

Proof. Suppose first that F is nonlattice. Let $x \in \mathbb{R}$ be fixed and for a given $\epsilon > 0$, choose T such that $d\bigl(y, \mathrm{supp}(U_+)\bigr) < \epsilon$ for all $y \geq T$ (this is possible in view of Lemma 1.3 and V.5.1). Choose next $z < x - T$ with $z \in \mathrm{supp}(F^{*k})$ for some k and $u \in \mathrm{supp}(U_+)$ with $|u - (x - z)| < \epsilon$. Then $z + u$ is clearly in $\mathrm{supp}(U)$ so that $d\bigl(x, \mathrm{supp}(U)\bigr) < \epsilon$ which letting $\epsilon \downarrow 0$ proves that $x \in \mathrm{supp}(U)$. The lattice case is similar. □

Now if F is lattice, say aperiodic on $\mathbb{Z}$, we may view $\{S_n\}$ as a Markov chain on $\mathbb{Z}$, and irreducibility follows from Lemma 2.5. Hence by I.1.2 we have recurrence if $\sum_0^\infty \mathbb{P}(S_n = 0) = \infty$ and transience otherwise. Also in the nonlattice case a similar dichotomy holds:

Theorem 2.6 *For any nonlattice random walk, one of the following two possibilities occur:*
(i) (TRANSIENT CASE) *for any bounded interval J, we have $U(J) < \infty$ and $\mathbb{P}(S_n \in J \text{ i.o.}) = 0$. That is, $|S_n| \to \infty$ a.s.;*
(ii) (RECURRENT CASE) *for any nondegenerate interval J, we have $U(J) = \infty$ and $\mathbb{P}(S_n \in J \text{ i.o.}) = 1$.*

Proof. We shall show that (i′) $U(-\epsilon, \epsilon) < \infty$ for some $\epsilon > 0$, (ii′) $U(-\epsilon, \epsilon) = \infty$ for all $\epsilon > 0$ imply the conclusions of (i), resp. (ii). Define

$$I \;=\; (x - \epsilon/2, x + \epsilon/2), \quad \tau \;=\; \inf\{n : S_n \in I\}.$$

In case (i′), we have on $\{\tau < \infty\}$ that $(S_\tau - \epsilon, S_\tau + \epsilon)$ covers I. Thus the strong Markov property (see I.8.2) and the spatial homogeneity of the random walk implies that $U(I) \le U(-\epsilon, \epsilon) < \infty$. But then for any bounded interval J we have $U(J) < \infty$, since J can be covered by a finite number of intervals of length ϵ. Also, by the Borel–Cantelli lemma only a finite number of the events $A_n = \{S_n \in J\}$ can occur since $\sum \mathbb{P}(A_n) = U(J) < \infty$. Thus (i) holds.

In case (ii′), it suffices to show that $U(I) = \infty$ and $\mathbb{P}(S_n \in I \text{ i.o.}) = 1$ (since x, ϵ are arbitrary). By Lemma 2.5 we have $\mathbb{P}(\tau < \infty) > 0$. Define

$$\begin{aligned} q(\delta) &= \mathbb{P}\big(|S_n| < \delta \text{ for some } n \ge 1\big), \\ p_k(\delta) &= \mathbb{P}\big(|S_n| < \delta \text{ at least } k \text{ times}\big). \end{aligned}$$

Applying the strong Markov property at the time of the kth visit to $(-\delta, \delta)$ shows that $p_{k+1}(\delta) \le p_k(\delta)q(2\delta)$ which in conjunction with $\sum p_k(\delta) = U(-\delta, \delta) = \infty$ yields $q(2\delta) = 1$. But from $q(\delta) = 1$ for all $\delta > 0$ and the strong Markov property applied to τ one easily gets

$$\mathbb{P}\big(S_n \in I \text{ for at least two } n \,\big|\, \tau < \infty\big) \;=\; 1.$$

Repeating the argument, we get $\mathbb{P}(S_n \in I \text{ i.o.} \,|\, \tau < \infty) = 1$, i.e. $\mathbb{P}(S_n \in I \text{ i.o.}) = \mathbb{P}(\tau < \infty) > 0$. Since $\{S_n \in I \text{ i.o.}\}$ is an exchangeable event, the Hewitt–Savage 0–1 law implies $\mathbb{P}(S_n \in I \text{ i.o.}) = 1$ and therefore also $U(I) = \infty$. The lattice case is entirely similar. □

Corollary 2.7 *The random walk is transient if $\mathbb{E}X \ne 0$ and recurrent if $\mathbb{E}X = 0$ or, more generally, if the weak LLN $S_n/n \xrightarrow{\mathbb{P}} 0$ holds.*

Proof. That $\mathbb{E}X \ne 0$ implies transience is obvious since then S_n eventually has the same sign as $\mathbb{E}X$. Recalling the interpretation of U, we have

$$\sum_{k=-m}^{m-1} U[k, k+1] \;\ge\; U[-m, m],$$

$$U[k, k+1] \;\le\; U[-1, 1]\mathbb{P}(S_n \in [k, k+1] \text{ for some } n) \;\le\; U[-1, 1].$$

Hence letting $R_n = |S_n|/n$, we have for each K and m that

$$\begin{aligned} U[-1,1] &\geq \frac{1}{2m}U[-m,m] \geq \frac{1}{2m}\sum_{n=1}^{Km}\mathbb{P}(|S_n|<m) \\ &\geq \frac{1}{2m}\sum_{n=1}^{Km}\mathbb{P}(R_n<K^{-1}). \end{aligned}$$

But $R_n \xrightarrow{\mathbb{P}} 0$ implies $\sum_1^N \mathbb{P}(R_n < \epsilon)/N \to 1$ for all $\epsilon > 0$. Hence letting $m \to \infty$ yields $U[-1,1] \geq K/2$. Thus indeed $U[-1,1] = \infty$ and recurrence follows. □

Problems

2.1 Explain in the case $\mathbb{E}X < 0$ how the expression $(1 - \|G_+\|)U_+$ for the distribution of M is connected to Theorem 2.3(b),(c) and the basic formula VI.(1.1) for the limits of regenerative processes.

2.2 Let $Y_0^{(1)}$, $Y_0^{(2)}$ be initial delays for a discrete–time renewal process with infinite mean, i.e. $\mu = \sum_1^\infty nf_n = \infty$. Let $c = f_1 + \cdots + f_N$ with N chosen such that at least two terms are nonzero, and define

$$g_n = f_n/c,\ n \leq N, \quad h_n = f_n/(1-c),\ n > N.$$

Consider independent sequences $\{U_n^{(1)}\}$, $\{U_n^{(2)}\}$, $\{V_n\}$, $\{B_n\}$ of i.i.d. r.v.'s such that the $U_n^{(i)}$ are governed by $\{g_n\}$, the V_n by $\{h_n\}$ and $\mathbb{P}(B_n = 1) = 1 - \mathbb{P}(B_n = 0) = c$, and define

$$Y_n^{(i)} = B_n U_n^{(i)} + (1-B_n)V_n, \quad S_n^{(i)} = \sum_{k=0}^{n} Y_k^{(i)}.$$

Show that $\{S_n^{(1)}\}$, $\{S_n^{(2)}\}$ are renewal processes governed by $\{f_n\}$, and that $\sigma = \inf\{n : S_n^{(1)} = S_n^{(2)}\}$ is a.s. finite [this construction is known as the *Ornstein coupling*].

2.3 Show that if $\mathbb{E}X > 0$, then $\mathbb{E}\tau(u)/u \sim u/\mathbb{E}X$ as $u \to \infty$. [*Hint:* The elementary renewal theorem applied to the ascending ladder height renewal process.]

Notes The systematic use of ladder processes is largely due to Feller (1971).

3 Wiener–Hopf Factorization

The expressions given in Theorems 2.1 and 2.2 for the distributions of $B(\infty)$ and M indicate that it is of major importance to compute G_+ and, for symmetry reasons, G_-. This problem certainly does not appear to be easy, and in fact the known ways (presented in the next section) to represent G_+, G_- in terms of F without imposing additional conditions on F seem too complicated to be of much use. However, in some situations it is easy or at least possible to compute *one* of G_+, G_- (we study some main examples in Section 5, but for a simple example take F concentrated on $\mathbb{Z}$ with $\mathbb{E}X \geq 0$, $\mathbb{P}(X \geq 2) = 0$; then obviously G_+ is the one–point distribution at 1). The set of formulas in the following main result then allows calculation of the other ladder height distribution and thereby the distributions of M, $B(\infty)$, etc.:

Theorem 3.1 (a) $F = G_+ + G_- - G_+ * G_-$.
(b) $U_- * F(A) \;=\; \begin{cases} G_+(A) & A \subseteq (0,\infty), \\ U_-(A) & A \subseteq (-\infty,0). \end{cases}$

In terms of ch.f.'s (or m.g.f.'s, when defined), we may rewrite (a) as

$$1 - \widehat{F} \;=\; (1-\widehat{G}_+)(1-\widehat{G}_-) \;=\; (1-\widehat{G}_+^{\mathrm{s}})(1-\widehat{G}_-^{\mathrm{w}}). \tag{3.1}$$

This formula (and some generalizations like Problem 4.2) is known in the literature as the *Wiener–Hopf factorization identity*, and we shall refer to (a) in the same way. We see that, knowing G_-, we can solve (3.1) for $\widehat{G}_+$. Alternatively, G_+ can be computed by (b), and this is frequently more appealing.

Proof. Consider $\mathbb{E}\sum_0^{\tau_+} I(S_n \in A)$, $A \subseteq \mathbb{R}$. By splitting into the contributions from $n = 0, \ldots, \tau_+ - 1$ and $n = \tau_+$, the sum becomes $R_+(A) + G_+(A)$ where $R_+(A) = \mathbb{E}\sum_0^{\tau_+ - 1} I(S_n \in A)$. Splitting instead into the contributions from $\{n = 0\}$ and $\{n = 1, \ldots, \tau_+\}$, we get $\delta_0(A) + (R_+ * F)(A)$ where δ_0 is the distribution degenerate at 0 (for a formal proof, see I.3.3), and so, using that $R_+ = U_-$ by Theorem 2.3(b), we obtain

$$U_- + G_+ \;=\; \delta_0 + U_- * F. \tag{3.2}$$

Since $U_- * G_- = \sum_1^\infty G_-^{*n} = U_- - \delta_0$, convolving (3.2) with G_- to the left yields

$$U_- - \delta_0 + G_- * G_+ \;=\; G_- + U_- * F - F$$

, and subtracting this identity from (3.2) produces (a). For (b), evaluate (3.2) at A and note that $U_-(A) = \delta_0(A) = 0$ when $A \subseteq (0,\infty)$, $G_+(A) = \delta_0(A) = 0$ when $A \subseteq (-\infty,0)$. □

If $\zeta > 0$, then obviously some asymmetry is inherent in Theorem 3.1, and variants of the formulas may be required. For example:

Corollary 3.2 (a) $F = G_+^{\mathrm{w}} + G_-^{\mathrm{s}} - G_+^{\mathrm{w}} * G_-^{\mathrm{s}}$.
(b) $U_+ * F(A) = G_-(A)$, $A \subseteq (-\infty,0]$.

Proof. Sign reversion immediately yields (a). Convolving Theorem 3.1(a) with U_+ yields

$$U_+ * F \;=\; U_+ - \delta_0 + U_+ * G_- - (U_+ - \delta_0) * G_- \;=\; U_+ + G_- - \delta_0.$$

Since $U_+ - \delta_0 = \sum_1^\infty G_+^{*n}$ is concentrated on $(0,\infty)$, it follows that $(U_+ * F)(A) = G_-(A)$, $A \subseteq (-\infty,0]$. □

Notes For an alternative proof of Theorem 3.1(a), see Kennedy (1994). The original work of Wiener and Hopf is analytic in flavour and deals with solution of integral equations of the form

$$Z(x) \;=\; z(x) + \int_{-\infty}^{x} Z(x-y)\,\mu(\mathrm{d}y), \quad x \ge 0, \tag{3.3}$$

where μ is a measure. Note the $x \ge 0$ in (3.3) so that the equation is not just a renewal (convolution) equation on the whole line. Also, the Lindley equation in III.6 corresponds to $z \equiv 0$. Asmussen (1998c) gives references to the mathematical literature and studies existence, uniqueness and asymptotic properties of solutions to (3.3) by probabilistic methods in the special case where $z, Z \ge 0$ and F is a probability measure.

We return to Wiener–Hopf factorization in the setting of Markov additive processes in XI.2f. Another important further direction is Lévy processes $\{S_t\}_{t\ge 0}$. We do not treat this case but refer to Bingham (1976), Bertoin (1996) and Sato (1999); the flavour is somewhat different because say $\tau_+ = \inf\{t > 0 : S_t > 0\}$ is zero for an abundance of interesting special cases, and so there are no nontrivial analogues of G_+, G_-.

Problems

3.1 By Theorem 2.3, one can rewrite (3.1) as $(1-\widehat{F})^{-1} = \widehat{U}_+\widehat{R}_+$ where R_+ is the pre–τ_+ occupation measure. Give the probabilistic interpretation of this formula when $\mathbb{E}X \neq 0$ and thereby an alternative short proof of (3.1).
3.2 Find the distribution of $\inf\{S_1, S_2, \ldots\}$. [*Hint:* The forms on $(-\infty, 0]$ and $(0,\infty)$ are very different.]
3.3 Show that $\mathbb{P}(M > x)$, $\mathbb{E}[\tau(x);\, \tau(x) < \infty]$, $\mathbb{E}[\mathrm{e}^{\alpha\tau(x)};\, \tau(x) < \infty]$ and $\mathbb{P}(B(x) \le y)$ all satisfy Wiener–Hopf integral equations of the form (3.3).
3.4 Let $U = \sum_0^\infty F^{*n}$ be as in (2.2). Show that $U(x, x+a]$ is bounded by $c_1 + c_2 a$ and has limit $a/\mathbb{E}X$ as $x \to \infty$ when $\mathbb{E}X > 0$ and F is nonlattice.

4 The Spitzer–Baxter Identities

The theory to be presented is a classical cornerstone of probability theory as a whole and an instructive example of both the merits and the deficits of transform methods. To illustrate the scope and flavour, we state two of the main results:

Theorem 4.1 *For $|r| < 1$ and $t \in \mathbb{R}$*

$$1 - \mathbb{E}\bigl[r^{\tau_+}\mathrm{e}^{\mathrm{i}tS_{\tau_+}}\bigr] \;=\; \exp\Bigl\{-\sum_{n=1}^{\infty}\frac{r^n}{n}\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\, S_n > 0\bigr]\Bigr\} \tag{4.1}$$

[recall that in expressions like the l.h.s. of (4.1), the integration is understood to be carried out on $\{\tau_+ < \infty\}$ only].

Theorem 4.2 (SPITZER'S IDENTITY) *For* $|r| < 1$ *and* $t \in \mathbb{R}$

$$\sum_{n=0}^{\infty} r^n \mathbb{E}e^{\mathrm{i}tM_n} \;=\; \exp\Bigl\{\sum_{n=1}^{\infty} \frac{r^n}{n} \mathbb{E}e^{\mathrm{i}tS_n^+}\Bigr\} \tag{4.2}$$

and, provided $M < \infty$ *a.s.,*

$$\mathbb{E}e^{\mathrm{i}tM} \;=\; \exp\Bigl\{\sum_{n=1}^{\infty} \frac{1}{n}\bigl(\mathbb{E}e^{\mathrm{i}tS_n^+} - 1\bigr)\Bigr\}. \tag{4.3}$$

It is seen that in a certain sense a complete solution of the random walk problems has been provided: by uniqueness theorems for transforms, the distributions of (τ_+, S_{τ_+}), M_n, M are in principle determined by (4.1)–(4.3), and the solutions are also explicit in the sense that knowing F, we can in principle also evaluate the distribution F^{*n}, thereby expressions such as $\mathbb{E}e^{\mathrm{i}tS_n^+}$ and finally by summation the desired transforms. However, the weaknesses of the approach should also be apparent. As an example, one needs only to think of the $M/M/1$ waiting time W, where $W \overset{\mathscr{D}}{=} M$ with F doubly exponential (cf. III.6), and it was found in III.9 that W is exponential. The simplicity of this result should be compared with the effort required to compute first the F^{*n} and next (4.3), and it is strongly indicated that for even only slightly more general $GI/G/1$ queues the computational difficulties are formidable.

Proof of Theorem 4.1. We let r be fixed throughout and define

$$\begin{aligned}
\beta_n(t) &= \mathbb{E}\bigl[e^{\mathrm{i}tS_n}; \tau_+ = n\bigr], \quad \gamma_n(t) = \mathbb{E}\bigl[e^{\mathrm{i}tS_n}; \tau_+ > n\bigr],\\
\beta(t) &= \sum_{n=1}^{\infty} r^n \beta_n(t) \;=\; \mathbb{E}\bigl[r^{\tau_+} e^{\mathrm{i}tS_{\tau_+}}\bigr], \quad \gamma(t) = \sum_{n=1}^{\infty} r^n \gamma_n(t).
\end{aligned}$$

With $\widehat{F}$ the characteristic function of F, we then have

$$\beta_n(t) + \gamma_n(t) \;=\; \mathbb{E}\bigl[e^{\mathrm{i}tS_n}; \tau_+ \geq n\bigr] \;=\; \widehat{F}[t]\gamma_{n-1}(t),$$

and since $\gamma_0 = 1$, it follows by multiplying by r^n and summing that $\beta + \gamma = r(1+\gamma)\hat{F}$. Equivalently, $1 - r\widehat{F} = (1-\beta)/(1+\gamma)$, and taking logarithms and expanding we get

$$\sum_{n=1}^{\infty} \frac{r^n}{n} \widehat{F}^n \;=\; \sum_{n=1}^{\infty} \frac{\beta^n}{n} - \sum_{n=1}^{\infty} (-1)^n \frac{\gamma^n}{n} \tag{4.4}$$

if r is so small, say $|r| < r_0$, that $|\beta(t)| < 1$ and $|\gamma(t)| < 1$ for all t. Obviously, β and γ are the characteristic functions of bounded measures $\widetilde{\varphi}, \widetilde{\psi}$ supported by $(0,\infty)$, resp. $(-\infty, 0]$. Thus also $\varphi = \sum_1^\infty \widetilde{\varphi}^{*n}/n$ is supported by $(0,\infty)$ and $\psi = \sum_1^\infty (-1)^n \widetilde{\psi}^{*n}/n$ by $(-\infty, 0]$, and we may rewrite (4.4) as $\widehat{H} = \widehat{\varphi} - \widehat{\psi}$ where $H = \sum_1^\infty r^n F^{*n}/n$. By the uniqueness theorem, it therefore follows that H and φ coincide on $(0,\infty)$. Taking transforms yields

$$\sum_{n=1}^{\infty}\frac{r^n}{n}\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\,S_n>0\bigr] \;=\; \widehat{\varphi}(t) \;=\; \sum_{n=1}^{\infty}\frac{\beta^n(t)}{n} \;=\; -\log\bigl(1-\beta(t)\bigr)$$

which is the same as (4.1). The truth of (4.1) for $r_0\le|r|<1$ follows by an analytic continuation argument. □

By just the same argument:

Corollary 4.3 *The formula* (4.1) *remains valid if the pair of qualifiers* $(\tau_+, S_n>0)$ *is replaced by any of* $(\tau_+^{\mathrm{w}}, S_n\ge 0)$, $(\tau_-, S_n\le 0)$, $(\tau_-^{\mathrm{s}}, S_n<0)$.

Corollary 4.4 $1-\mathbb{E}r^{\tau_+} \;=\; \exp\Bigl\{-\sum_{n=1}^{\infty}\frac{r^n}{n}\mathbb{P}(S_n>0)\Bigr\}$, $|r|<1$.
Furthermore

$$\frac{1}{\mathbb{E}\tau_-} \;=\; 1-\|G_+\| \;=\; \exp\Bigl\{-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{P}(S_n>0)\Bigr\} \tag{4.5}$$

and the assertions (i) $S_n\overset{\text{a.s.}}{\to}-\infty$, (ii) $M<\infty$, (iii) $\|G_+\|<1$ *and* (iv) $\sum_1^{\infty}\mathbb{P}(S_n>0)/n<\infty$ *are equivalent.*

Proof. The first statement follows just by letting $t=0$ in (4.1). The first identity in (4.5) has been shown in Theorem 2.4, and the second follows from $\|G_+\|=\lim_{r\uparrow 1}\mathbb{E}r^{\tau_+}$. Finally the last statement follows from Theorem 2.4 and (4.5). □

Proof of (4.2). Define

$$\begin{aligned} A_{n,k} \;&=\; \bigl\{S_k=M_n;\,S_\ell<M_n,\,\ell=0,\dots,k-1\bigr\},\quad k\le n,\\ \psi_n(t) \;&=\; \mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\,A_{n,0}\bigr] \;=\; \mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\,M_n=0\bigr]. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tM_n}\exp\{\mathrm{i}u(S_n-M_n)\}\bigr] \;&=\; \sum_{k=0}^{n}\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_k}\exp\{\mathrm{i}u(S_n-S_k)\}\,;\,A_{n,k}\bigr]\\ &=\; \sum_{k=0}^{n}\psi_{n-k}(u)\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_k};\,A_{k,k}\bigr]. \end{aligned}$$

Letting $u=0$ we obtain

$$\sum_{n=0}^{\infty}r^n\mathbb{E}\mathrm{e}^{\mathrm{i}tM_n} \;=\; \sum_{n=0}^{\infty}r^n\psi_n(0)\cdot\sum_{k=0}^{\infty}r^k\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_k};\,A_{k,k}\bigr] \;=\; A_1\cdot A_2(t)$$

(say). Here

$$A_1 \;=\; \sum_{n=0}^{\infty}r^n\mathbb{P}(M_n=0) \;=\; \sum_{n=0}^{\infty}r^n\mathbb{P}(\tau_+>n) \;=\; \frac{1}{1-r}(1-\mathbb{E}r^{\tau_+})$$

$$
\begin{aligned}
&= \exp\Bigl\{\sum_{n=1}^{\infty}\frac{r^n}{n}\bigl(1-\mathbb{P}(S_n>0)\bigr)\Bigr\} \text{ (using Corollary 4.4)},\\
A_2(t) &= \sum_{n=0}^{\infty}\mathbb{E}\bigl[r^{\tau_+(n)}\mathrm{e}^{\mathrm{i}tS_{\tau_+(n)}}\bigr] = \sum_{n=0}^{\infty}\Bigl(\mathbb{E}\bigl[r^{\tau_+}\mathrm{e}^{\mathrm{i}tS_{\tau_+}}\bigr]\Bigr)^n\\
&= \Bigl(1-\mathbb{E}\bigl[r^{\tau_+}\mathrm{e}^{\mathrm{i}tS_{\tau_+}}\bigr]\Bigr)^{-1} = \exp\Bigl\{\sum_{n=1}^{\infty}\frac{r^n}{n}\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\,S_n>0\bigr]\Bigr\}
\end{aligned}
$$

and the proof is completed by observing that

$$\mathbb{P}(S_n\le 0)+\mathbb{E}[\mathrm{e}^{\mathrm{i}tS_n};\,S_n>0] = \mathbb{E}\mathrm{e}^{\mathrm{i}tS_n^+}. \qquad \square$$

Proof of (4.3). If $M<\infty$, then (iv) of Corollary 4.4 permits us to let $r\uparrow 1$ and use dominated convergence in (4.1) to get

$$1-\widehat{G}_+[t] = 1-\mathbb{E}\mathrm{e}^{\mathrm{i}tS_{\tau_+}} = \exp\Bigl\{-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{E}\bigl[\mathrm{e}^{\mathrm{i}tS_n};\,S_n>0\bigr]\Bigr\} \tag{4.6}$$

and since the characteristic function of M is $\bigl(1-\widehat{G}_+[0]\bigr)/\bigl(1-\widehat{G}_+[t]\bigr)$ by Theorem 2.2, (4.3) follows easily. □

In the further development of the theory, one discovers that a certain care is needed to give rigorous proofs of expected results. For example, one might ask whether (4.6) also holds when $\|G_+\|=1$, whether, say, the Laplace transform is obtained by replacing it by $\beta<0$ and whether the expressions for the moments which come out by formal differentiation are correct. We shall not go into these points, but give a direct proof of a result of the last type:

Proposition 4.5 $\mathbb{E}M_n = \sum_{k=1}^{n}\frac{1}{k}\mathbb{E}S_k^+, \quad \mathbb{E}M = \sum_{k=1}^{\infty}\frac{1}{k}\mathbb{E}S_k^+ .$

Proof. Letting $F_n=\{S_n>0\}$, $G_n=\{M_n>0, S_n\le 0\}$ we have with $K=\max\{S_k-S_1 : k=1,\dots,n\}$ that

$$
\begin{aligned}
\mathbb{E}M_n &= \mathbb{E}[M_n;\,M_n>0] = \mathbb{E}[M_n;\,F_n]+\mathbb{E}[M_n;\,G_n]\\
&= \mathbb{E}[X_1;\,F_n]+\mathbb{E}[K;\,F_n]+\mathbb{E}[M_n;\,G_n].
\end{aligned}
$$

By symmetry arguments, the two first terms are $\mathbb{E}[S_n;\,F_n]/n=\mathbb{E}S_n^+/n$ and $\mathbb{E}[M_{n-1};\,F_n]$, respectively, whereas the last is

$$\mathbb{E}[M_{n-1};\,G_n] = \mathbb{E}\bigl[M_{n-1};\,M_{n-1}>0, S_n\le 0\bigr] = \mathbb{E}[M_{n-1};\,S_n\le 0].$$

Hence $\mathbb{E}M_n=\mathbb{E}S_n^+/n+\mathbb{E}M_{n-1}$, and the desired expression for $\mathbb{E}M_n$ follows by iteration. For $\mathbb{E}M$, let $n\uparrow\infty$ and use monotone convergence. □

Note that even simple conditions for $\mathbb{E}M$ to be finite are not at all apparent from Proposition 4.5. We return to the problem in X.2.

Problems

4.1 Show that (4.6) also holds if $\|G_-\| < 1$. [*Hint:* Find first $1 - \widehat{G}_-$ and use Wiener–Hopf factorization.]
4.2 Let $\widehat{g}_+(r,t)$ denote the expression (4.1) and $\widehat{g}_-(r,t)$ the same thing with τ_+ replaced by τ_-. Show that $1 - r\widehat{F}[t] = \widehat{g}_+(r,t)\widehat{g}_-(r,t)$.

Notes The results of this section were found by Baxter and Spitzer around 1960. Good references are Feller (1971, Chs. XII and XVIII), Chung (1974), Spitzer (1976), Woodroofe (1982) and Siegmund (1985). For a more recent study in the area, see Grübel (1988).

5 Explicit Examples. $M/G/1$, $GI/M/1$, $GI/PH/1$

5a. Lattice Distributions
5b. Skip–Free Distributions. $M/G/1$ and $GI/M/1$
5c. Distributions with a Rational Laplace Transform
5d. Phase–Type Distributions. $GI/PH/1$

We consider a random walk with $\mu = \mathbb{E}X < 0$, as is the typical case in queueing theory (the case $\mu > 0$ is essentially symmetric, whereas some modifications may be required when $\mu = 0$). We shall in some cases be satisfied by evaluating either G_+ or G_-, since it is then obvious how to proceed for, say, the distribution of M.

The following simple observation is often the key:

Lemma 5.1 *Let $F^{(z)}$ denote the distribution of the overshoot $X \mid X > z$ over z, $\overline{F}^{(z)}(x) = \overline{F}(x+z)/\overline{F}(z)$. Then G_+ is a* (defective) *mixture of the $F^{(z)}$, $z > 0$. If $X = U - T$ with $U, T > 0$ independent with distributions A, resp. B, then G_+ is a* (defective) *mixture of the $B^{(z)}$.*

Proof. Conditioning upon $X_1, \ldots, X_{n-1}$ gives for $x > 0$ that

$$\begin{aligned}
\mathbb{P}(S_{\tau_+} > x) &= \sum_{n=1}^{\infty} \mathbb{P}(S_{\tau_+} > x, \tau_+ = n) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\big[\overline{F}(x - S_{n-1});\ \tau_+ > n-1\big] \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left[\overline{F}(-S_{n-1})\overline{F}^{(-S_{n-1})}(x);\ \tau_+ > n-1\right].
\end{aligned}$$

Hence $G_+ = \int_0^\infty F^{(z)}\,\nu(\mathrm{d}z)$ where

$$\nu(A) \;=\; \sum_{n=1}^{\infty} \mathbb{E}\big[\overline{F}(-S_{n-1});\ \tau_+ > n-1, -S_{n-1} \in A\big], \quad A \subseteq (0,\infty).$$

For the last statement, just note that $F^{(z)}$ must be a mixture of the $B^{(y)}$, $y > z$. □

5a Lattice Distributions

We consider first the lattice case where w.l.o.g. we may assume F to be aperiodic on $\mathbb{Z}$ with point probabilities $f_k = \mathbb{P}(X = k)$. Then also G_+, G_- are concentrated on $\mathbb{Z}$, and we write $g_{+,k}, g_{-,k}$ for the point probabilities. In order for explicit results to come out, we will assume F to have bounded support in one direction, but one can in fact generalize to a discrete phase–type form in one direction along the lines of Sections 5c, 5d.

We first present a classical argument based on transforms, in this context the generating functions $\widehat{F}[s] = \mathbb{E}s^X$, $\widehat{G}_+[s]$, $\widehat{G}_-[s]$. Here $\widehat{F}[s]$ is always defined for $|s| = 1$ but may have larger domain, whereas $\widehat{G}_+[s]$ is always defined for s in the closed unit disc $\Delta = \{s \in \mathbb{C} : |s| \le 1\}$ and $\widehat{G}_-[s]$ for $s^{-1} \in \Delta$.

The essence of the argument is to recognize the form of either G_+ or G_- by a probabilistic argument (cf. Lemma 5.1) and identifying some key constants via the relation between the roots of the equations $\widehat{F}[s] = 1$, $\widehat{G}_+[s] = 1$, $\widehat{G}_-[s] = 1$ provided by the Wiener–Hopf factorization identity (3.1) stating that $1 - \widehat{F}[s] = (1 - \widehat{G}_+[s])(1 - \widehat{G}_-[s])$. Here we shall say that the equation $\varphi(s) = 0$ (where $\varphi : D \to \mathbb{C}$ is a continuous function on a complex domain $D \subseteq \mathbb{C}$) has the roots $\alpha_1, \ldots, \alpha_r$ in $D_1 \subseteq D$ if $\varphi(s) = (z - \alpha_1) \cdots (z - \alpha_r)\psi(z)$ with $\alpha_1, \ldots, \alpha_r \in D_1$ and ψ continuous and nonzero on D_1. Note that some α_i may coincide, corresponding to multiple roots.

Theorem 5.2 *Assume that $f_r > 0$, $f_{r+1} = f_{r+2} = \cdots = 0$ for some $r = 1, 2, \ldots$. Then the equation $\widehat{F}[s] = 1$ has exactly r roots $\alpha_1, \ldots, \alpha_r \in \mathbb{C}\backslash\Delta$ outside the unit circle, and these determine G_+ by means of*

$$1 - \widehat{G}_+[s] \;=\; \left(1 - \frac{s}{\alpha_1}\right) \cdots \left(1 - \frac{s}{\alpha_r}\right). \tag{5.1}$$

Proof. Clearly, G_+ is concentrated on $\{1, \ldots, r\}$ with $g_{+,r} \ge f_r > 0$. Thus $\widehat{G}_+[s]$ is a polynomial of degree r with $\widehat{G}_+[0] = 0$ so that we may write $1 - \widehat{G}_+[s]$ in the form (5.1). Further, for $\alpha \in \Delta$ we have $\left|\widehat{G}_+[\alpha]\right| \le \mathbb{P}(\tau_+ < \infty) < 1$, so that taking $s = \alpha_k$ in (5.1) shows that $\alpha_k \notin \Delta$.

The factorization identity (3.1) now takes the form

$$1 - \widehat{F}[s] \;=\; \left(1 - \widehat{G}_-[s]\right) \prod_{k=1}^{r} (1 - s/\alpha_k).$$

Since G_- is concentrated on $\{0, -1, -2, \ldots\}$, $1 - \widehat{G}_-[s]$ is finite, continuous and nonzero in $\mathbb{C}\backslash\Delta$. This shows the assertion concerning the roots of $\widehat{F}[s] = 1$. □

Theorem 5.3 *Assume that* $f_{-r} > 0$, $f_{-r-1} = f_{-r-2} = \cdots = 0$ *for some* $r = 1, 2, \ldots$. *Then the equation* $\widehat{F}[s] = 1$ *has exactly* r *roots* $\alpha_1, \ldots, \alpha_r \in \Delta$ *in the unit disc, one of which is* 1 *and the rest are inside the unit circle, say* $\alpha_r = 1$, $\alpha_1, \ldots, \alpha_{r-1} \in \operatorname{int}(\Delta)$. *These roots determine* G_- *by means of*

$$1 - \widehat{G}_-[s] \;=\; \frac{(-1)^{r+1} f_{-r}}{\alpha_1 \cdots \alpha_{r-1}} \Bigl(1 - \frac{1}{s}\Bigr)\Bigl(1 - \frac{\alpha_1}{s}\Bigr) \cdots \Bigl(1 - \frac{\alpha_{r-1}}{s}\Bigr). \qquad (5.2)$$

Proof. Clearly, G_- is concentrated on $\{0, -1, \ldots, -r\}$ with $g_{-,-r} = f_{-r} > 0$ (note that $S_{\tau_-} = -r$ can only occur if $X_1 = -r$). Thus $\widehat{G}_-[s]$ is a polynomial of degree r in $1/s$ with coefficient f_{-r} to s^{-r} so that we may write $1 - \widehat{G}_-[s]$ in the form (5.2); that one α_k is 1 follows since G_- has mass $\widehat{G}_-[1] = 1$, and that not more than one is has absolute value 1 follows by aperiodicity. The rest of the proof is similar to that of Theorem 5.2. □

We next present a martingale approach, going back to Wald, which gives the point probabilities of G_+, G_- using matrix inversion rather than transform inversion.

Theorem 5.4 *In the set–up of Theorem 5.2,*

$$\begin{pmatrix} g_{+,1} \\ \vdots \\ g_{+,r} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^r \\ \vdots & & & \vdots \\ \alpha_r & \alpha_r^2 & \ldots & \alpha_r^r \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \qquad (5.3)$$

Proof. From $\widehat{F}[\alpha_k] = 1$ it follows (cf. III.8.8–8.9) that $\bigl\{\alpha_k^{S_n}\bigr\}$ is a (complex–valued) martingale. Letting $g_{+,j,n} = \mathbb{P}(S_{\tau_+} = j, \tau_+ \le n)$, optional stopping at $\tau_+ \wedge n$ yields

$$1 \;=\; \alpha_k^{S_0} \;=\; \mathbb{E}\alpha_k^{S_{\tau_+ \wedge n}} \;=\; \mathbb{E}\bigl[\alpha_k^{S_n};\, \tau_+ > n\bigr] + \sum_{j=1}^{r} \alpha_k^j g_{+,j,n}.$$

However, since $S_n \le 0$ on $\{\tau_+ > n\}$, it follows from $|\alpha_k| > 1$, $S_n \overset{\text{a.s.}}{\to} -\infty$ and dominated convergence that the first term converges to zero and since $g_{+,j,n} \uparrow g_{+,j}$, we get $1 = \sum_1^r \alpha_k^j g_{+,j}$. The solution of the r linear equations obtained by letting k vary is indeed as asserted. □

5b Skip–Free Distributions. $M/G/1$ and $GI/M/1$

A particular simple and important case is the skip–free one. We say that $\{S_n\}$ is *upward skip–free* or *right–continuous* if $r = 1$ in the setting of Theorem 5.2, and *downward skip–free* or *left–continuous* if $r = 1$ in Theorem 5.3.

Corollary 5.5 *In the upward skip–free lattice case:*
(a) G_+ *is concentrated at 1. The point mass* $\theta = \|G_+\|$ *can be evaluated as*

the unique solution < 1 *of*

$$1 \;=\; \widehat{F}[1/\theta] \;=\; \frac{1}{\theta}f_1 + f_0 + \theta f_{-1} + \theta^2 f_{-2} + \cdots; \tag{5.4}$$

(b) *The distribution of* M *is geometric with parameter* θ*,* $\mathbb{P}(M = n) = (1-\theta)\theta^n$*,* $n = 0, 1, \ldots$*;*
(c) G_- *is given by the point probabilities* $g_{-,n} = \sum_{-\infty}^{n} \theta^{n-k} f_k$*,* $n = 0, -1, -2, \ldots$

Proof. Part (a) follows immediately either from Theorem 5.2 (use $\|G_+\| = \hat{G}_+[1] = 1/\alpha_1$) or from Theorem 5.4 (use $g_{+,1} = 1/\alpha_1$). Then $U_+ = \sum_0^\infty G_+^{*n}$ has point mass θ^n at n, from which (b) follows by Theorem 2.2 and (c) by $G_- = (U_+ * F)\big|_{(-\infty,0]}$, cf. Corollary 3.2(b), which yields

$$g_{-,n} \;=\; \sum_{\ell=0}^{\infty} \theta^\ell f_{n-\ell} \;=\; \sum_{k=-\infty}^{n} \theta^{n-k} f_k, \quad n = 0, -1, -2, \ldots. \qquad \square$$

Corollary 5.6 *In the downward skip–free lattice case:*
(a) G_- *is concentrated at* $\{-1, 0\}$ *with point probabilities* $g_{-,-1} = f_{-1}$*,* $g_{-,0} = 1 - f_{-1}$*.*
(b) G_+ *is given by the point probabilities* $g_{+,n} = r_n/f_{-1}$*,* $n = 1, 2, \ldots$*, where* $r_n = f_n + f_{n+1} + \cdots$*; furthermore* $\|G_+\| = 1 + \mu/f_{-1}$*.*
(c) *The point probabilities* $\nu_n = \mathbb{P}(M = n)$ *are given by* $\nu_0 = -\mu/f_{-1}$*,*

$$\nu_n \;=\; \frac{r_n}{f_{-1}}\nu_0 + \frac{r_{n-1}}{f_{-1}}\nu_1 + \cdots + \frac{r_1}{f_{-1}}\nu_{n-1}, \quad n \geq 1.$$

Proof. That $g_{-,n} = 0$ for $n = -2, -3, \ldots$ is clear by left–continuity, which also implies that $S_{\tau_-} = -1$ can only occur if $X_1 = -1$. This shows (a). Noting that the expected number of visits of $\{S_{\tau_-(n)}\}$ to any $k = 0, -1, \ldots$ is geometric, we then obtain $U_-(\{k\}) = \sum_1^\infty n(1 - f_{-1})^{n-1} f_{-1} = 1/f_{-1}$, The first part of (b) then follows from $G_+ = (U_- * F)\big|_{(0,\infty)}$, cf. Theorem 3.1(b), and the second from (2.1) and $\mathbb{E}S_{\tau_-} = -f_{-1}$. For (c), it is now clear that $\nu_0 = 1 - \|G_+\| = -\mu/f_{-1}$, and the recursion formula for $\nu_1, \nu_2, \ldots$ follows since by Theorem 2.2 $\{\nu_n\}$ is proportional to the renewal sequence governed by the $\{g_{+,n}\}$, cf. I.(2.1). $\square$

An important example to which Corollary 5.5 applies is the number of customers in $GI/M/1$ just before arrivals, cf. III.6.2, and similarly Corollary 5.6 provides a road to the distribution of the number of customers in $M/G/1$ just after departures, cf. Problem III.6.3.

For continuous distributions, there may at a first sight appear to be no natural analogue of the concept of skip–freeness. However, if one thinks more specifically in terms of queues, the netput process of the $M/G/1$ workload process (a compound Poisson process with drift) has properties similar to a downward skip–free lattice random walk, and obviously its ladder height distribution is the same as that of the random walk driving the

Lindley process describing the waiting times. This motivates that one gets a simple solution in this context (the random walk setting can be generalized as in Problem 5.1 at the cost of somewhat more lengthy computations). Recall that B_0 is the stationary recurrence time distribution from renewal theory, i.e. the distribution with density $b_0(x) = \overline{B}(x)/\mu_B$.

Theorem 5.7 *Consider the $M/G/1$ queue with arrival intensity β, service time distribution B and $\rho = \beta\mu_B < 1$. Then:*
(a) *Consider a random walk with $X_n = U_n - T_n$ where U_n is the nth service time and T_n the nth interarrival time. Then G_+ is absolutely continuous with density $g_+(x) = \beta\overline{B}(x) = \rho b_0(x)$.*
(b) *In the steady state, the actual waiting time W and the workload V have a common distribution given by*

$$\mathbb{P}(W \le x) \;=\; \mathbb{P}(V \le x) \;=\; (1-\rho)\sum_{n=0}^{\infty} \rho^n B_0^{*n}(x). \tag{5.5}$$

Equivalently, $W \overset{\mathscr{D}}{=} V \overset{\mathscr{D}}{=} Y_1 + \cdots + Y_N$ where $N, Y_1, Y_2, \ldots$ are independent such that N is geometric with $\mathbb{P}(N = n) = (1-\rho)\rho^n$ and $Y_1, Y_2, \ldots$ are i.i.d. with common distribution B_0. In particular, writing $\mu_B^{(k)} = \mathbb{E}U_n^k$, the two first moments are

$$\mathbb{E}W \;=\; \mathbb{E}V \;=\; \frac{\rho\mu_B^{(2)}}{2(1-\rho)\mu_B}, \tag{5.6}$$

$$\mathbb{E}W^2 \;=\; \mathbb{E}V^2 \;=\; \frac{\rho\mu_B^{(3)}}{3(1-\rho)\mu_B} + \frac{\beta^2\mu_B^{(2)^2}}{2(1-\rho)^2}. \tag{5.7}$$

Proof. Sign reversion in Lemma 5.1 shows that G_- is exponential with intensity β, so that the descending ladder height renewal process is Poisson on $(-\infty, 0)$ (and as always has an epoch at 0). Therefore $U_-(\mathrm{d}x) = \delta_0 + \beta\mathrm{d}x$.

The form of X_n shows that F has a density $f(x)$ for $x > 0$ and that

$$\begin{aligned} f(x) &= \int_x^\infty \beta \mathrm{e}^{-\beta(y-x)}\, B(\mathrm{d}y), \\ \overline{F}(x) &= \int_x^\infty \left(1 - \mathrm{e}^{-\beta(y-x)}\right) B(\mathrm{d}y) \;=\; \overline{B}(x) - f(x)/\beta. \end{aligned}$$

Hence by Theorem 3.1(b), $g_+(x)$ is given by

$$U_- * f(x) \;=\; f(x) + \int_0^\infty f(x+z)\beta\,\mathrm{d}z \;=\; f(x) + \beta\overline{F}(x) \;=\; \beta\overline{B}(x)$$

for $x > 0$, showing (a). That $W \overset{\mathscr{D}}{=} V$ is a consequence of PASTA, see VII.6 (cf. also III.9.2), and (5.5) then follows from (a) and Theorem 2.2. Since the mean of B_0 is $\mu_B^{(2)}/2\mu_B$, (5.5) then gives

$$\mathbb{E}W \;=\; (1-\rho)\sum_{n=0}^{\infty} \rho^n n \mu_B^{(2)}/2\mu_B,$$

which is the same as (5.6). The proof of (5.7) is similar (though a little more lengthy) and left to the Problems. □

The formula (5.5) (or its transform version, Problem 5.2, or sometimes just (5.6)) is commonly referred to as the *Pollaczeck–Khinchine formula.*

Theorem 5.8 *Consider the $GI/M/1$ queue with service intensity δ, interarrival time distribution A and $\rho = (\delta\mu_A)^{-1} < 1$. Then:*
(a) *Consider a random walk with $X_n = U_n - T_n$ where U_n is the nth service time and T_n the nth interarrival time. Then the equation*

$$1 \;=\; \widehat{F}[\eta] \;=\; \mathbb{E}e^{\eta X_n} \;=\; \mathbb{E}e^{\eta(U_n - T_n)} \;=\; \frac{\delta}{\delta - \eta}\int_0^\infty e^{-\eta y}\,A(dy) \qquad (5.8)$$

has a unique positive solution. Further G_+ is defective exponential with intensity δ and total mass $\theta = 1 - \eta/\delta = \int_0^\infty e^{-\eta y}\,A(dy)$. That is, the density is $\theta\delta e^{-\delta x}$.
(b) *In the steady state, the distribution of the actual waiting time W is given by $\mathbb{P}(W > x) = \theta e^{-\eta x}$.*

Proof. The existence and uniqueness of η is easy. It follows from Lemma 5.1 that G_+ is defective exponential with intensity δ and mass say $\widetilde{\theta}$. The factorization identity (3.1) then means

$$0 \;=\; 1 - \widehat{F}[\eta] \;=\; \bigl(1 - \widehat{G}_+[\eta]\bigr)\bigl(1 - \widehat{G}_-[\eta]\bigr) \;=\; \Bigl(1 - \tilde{\theta}\frac{\delta}{\delta - \eta}\Bigr)\bigl(1 - \widehat{G}_-[\eta]\bigr).$$

Here $\widehat{G}_-[\eta] < 1$ since $\eta > 0$ so that the first factor on the r.h.s. must be 0 which yields $\widetilde{\theta} = (\delta - \eta)/\delta = \theta$, showing (a).

A ladder step terminates with intensity δ and is the last w.p. $1 - \theta$. Hence the failure rate of W given $W > 0$ is $\delta(1 - \theta) = \eta$, and since clearly $\mathbb{P}(W > 0) = \theta$, (b) follows. □

Problems

5.1 Assume F has a density on $(-\infty, 0]$ of the form $\alpha e^{\beta x}$ with $\alpha < \beta$. Show that G_- is exponential with intensity β and that $G_+(x) = F(x) + \beta[\mu - \int_x^\infty \overline{F}(y)\,dy]$, $x > 0$, $\|G_+\| = 1 + \mu\beta$.
5.2 Find the m.g.f. $\widehat{G}_+$ in $M/G/1$, both directly from Theorem 5.7(a) and by Wiener–Hopf factorization. Show hereby that the m.g.f. $\mathbb{E}e^{sW}$ is $(1 - \rho)s/(s - \beta + \beta\hat{B}[s])$.
5.3 Show (5.7).

5c *Distributions with a Rational Laplace Transform*

We now assume $X = U - T$ with U, T independent, U having distribution $B \in \mathscr{RLT}$ (cf. III.4) and T having a general distribution. We write the m.g.f. (not Laplace transform!) $\widehat{B}[s]$ of B as $p(s)/q(s)$ where p, q are polynomials without common roots and degree say d of q; we can assume w.l.o.g. that the coefficient to s^d in $q(s)$ is 1, $q(s) = s^d + q_{d-1}s^{d-1} + \cdots + q_0$. The radius of convergence of $\widehat{B}[s]$ is finite, cf. (5.9) below, but $p(s)/q(s)$ is an analytic continuation of $\widehat{B}[s]$ to the domain $\Omega_B = \{s \in \mathbb{C} : q(s) \neq 0\}$. Since $\widehat{A}[-s]$ is always defined when $\Re(s) > 0$, $\widehat{A}[-s]p(s)/q(s)$ is therefore an analytic continuation of $\widehat{F}[s] = \widehat{A}[-s]\hat{B}[s]$ to $\Omega = \{s \in \Omega_B : \Re(s) > 0\}$.

Lemma 5.9 *For any $z > 0$, the conditional distribution of U given $U > z$ has a m.g.f. of the form $p^{(z)}(s)/q(s)$ where $p^{(z)}$ is a polynomial.*

Proof. If we write $q(s) = (s - t_1)^{d_1} \cdots (s - t_k)^{d_k}$ where $d_i > 0$, $d_1 + \cdots + d_k = d$, fractional expansion of $p(s)/g(s)$ and transform inversion shows that the density $b(x)$ of B has the form

$$\sum_{i=1}^{k}\sum_{j=1}^{d_i} c_{ij}x^{j-1}\mathrm{e}^{t_i x}. \tag{5.9}$$

To get the density $b^{(z)}(x)$ of U given $U > z$, we must replace x by $x+z$ and divide by $\mathbb{P}(U > z)$ which after expanding $(x+z)^{j-1}$ shows that $b^{(z)}(x)$ has a similar form, only with changed c_{ij}. Therefore the m.g.f. has the asserted form. □

Theorem 5.10 *Assume $\mu = \mathbb{E}X = \mathbb{E}U - \mathbb{E}T < 0$ and that the function*

$$1 - \widehat{F}[s] \;=\; 1 - \widehat{A}[-s]\widehat{B}[s] \;=\; 1 - \widehat{A}[-s]\frac{p(s)}{q(s)} \tag{5.10}$$

has d roots $s_1, \ldots, s_d$ in Ω. Then $\widehat{G}_+[s] = 1 - \dfrac{1}{q(s)}\displaystyle\prod_{i=1}^{d}(s - s_i)$.

Proof. By Lemmas 5.1, 5.9 , $1 - \widehat{G}_+[s]$ must be of the form $p_+(s)/q(s)$ where p_+ is a polynomial. From $\widehat{G}_+[s] \to 0$, $s \to -\infty$ and $q(s) = s^d + \cdots$ it then follows that $p_+(s) = s^d + \cdots$, i.e. we can write $p_+(s) = \prod_{i=1}^{d}(s - \widetilde{s}_i)$. The Wiener–Hopf factorization identity $1 - \widehat{F}[s] = \bigl(1 - \widehat{G}_-[s]\bigr)\bigl(1 - \widehat{G}_+[s]\bigr)$ (valid by analytic continuation for all $s \in \Omega$) becomes

$$\psi(s)\prod_{i=1}^{d}(s - s_i) \;=\; \bigl(1 - \widehat{G}_-[s]\bigr)\frac{1}{q(s)}\prod_{i=1}^{d}(s - \widetilde{s}_i)$$

for a suitable function ψ, cf. the definition preceding Theorem 5.2. Since $\bigl|\widehat{G}_-[s]\bigr| < 1$ for $\Re(s) > 0$, the result follows easily (cf. also the parallel proof of Theorem 5.3). □

5d Phase–Type Distributions. $GI/PH/1$

In the setting of the preceding subsection, we next strengthen the assumption of B having a rational transform to B being phase–type, say with representation $(E, \boldsymbol{\alpha}, \boldsymbol{T})$. Of course, Theorem 5.10 applies to this case as well, but we shall present an alternative more probabilistic approach.

Proposition 5.11 *G_+ is* (defective) *phase–type with representation* $(\boldsymbol{\alpha}_+, \boldsymbol{T})$ *for some vector* $\boldsymbol{\alpha}_+ = (\alpha_{+;j})$.
[defective means $\boldsymbol{\alpha}_+\mathbf{1} < 1$.]

Proof. By Lemma 5.1, $G_+ = \int_0^\infty B^{(z)}\,\nu(\mathrm{d}z)$ for some measure ν. But by Problem III.4.4, $B^{(z)}$ is phase–type with representation $(\boldsymbol{\beta}^{(z)}, \boldsymbol{T})$ for some $\boldsymbol{\beta}^{(z)}$ [in fact, $\boldsymbol{\beta}^{(z)} = \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}z}/\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}z}\mathbf{1}$]. This is easily seen to imply the assertion with $\boldsymbol{\alpha}_+ = \int_0^\infty \boldsymbol{\beta}^{(z)}\,\nu(\mathrm{d}z)$. □

Thus, the problem is to evaluate $\boldsymbol{\alpha}_+$. To this end, we define a process $\{m_x\}$ as in Fig. 5.1.

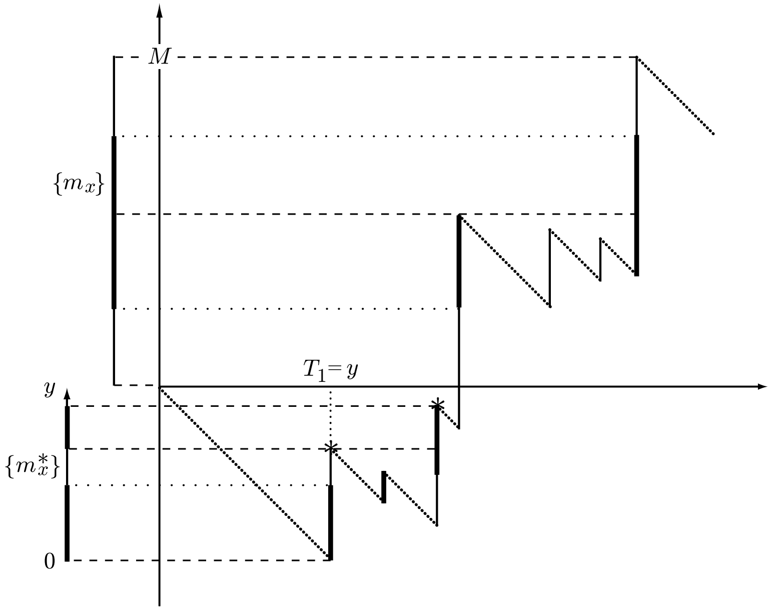

Figure 5.1

In Fig. 5.1, we have assumed two phases represented by thick and thin lines. The process depicted, say $\{R_t\}$, is the netput process for the workload, i.e. the process that jumps by the service time when a customer arrives and decreases linearly between arrivals; obviously, $\{R_t\}$ has the same ascending ladder height distribution as the given random walk $\{S_n\}$, which corresponds to the values just after jumps. The thin and thick lines in the jumps correspond to the phases in the Markov processes generating the service times, and m_x is the phase in which level x is upcrossed. We let ω_+ be the time of the first upcrossing of level 0 so that $S_{\tau_+} = R_{\omega_+}$.

Theorem 5.12 (a) $\boldsymbol{\alpha}_+$ *is the (defective) distribution of* m_0*;*
(b) $\{m_x\}$ *is a (terminating) Markov process on* E*, with intensity matrix* $\boldsymbol{Q}$ *given by* $\boldsymbol{Q} = \boldsymbol{T} + \boldsymbol{t}\boldsymbol{\alpha}_+$*;*
(c) $\boldsymbol{\alpha}_+$ *satisfies* $\boldsymbol{\alpha}_+ = \varphi(\boldsymbol{\alpha}_+)$*, where*

$$\varphi(\boldsymbol{\alpha}_+) \;=\; \boldsymbol{\alpha}\widehat{A}[\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha}_+] \;=\; \boldsymbol{\alpha}\int_0^\infty \mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha}_+)y}A(\mathrm{d}y). \tag{5.11}$$

(d) $\boldsymbol{\alpha}_+$ *can be computed by iteration of* (5.11)*, i.e. by* $\boldsymbol{\alpha}_+ = \lim_{n\to\infty}\boldsymbol{\alpha}_+^{(n)}$ *where*

$$\boldsymbol{\alpha}_+^{(0)} = \boldsymbol{0},\ \boldsymbol{\alpha}_+^{(1)} = \varphi\bigl(\boldsymbol{\alpha}_+^{(0)}\bigr),\ \ \boldsymbol{\alpha}_+^{(2)} = \varphi\bigl(\boldsymbol{\alpha}_+^{(1)}\bigr),\ \ \ldots. \tag{5.12}$$

Proof. Part (a) is clear. The proof of (b) is similar to that used for phase–type renewal processes in III.5: in between ladder heights, $\{m_x\}$ is governed by $\boldsymbol{T}$. The vector of intensities generating ladder heights in the different states is $\boldsymbol{t}$, and $\{m_x\}$ either terminates at a ladder height or is restarted according to $\boldsymbol{\alpha}_+$. Collecting terms, (b) follows.

For (c), we condition upon $T_1 = y$ and define $\{m_x^*\}$ from $\{R_{t+y} - R_{y-}\}$ in the same way as $\{m_x\}$ is defined from $\{R_t\}$; cf. Fig. 5.1. Then $\{m_x^*\}$ is Markov with the same transition intensities as $\{m_x\}$, but with initial distribution $\boldsymbol{\alpha}$ rather than $\boldsymbol{\alpha}_+$. Also, obviously $m_0 = m_y^*$. Since the conditional distribution of m_y^* given $T_1 = y$ is $\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{Q}y}$, it follows by integrating y out that the distribution $\boldsymbol{\alpha}_+$ of m_0 is given by the final expression in (5.11).

In (d), note first that the term $\boldsymbol{t}\boldsymbol{\beta}$ in $\varphi(\boldsymbol{\beta})$ represents feedback with rate vector $\boldsymbol{t}$ and feedback probability vector $\boldsymbol{\beta}$. Hence $\varphi(\boldsymbol{\beta})$ (defined on the domain of subprobability vectors $\boldsymbol{\beta}$) is an increasing function of $\boldsymbol{\beta}$. In particular, $\boldsymbol{\alpha}_+^{(1)} \geq \boldsymbol{0} = \boldsymbol{\alpha}_+^{(0)}$ implies

$$\boldsymbol{\alpha}_+^{(2)} \;=\; \varphi\bigl(\boldsymbol{\alpha}_+^{(1)}\bigr) \;\geq\; \varphi\bigl(\boldsymbol{\alpha}_+^{(0)}\bigr) \;=\; \boldsymbol{\alpha}_+^{(1)}$$

and (by induction) that $\bigl\{\boldsymbol{\alpha}_+^{(n)}\bigr\}$ is an increasing sequence such that $\lim_{n\to\infty}\boldsymbol{\alpha}_+^{(n)}$ exists. Similarly, $\boldsymbol{0} = \boldsymbol{\alpha}_+^{(0)} \leq \boldsymbol{\alpha}_+$ yields

$$\boldsymbol{\alpha}_+^{(1)} \;=\; \varphi\bigl(\boldsymbol{\alpha}_+^{(0)}\bigr) \;\leq\; \varphi\left(\boldsymbol{\alpha}_+\right) \;=\; \boldsymbol{\alpha}_+$$

and by induction that $\boldsymbol{\alpha}_+^{(n)} \leq \boldsymbol{\alpha}_+$ for all n. Thus, $\lim_{n\to\infty}\boldsymbol{\alpha}_+^{(n)} \leq \boldsymbol{\alpha}_+$.

To prove the converse inequality, we let $F_n = \{T_1 + \cdots + T_{n+1} > \omega_+\}$ be the event that $\{R_t\}$ has at most n arrivals in $[T_1, \omega_+]$, and let $\widetilde{\alpha}_{+;i}^{(n)} = \mathbb{P}(m_{T_1}^* = i;\, F_n)$. Obviously, $\widetilde{\boldsymbol{\alpha}}_+^{(n)} \uparrow \boldsymbol{\alpha}_+$, so to complete the proof it suffices to show that $\widetilde{\boldsymbol{\alpha}}_+^{(n)} \leq \boldsymbol{\alpha}_+^{(n)}$ for all n. For $n = 0$, both quantities are just $\boldsymbol{0}$. Assume the assertion shown for $n-1$, and define a subexcursion of $\{R_t\}$ as the segment from just after an arrival time, say σ_k (here $\sigma_0 = 0$), until level R_{σ_k} is upcrossed again (thus on Fig. 5.1 there are two subexcursions in $[0, \omega_+]$). Then each subexcursion before time ω_+ can contain at most $n-1$

arrivals (n arrivals are excluded because of the initial arrival at time T_1). It follows that on F_n the feedback to $\{m_x^*\}$ after each ladder step cannot exceed $\widetilde{\boldsymbol{\alpha}}_+^{(n-1)}$ so that

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_+^{(n)} &\le \widetilde{\boldsymbol{\alpha}} \int_0^\infty \mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\widetilde{\boldsymbol{\alpha}}_+^{(n-1)})y} A(\mathrm{d}y) \\ &\le \boldsymbol{\alpha} \int_0^\infty \mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha}_+^{(n-1)})y} A(\mathrm{d}y) \;=\; \varphi\big(\boldsymbol{\alpha}_+^{(n-1)}\big) \;=\; \boldsymbol{\alpha}_+^{(n)}.\end{aligned}$$

□

Corollary 5.13 *The maximum M of the random walk or, equivalently, the $GI/PH/1$ waiting time, is the lifetime of $\{m_x\}$. In particular, $\mathbb{P}(M = 0) = 1 - \boldsymbol{\alpha}_+\boldsymbol{1}$ and M is absolutely continuous on $(0,\infty)$ with density $\boldsymbol{\alpha}_+\mathrm{e}^{(\boldsymbol{T}+\boldsymbol{t}\boldsymbol{\alpha}_+)x}$.*

Corollary 5.14 *Consider the $M/PH/1$ queue with arrival intensity β and phase representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ of the service time distribution B. Then the steady–state waiting time W is phase–type with representation $(\boldsymbol{\alpha}_+, \boldsymbol{Q})$ where $\boldsymbol{\alpha}_+ = -\beta\boldsymbol{\alpha}\boldsymbol{T}^{-1}$, $\boldsymbol{Q} = \boldsymbol{T} + \boldsymbol{t}\boldsymbol{\alpha}_+$. In particular, $\mathbb{P}(W = 0) = 1 - \boldsymbol{\alpha}_+\boldsymbol{1}$ and W is absolutely continuous on $(0,\infty)$ with density $\boldsymbol{\alpha}_+\mathrm{e}^{\boldsymbol{Q}x}$.*

Proof. By III.5.3, the stationary excess distribution B_0 is phase–type with representation $(\boldsymbol{\nu}, \boldsymbol{T})$ where $\boldsymbol{\nu} = -\boldsymbol{\alpha}\boldsymbol{T}^{-1}/\mu_B$. Since $G_+ = \rho B_0$ by Theorem 5.7(a), it follows that $\boldsymbol{\alpha}_+ = \rho\boldsymbol{\nu} = -\beta\boldsymbol{\alpha}\boldsymbol{T}^{-1}$. □

We finally give the link to Theorem 5.10. Let d denote the number of phases.

Corollary 5.15 *Suppose $\mu < 0$, that the equation $\widehat{F}[s] = 1$ has d distinct roots $\rho_1, \ldots, \rho_d$ in the domain $\Re(s) > 0$, and define $\boldsymbol{h}_i = (-\rho_i\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}$, $\boldsymbol{Q} = \boldsymbol{D}\boldsymbol{C}^{-1}$ where $\boldsymbol{C}$ is the matrix with columns $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d$, $\boldsymbol{D}$ that with columns $-\rho_1\boldsymbol{h}_1, \ldots, -\rho_d\boldsymbol{h}_d$. Then G_+ is phase-type with representation $(\boldsymbol{\alpha}_+, \boldsymbol{T})$ with $\boldsymbol{\alpha}_+ = \boldsymbol{\alpha}(\boldsymbol{Q}-\boldsymbol{T})/\boldsymbol{\alpha}\boldsymbol{t}$. Further, letting $\boldsymbol{\nu}_i$ be the left eigenvector of $\boldsymbol{Q}$ corresponding to $-\rho_i$ and normalized by $\boldsymbol{\nu}_i\boldsymbol{h}_i = 1$, $\boldsymbol{Q}$ has diagonal form*

$$\boldsymbol{Q} \;=\; -\sum_{i=1}^d \rho_i\,\boldsymbol{\nu}_i \otimes \boldsymbol{h}_i \;=\; -\sum_{i=1}^d \rho_i\,\boldsymbol{h}_i\boldsymbol{\nu}_i\,. \tag{5.13}$$

Proof. Since $\Re(\rho_i) > 0$ and G_- is concentrated on $(-\infty, 0)$, we have $\big|\widehat{G}_-[\rho_i]\big| < 1$, and hence $\widehat{G}_+[\rho_i] = 1$ by the Wiener-Hopf factorization identity (3.1), which according to Proposition 5.11 means that $\boldsymbol{\alpha}_+(-\rho_i\boldsymbol{I}-\boldsymbol{T})^{-1}\boldsymbol{t}$ $= 1$. Hence

$$\begin{aligned}\boldsymbol{Q}\boldsymbol{h}_i &= (\boldsymbol{T} + \boldsymbol{t}\boldsymbol{\alpha}_+)\boldsymbol{h}_i \;=\; \boldsymbol{T}(-\rho_i\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t} + \boldsymbol{t} \\ &= (\boldsymbol{T} + \rho_i\boldsymbol{I} - \rho_i\boldsymbol{I})(-\rho_i\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t} + \boldsymbol{t} \;=\; -\boldsymbol{t} - \rho_i\boldsymbol{h} + \boldsymbol{t} \;=\; -\rho_i\boldsymbol{h}\,.\end{aligned}$$

It follows that the matrix $\boldsymbol{Q}$ in Theorem 5.12 has the d distinct eigenvalues $-\rho_1, \ldots, -\rho_d$ with corresponding right eigenvectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d$. This immediately implies that $\boldsymbol{Q}$ has the form $\boldsymbol{D}\boldsymbol{C}^{-1}$ and the last assertion on the diagonal form. Given $\boldsymbol{T}$ has been computed, we get

$$\frac{1}{\boldsymbol{\alpha t}}\boldsymbol{\alpha}(\boldsymbol{Q}-\boldsymbol{T}) \;=\; \frac{1}{\boldsymbol{\alpha t}}\boldsymbol{\alpha t}\boldsymbol{\alpha}_{+} \;=\; \boldsymbol{\alpha}_{+}\,. \qquad \square$$

Notes Results like those of the present section have a long history. In particular, some early references are Wald (1947) for Section 5a and Täcklind (1942) and Smith (1953) for Section 5c. An ingredient that is often met is *Rouché's theorem*, a classical result from complex analysis giving a criterion for two complex functions to have the same number of zeros within the unit circle.

This complex plane approach is often met with criticism for a number of reasons such as to lack probabilistic interpretation and to only give transform solutions. In queueing theory, an alternative approach (the matrix–geometric method) has been developed largely by M.F. Neuts and his students, starting around 1975. For surveys, see Neuts (1981, 1989) and Latouche and Ramaswami (1999) (we cover the basics in XI.3–4). Here phase-type assumptions are essential, but the models solved are basically Markov chains and processes with countably many states (e.g. queue length processes). Asmussen (1992a) modified some of the ideas to deal with waiting times and workloads, and Section 5d is from that paper (the $GI/PH/1$ waiting time was studied earlier by Sengupta, 1989, by different means; see further XI.3d).

Concerning further explicit distributions of ladder heights, random walk maxima, etc. it is remarkable that even the case of a Gaussian F presents major difficulties so that the available results (Lotov, 1996, Chang and Perez, 1997) are recent and not of a very simple form.

IX
Lévy Processes, Reflection and Duality

1 Lévy Processes

By a *Lévy process* we understand a real–valued continuous–time stochastic process $\{X_t\}_{t\geq 0}$ with stationary independent increments and $X_0 = 0$, cf. III.7. Simple examples are: Brownian motion with drift θ and variance constant σ^2, for which

$$\mathbb{E}X_t = t\theta, \quad \mathbb{V}ar X_t = t\sigma^2, \quad \mathbb{E}\mathrm{e}^{\alpha X_t} = \mathrm{e}^{t\kappa(\alpha)} \tag{1.1}$$

where $\kappa(\alpha) = \theta\alpha + \sigma^2\alpha^2/2$; and a compound Poisson process $\sum_1^{N_t} U_i$. Here $\{N_t\}$ is a Poisson process with rate β and the U_i are independent, mutually and of $\{N_t\}$, with common distribution B, and

$$\mathbb{E}X_t = t\beta\mathbb{E}U, \quad \mathbb{V}ar X_t = t\beta\mathbb{E}U^2, \quad \mathbb{E}\mathrm{e}^{\alpha X_t} = \mathrm{e}^{t\kappa(\alpha)} \tag{1.2}$$

where $\kappa(\alpha) = \beta(\widehat{B}[\alpha] - 1)$, $\widehat{B}[\alpha] = \mathbb{E}\mathrm{e}^{\alpha U}$. Of course, a linear combination of independent Lévy processes is again a Lévy process.

Recall that a probability measure μ is called *infinitely divisible* (i.d.) if for each $n = 1, 2, \ldots$ there exists μ_n such that $\mu = \mu_n^{*n}$ (the nth–fold convolution of μ_n). If $\{X_t\}$ is a Lévy process, then for each t the distribution of X_t is i.d., since the terms in the expansion $X_t = \sum_1^n (X_{tk/n} - X_{t(k-1)/n})$ are i.i.d. This connection plays a crucial role in the theory of Lévy processes.

Lemma 1.1 *If μ is i.d., then the ch.f. $\widehat{\mu}[\mathrm{i}t] = \int \mathrm{e}^{\mathrm{i}tx}\mu(\mathrm{d}x)$ is nonzero for each real t, and there exists a unique continuous complex function $\kappa(\cdot)$ such that $\widehat{\mu}[\mathrm{i}t] = \mathrm{e}^{\kappa(\mathrm{i}t)}$.*

Proof. If μ is i.d., then the symmetrized distribution $\mu^{\#}$ defined by $\mu^{\#}(A) = \int \mu(A+x)\,\mu(\mathrm{d}x)$ is i.d. corresponding to $\mu_n^{\#} = (\mu_n)^{\#}$. Since $\widehat{\mu}^{\#} = |\widehat{\mu}|^2$, it follows that

$$I(\widehat{\mu}[\mathrm{i}t] \neq 0) \;=\; \lim_{n\to\infty} \bigl|\widehat{\mu}[\mathrm{i}t]\bigr|^{2/n} \;=\; \lim_{n\to\infty} \widehat{\mu}_n^{\#}[\mathrm{i}t].$$

Thus $I(\widehat{\mu}[\mathrm{i}t] \neq 0)$ is limit of ch.f.'s, and since $I(\widehat{\mu}[\mathrm{i}t] \neq 0)$ is continuous at $t=0$ because a ch.f. is continuous and 1 at 0, $I(\widehat{\mu}[\mathrm{i}t] \neq 0)$ is a ch.f., cf. e.g. Chung (1974) p. 161. The continuity of a ch.f. and $\widehat{\mu}[0] = 1$ therefore implies that $I(\widehat{\mu}(\mathrm{i}t) \neq 0)$ must be identically 1. The last statement concerning $\kappa(\cdot)$ follows from general facts on the complex logarithm. □

In view of this result, we can define the *Lévy exponent* $\kappa(\cdot)$ of a Lévy process by taking μ as the distribution of X_1 (that is, $\kappa(\alpha) = \log \mathbb{E}\mathrm{e}^{\alpha X_1}$), and extend $\kappa(\cdot)$ to $\Theta = \bigl\{\alpha \in C : \mathbb{E}\mathrm{e}^{\Re(\alpha)X_1} < \infty\bigr\}$ in such a way that $\mathbb{E}\mathrm{e}^{\alpha X_1} = \mathrm{e}^{\kappa(\alpha)}$ not only for $\Re(\alpha) = 0$ but for all $\alpha \in \Theta$ (cf. Problem 1.2).

Proposition 1.2 *For a Lévy process, the following statements are equivalent:* (i) *the distribution of X_t is a measurable function of t;* (ii) $\{X_t\}$ *is stochastically continuous;* (iii) $\{X_t\}$ *has a version with D–paths;* (iv) $\mathbb{E}\mathrm{e}^{\alpha X_t} = \mathrm{e}^{t\kappa(\alpha)}$ *for all $t \ge 0$ and all $\alpha \in \Theta$.*

Proof. Recall that $\{X_t\}$ being stochastically continuous means that for all t, $X_s \xrightarrow{\mathbb{P}} X_t$ as $s \to t$; for a Lévy process, this is easily seen to be equivalent to $X_s \xrightarrow{\mathbb{P}} 0$ as $s \to 0$. In conjunction with the spatial homogeneity, the implication (ii)⇒(iii) then follows by general Markov process theory (see, e.g., Sato, 1999, p. 59) and also (i) holds generally in D so that (iii)⇒(i).

For (i)⇒(iv), let $f(t) = \mathbb{E}\mathrm{e}^{\alpha X_t}$. Then under (i), $f(\cdot)$ is measurable, and the functional equation $f(s+t) = f(s)f(t)$ (which holds because of the stationary independent increments) then implies $f(t) = f(1)^t = \mathrm{e}^{t\kappa(\alpha)}$. Finally, (iv)⇒(ii) is trivial. □

We will usually assume that a Lévy process satisfies one of the minor equivalent conditions in Proposition 1.2.

Recall the *Lévy–Khintchine representation* of an i.d. distribution:

Theorem 1.3 *Let μ be i.d. with $\kappa(\cdot)$ as in Lemma 1.1. Then there exists $\theta \in \mathbb{R}$, $\sigma \ge 0$ and a nonnegative measure ν on $\mathbb{R}$ with*

$$\int_{-\epsilon}^{\epsilon} x^2\,\nu(\mathrm{d}x) < \infty, \quad \nu([-\epsilon,\epsilon]^c) = \int_{\{x:\,|x|>\epsilon\}} \nu(\mathrm{d}x) < \infty \tag{1.3}$$

for all $\epsilon > 0$, such that

$$\kappa(\alpha) \;=\; \theta\alpha + \sigma^2\alpha^2/2 + \int_{-\infty}^{\infty} \bigl[\mathrm{e}^{\alpha x} - 1 - \alpha x I(|x| \le 1)\bigr]\,\nu(\mathrm{d}x), \quad \alpha \in \Theta. \tag{1.4}$$

[$I(|x| \le 1)$ could be replaced by $I(|x| \le \epsilon)$ for any $\epsilon > 0$ by changing θ appropriately]. The proof is given below. The measure ν is referred to as the *Lévy measure* and (ν, θ, σ^2) to as the *characteristic triplet*. The Lévy measure is unique and additive under convolutions, as is θ and σ^2. Note that if ν satisfies the stronger requirement

$$\int_{-\epsilon}^{\epsilon} |x|\,\nu(\mathrm{d}x) < \infty, \quad \nu([-\epsilon,\epsilon]^c) < \infty, \tag{1.5}$$

then (1.4) can be rewritten in the form

$$\kappa(\alpha) \;=\; \theta\alpha \;+\; \sigma^2\alpha^2/2 \;+\; \int_{-\infty}^{\infty} [\mathrm{e}^{\alpha x} - 1]\,\nu(\mathrm{d}x), \quad \alpha \in \Theta \tag{1.6}$$

(replace θ by $\theta - \int_{-1}^{1} x\,\nu(\mathrm{d}x)$). Here the first two terms correspond to a Brownian motion with drift θ and variance constant σ^2 and, when $|\nu| < \infty$, the last to an independent compound Poisson process with $\beta = |\nu|$, $B = \nu/|\nu|$, cf. (1.1), (1.2). We proceed to the construction and interpretation of the process in the general case.

Proposition 1.4 *Let ν be a nonnegative measure satisfying* (1.5) *and let $M(\mathrm{d}t, \mathrm{d}x)$ be a Poisson process on $[0,\infty) \times \mathbb{R}$ with intensity measure $\mathrm{d}t \otimes \nu(\mathrm{d}x)$. Then*

$$X_t \;=\; \int_0^t \int_{\mathbb{R}} x\,M(\mathrm{d}s, \mathrm{d}x) \tag{1.7}$$

is a Lévy process with D–paths and Lévy exponent $\int_{-\infty}^{\infty} [\mathrm{e}^{\alpha x} - 1]\,\nu(\mathrm{d}x)$.

Proof. By the additivity property noted above, we may write $\{X_t\}$ as an independent sum of terms corresponding to components of ν concentrated on different subsets, say $\{|x| \geq 1\}$, $\{-1 < x < 0\}$ and $\{0 < x < 1\}$. The first term is compound Poisson and the assertion is trivial, so by symmetry, we can assume w.l.o.g. that ν is concentrated on $(0,1)$. Then the expectation of the r.h.s. of (1.7) is $t\int_0^1 x\,\nu(\mathrm{d}x)$, which is finite by (1.5). Hence X_t is welldefined and finite, and the property of D–paths follows by monotone convergence in (1.7). Finally, if we let $X_t^{(\epsilon)} = \int_0^t \int_\epsilon^1 x\,M(\mathrm{d}s, \mathrm{d}x)$, then $\{X_t^{(\epsilon)}\}$ is compound Poisson and hence by (1.2) and monotone convergence,

$$\begin{aligned}
\log \mathbb{E}\mathrm{e}^{\alpha X_1} \;&=\; \lim_{\epsilon \downarrow 0} \log \mathbb{E}\mathrm{e}^{\alpha X_1^{(\epsilon)}} \;=\; \lim_{\epsilon \downarrow 0} \int_\epsilon^1 [\mathrm{e}^{\alpha x} - 1]\,\nu(\mathrm{d}x) \\
&=\; \int_0^1 [\mathrm{e}^{\alpha x} - 1]\,\nu(\mathrm{d}x).
\end{aligned}$$

□

The interpretation of (1.7) is that $\{X_t\}$ moves by jumps alone, that a jump of size x occurs with intensity $\nu(\mathrm{d}x)$, and that jumps of different sizes are independent. In the general case (1.3), the situation is more complicated since the r.h.s. of (1.7) need not converge, and one needs to subtract means and go to the limit (this procedure is known as *compensation*).

Proposition 1.5 *Let ν be a nonnegative measure satisfying* (1.3)*, let $\nu^{(\epsilon)}(A) = \nu\bigl(A \cap [-\epsilon,\epsilon]^c\bigr)$, $\mu^{(\epsilon)} = \int_{\{x:\,\epsilon<|x|\le 1\}} x\,\nu(\mathrm{d}x)$ and let $\bigl\{Y_t^{(\epsilon)}\bigr\}$ be a compound Poisson process with Lévy measure $\nu^{(\epsilon)}$, $X_t^{(\epsilon)} = Y_t^{(\epsilon)} - t\mu^{(\epsilon)}$. Then $X_t = \lim_{\epsilon\downarrow 0} X_t^{(\epsilon)}$ exists a.s. and is a Lévy process with Lévy exponent $\int_{-\infty}^{\infty}\bigl[\mathrm{e}^{\alpha x} - 1 - \alpha x I(|x|\le 1)\bigr]\,\nu(\mathrm{d}x)$.*

Proof. Assume w.l.o.g. that ν is concentrated on $[-1,1]$. Then by (1.2), $\mathbb{V}ar X_t^{(\epsilon)} = t\int_{\{x:\,\epsilon<|x|\le 1\}} x^2\,\nu(\mathrm{d}x)$. Also, clearly $\mathbb{E}X_t^{(\epsilon)} = 0$ and hence the family $\bigl\{X_t^{(\epsilon)}\bigr\}_{\epsilon>0}$ is L_2–bounded. The existence of a limit X_t as $\epsilon\downarrow 0$ therefore follows by verifying that it is a backward martingale, i.e. that

$$\mathbb{E}\bigl[X_t^{(\epsilon_1)}\,\big|\,X_t^{(\epsilon_2)}\bigr] \;=\; X_t^{(\epsilon_2)} \tag{1.8}$$

when $0 < \epsilon_1 < \epsilon_2$. But by general results on Poisson thinning, $Y_t^{(\epsilon_2)}$ is independent of $Y_t^{(\epsilon_1)} - Y_t^{(\epsilon_2)}$ which is readily seen to imply (1.8). The rest of the proof is easy by limit arguments. □

We now turn to the proof of Theorem 1.3.

Lemma 1.6 *For any i.d. distribution μ, there exists a stochastically continuous Lévy process $\{X_t\}$ such that X_1 has distribution μ.*

Proof. For each t, define $\widehat{\lambda}^{(t)}[\mathrm{i}s] = \mathrm{e}^{t\kappa(\mathrm{i}s)}$. The nth root μ_n of μ is unique and has ch.f. $\widehat{\mu}[\mathrm{i}s]^{1/n} = \widehat{\lambda}^{(1/n)}[\mathrm{i}s]$ by Lemma 1.1 and its proof. Taking k–fold convolutions and writing t as limit of rationals of the form k/n shows that $\widehat{\lambda}^{(t)}[\cdot]$ is a limit of ch.f.'s, hence (since $\widehat{\lambda}^{(t)}[\cdot]$ is continuous) the ch.f. of a probability distribution $\lambda^{(t)}$. Since the ch.f.'s are multiplicative in t, we have $\lambda^{(s+t)} = \lambda^{(s)} * \lambda^{(t)}$. Hence if for $0 = t_0 < t_1 < \cdots < t_m$ we define the distribution $\mathbb{P}_{t_0,\ldots,t_m}$ of $(X_{t_0},\ldots,X_{t_m})$ by the requirement that $X_{t_0} = 0$ a.s. and that the $X_{t_k} - X_{t_{k-1}}$ are independent with distributions $\lambda^{(t_k - t_{k-1})}$, we get a consistent family, the Kolmogorov extension of which is the distribution of a Lévy process. Stochastic continuity follows since $\widehat{\lambda}^{(t)}[\mathrm{i}s] \to 1$ as $t\downarrow 0$ and hence $X_t \xrightarrow{\mathscr{D}} 0$, and that X_1 has distribution μ is clear from $\widehat{\lambda}^{(1)}[\mathrm{i}s] = \widehat{\mu}[\mathrm{i}s]$. □

Lemma 1.7 *Let $\{X_t\}$ be a Lévy process such that the absolute value of a jump is bounded. Then $\Theta = \mathbb{R}$.*

Proof. Assume w.l.o.g. that $|X_t - X_{t-}| \le 1$ for all t. Define $\tau_0 = 0$, $\tau_{k+1} = \inf\{t > 0 : |X_{t+\tau_k} - X_{\tau_k}| \ge 1\}$, $k = 0,1,\ldots$. Then by the strong Markov property (Problem 1.1), the τ_k are i.i.d., with common distribution say F, and since $|X_{t+\tau_k} - X_{\tau_k}| \le 2$, we get

$$\mathbb{P}\bigl(|X_t| \ge 2n\bigr) \;\le\; \mathbb{P}(\tau_1 + \cdots + \tau_n \le t) \;=\; F^{*n}(t)\,.$$

For fixed t, $F^{*n}(t)$ decreases faster than any exponential, cf. V.2.5. From this, it follows that $\mathbb{E}\mathrm{e}^{\alpha X(t)} < \infty$ for all α. □

Lemma 1.8 *A Lévy process $\{X_t\}$ with continuous sample paths is Brownian motion (θ,σ^2) for some θ,σ^2.*

Proof. By Lemma 1.7, all cumulants of X_t exist (and are necessarily linear functions of t) so we may assume w.l.o.g. that $\mathbb{E}X_t = 0$, $\mathbb{V}ar X_t = t$ and then have to show that X_1 is standard normal. Writing X_1 as a sum of n i.i.d. terms distributed as X_h $(h = 1/n)$, this follows from the Feller–Lindeberg theorem provided we can show $\mathbb{E}[X_h^2;\, |X_h| > \epsilon] = \mathrm{o}(h)$ for each $\epsilon > 0$. Let $\tau(\delta) = \inf\{t > 0 : |X_t| \ge \delta\}$. Then by stochastic continuity, there is a constant c_1 such that $\mathbb{P}(|X_h| \ge \delta) \ge c_1\mathbb{P}(\tau(2\delta) \le h)$ for all small h. By Chebycheff, the l.h.s. can be bounded by c_2h, so that $\mathbb{P}(\tau(\epsilon) \le h)$, being bounded by $\mathbb{P}(\tau(\epsilon/2) \le h)^2$ because of continuity, is $\mathrm{O}(h^2)$. We then get

$$\begin{aligned}
&\mathbb{E}\bigl[X_h^2;\, |X_h| > \epsilon\bigr] \\
&\quad\le\; \mathbb{E}\bigl[X_h^2;\, \tau(\epsilon) \le h\bigr] \;=\; \mathbb{E}\bigl[\epsilon^2 + \mathbb{V}ar(X_h|\tau(\epsilon) \le h);\, \tau(\epsilon) \le h\bigr] \\
&\quad\le\; \bigl[\epsilon^2 + \mathrm{O}(h)\bigr]\mathbb{P}(\tau(\epsilon) \le h) \;=\; \mathrm{O}(h^2).
\end{aligned}$$

□

Given the i.d. distribution μ, we can find a Lévy process $\{X_t\}$ with D–paths such that X_1 has distribution μ, cf. Lemma 1.6 and Proposition 1.2. Let $N_t(A)$ be the number of jumps of size in A before t, $N_t^{(\epsilon)} = N_t\bigl([-\epsilon,\epsilon]^c\bigr)$, $Z_t^{(\epsilon)} = \int_0^t X_s\, dN_s^{(\epsilon)}$ (the sum of jumps of absolute size $> \epsilon$) and $Y_t^{(\epsilon)} = X_t - Z_t^{(\epsilon)}$. If $0 \notin \overline{A}$, then $N_t(A) < \infty$ for all t, and since $\{N_t(A)\}$ is a point process without multiple points and having stationary independent increments, $\{N_t(A)\}$ must be Poisson with rate $\nu(A)$ for some $\nu(A)$; that the set function $\nu(\cdot)$ is a measure is clear from $\nu(A) = \mathbb{E}N_1(A)$, and it is also obvious that $\nu([-\epsilon,\epsilon]^c) < \infty$. Similarly, $\bigl\{Z_t^{(\epsilon)}\bigr\}$ must be compound Poisson, and we denote its jump size distribution by $F^{(\epsilon)}$.

Lemma 1.9 $\bigl\{Z_t^{(\epsilon)}\bigr\}$ *and* $\bigl\{Y_t^{(\epsilon)}\bigr\}$ *are independent for each* $\epsilon > 0$.

Proof. Clearly, $\bigl\{Y_t^{(\epsilon)}\bigr\}_{0\le t\le T}$ has stationary independent increments conditionally upon $N_T^{(\epsilon)} = 0$, and accordingly, there exists $\kappa^{(\epsilon)}(\cdot)$ such that

$$\mathbb{E}\left[\mathrm{e}^{\alpha Y_t^{(\epsilon)}}\,\middle|\, N_T^{(\epsilon)} = 0\right] \;=\; \mathrm{e}^{\kappa^{(\epsilon)}(\alpha)},\quad 0 \le t \le T.$$

Now for $T > 0$ and a given n, define $J_k = \bigl((k-1)T/n, kT/n\bigr]$, $Y_{J_k}^{(\epsilon)} = Y_{kT/n}^{(\epsilon)} - Y_{(k-1)T/n}^{(\epsilon)}$, etc. Then

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{e}^{\mathrm{i}sY_T^{(\epsilon)}};\, N_T^{(\epsilon)} = 1, Z_T^{(\epsilon)} \le x\right] \qquad (1.9)\\
&\quad=\; \sum_{k=1}^{n}\mathbb{E}\left[\mathrm{e}^{\mathrm{i}sY_T^{(\epsilon)}};\, N_{J_k}^{(\epsilon)} = 1, N_{J_\ell}^{(\epsilon)} = 0, \ell \ne k, Z_{J_k}^{(\epsilon)} \le x\right]
\end{aligned}$$

$$= \quad \mathrm{e}^{(n-1)T/n\cdot\kappa^{(\epsilon)}(\mathrm{i}s)}\sum_{k=1}^{n}\mathbb{E}\Big[\mathrm{e}^{\mathrm{i}sY^{(\epsilon)}_{J_k}};\,N^{(\epsilon)}_{J_k}=1,N^{(\epsilon)}_{J_\ell}=0,\ell\neq k,Z^{(\epsilon)}_{J_k}\le x\Big].$$

Define $M_n=\sup\bigl\{|\mathrm{e}^{\mathrm{i}s(Y^{(\epsilon)}_u-Y^{(\epsilon)}_v)}-1|\bigr\}$ where the sup extends over all $u,v\le T$ such that $|u-t|\le 1/n$, $|v-t|\le 1/n$ for some epoch t of $\bigl\{N^{(\epsilon)}_t\bigr\}$. Then since each such t must be a continuity point of $\bigl\{Y^{(\epsilon)}_t\bigr\}$, we have $M_n\overset{\text{a.s.}}{\to}0$, $n\to\infty$, and hence by dominated convergence,

$$\begin{aligned}\Delta_n &= \left|\sum_{k=1}^{n}\mathbb{E}\Big[\mathrm{e}^{\mathrm{i}sY^{(\epsilon)}_{J_k}}-1;\,N^{(\epsilon)}_{J_k}=1,N^{(\epsilon)}_{J_\ell}=0,\ell\neq k,Z^{(\epsilon)}_{J_k}\le x\Big]\right|\\ &\le \mathbb{E}\big[M_n;\,N^{(\epsilon)}_T=1\big]\ \to\ 0.\end{aligned}$$

It follows that (1.9) is

$$\begin{aligned}&\mathrm{e}^{T\kappa^{(\epsilon)}(\mathrm{i}s)}\sum_{k=1}^{n}\mathbb{P}\left(N^{(\epsilon)}_{J_k}=1,N^{(\epsilon)}_{J_\ell}=0,\ell\neq k,Z^{(\epsilon)}_{J_k}\le x\right)\ +\ \mathrm{o}(1)\\ &\qquad\qquad = \ \mathrm{e}^{T\kappa^{(\epsilon)}(is)}\mathbb{P}\left(N^{(\epsilon)}_T=1,Z^{(\epsilon)}_T\le x\right).\end{aligned}$$

Similar arguments apply first to the case $N^{(\epsilon)}_T=m>1$, showing that $Z^{(\epsilon)}_T$ and $Y^{(\epsilon)}_T$ are independent, and next to the independence of $Z^{(\epsilon)}_{I_1},Y^{(\epsilon)}_{I_1},\ldots,Z^{(\epsilon)}_{I_r},Y^{(\epsilon)}_{I_r}$ for disjoint intervals $I_1,\ldots,I_r$. □

Proof of Theorem 1.3. Let $\nu^{(\epsilon)}$ be the restriction of ν to $[-\epsilon,\epsilon]^c$. Applying Lemma 1.9 repeatedly yields easily that $\bigl\{Z^{(\epsilon)}_t\bigr\}$ is a Lévy process with Lévy measure $\nu^{(\epsilon)}$ and independent of $\bigl\{Y^{(\epsilon)}_t\bigr\}$ for each $\epsilon>0$, in particular for $\epsilon=1$. Let $\widetilde{Y}^{(\epsilon)}_t=Y^{(\epsilon)}_t-t\theta^{(\epsilon)}$ where $\theta^{(\epsilon)}=\mathbb{E}Y^{(\epsilon)}_1$ (note that both the mean and the variance exist by Lemma 1.7) and $J^{(\epsilon)}_t=\widetilde{Y}^{(1)}_t-\widetilde{Y}^{(\epsilon)}_t$, $0<\epsilon<1$. Then as above, $\bigl\{J^{(\epsilon)}_t\bigr\}$ is a Lévy process, which is independent of $\bigl\{Y^{(\epsilon)}_t\bigr\}$ and where the Lévy measure is the restriction of ν to $\{x:\epsilon<|x|\le 1\}$. Hence

$$\mathbb{V}ar Y^{(1)}_t\ =\ \mathbb{V}ar J^{(\epsilon)}_t+\mathbb{V}ar Y^{(\epsilon)}_t\ \ge\ \mathbb{V}ar J^{(\epsilon)}_t\ =\ t\int_{\epsilon<|x|\le 1}x^2\nu(\mathrm{d}x).$$

Since the l.h.s. is finite, we can make two conclusions: that $\int_{-1}^{1}x^2\nu(\mathrm{d}x)<\infty$, so that ν is a legitimate Lévy measure; and that $\bigl\{J^{(\epsilon)}_t\bigr\}_{\epsilon>0}$ is a backward martingale having an a.s. limit $J^{(0)}_t$ (combine with the arguments around (1.8)). Therefore also $W_t=\lim\widetilde{Y}^{(\epsilon)}_t$ exists. But obviously, $\{W_t\}_{t\ge0}$ is continuous and since the property of stationary independent increments is easily seen to carry over in an a.s. limit, we conclude from Lemma 1.8 that $\{W_t\}$ is Brownian motion with suitable drift θ and suitable variance σ^2. Finally in the decomposition $X_t=W_t+\theta^{(1)}t+J^{(0)}_t+Z^{(1)}_t$, the components are independent Lévy processes with Lévy exponents

$$\theta\alpha+\alpha^2\sigma^2/2,\quad \theta^{(1)}\alpha,\quad \int_{-1}^{1}(\mathrm{e}^{\alpha x}-1-\alpha x)\nu(\mathrm{d}x),\quad \int_{|x|>1}(\mathrm{e}^{\alpha x}-1)\nu(\mathrm{d}x).$$

Adding shows that μ has a c.g.f. of the form given by the Lévy–Khintchine representation. □

Problems

1.1 Show that a Lévy process with D–paths is strong Markov.
1.2 Show that $\mathbb{E}\mathrm{e}^{\alpha X_1}\neq 0$ whenever the expectation is welldefined.

Notes Standard references on Lévy processes are Protter (1990), Bertoin (1996) and Sato (1999). A good impression of the many directions in which the topic has been developed and applied can be obtained from the volume edited by Barndorff–Nielsen *et al.* (2001).

An important special case is *α–stable processes* ($0<\alpha<2$), where $\nu(\mathrm{d}x)=c_\pm x^{-\alpha-1}$ with c_+ for $x>0$ and c_- for $x<0$, see Samorodnitsky and Taqqu (1994), another *subordinators* defined as nondecreasing Lévy processes (here $\theta\geq 0$, $\sigma^2=0$ and one has *spectral positivity*, i.e. $\nu(\mathrm{d}x)=0$, $x<0$).

2 Reflection and Loynes's Lemma

2a. Local Time and Reflection
2b. The Skorokhod Problem
2c. Stationarity and Loynes's Lemma
2d. Reflected Lévy Processes

2a Local Time and Reflection

Let $\{S_t\}$ be a real–valued stochastic process with a discrete or continuous nonnegative time parameter $t\in\mathbb{T}$ and $S_0=0$. For $x\geq 0$, we want to define and study the reflected (at 0) version $\{V_t(x)\}$ of $\{S_t\}$ starting from $V_0(x)=x$.

In discrete time, we can just as in III.6 define the increment as $X_n=S_n-S_{n-1}$ and let $V_0(x)=x$,

$$V_{n+1}(x)\;=\;\bigl(V_n(x)+X_n\bigr)^+ \qquad (2.1)$$

(the Lindley recursion). Some more care is needed to incorporate also the continuous–time case. Here we define

$$L_t\;=\;-\inf_{0\leq v\leq t}S_v\;=\;\sup_{0\leq v\leq t}-S_v,\qquad L_t(x)=(L_t-x)^+,$$

$$V_t(x)\;=\;x+S_t+L_t(x)\;=\;S_t+x\vee L_t. \qquad (2.2)$$

The process $\{L_t\}$ is called the *local time* and is nonnegative and increasing (in particular of bounded variation). It follows immediately that $L_t(x) = -x - \inf_{\tau_-(x)\le v\le t} S_v$ for $t \ge \tau_-(x)$ where $\tau_-(x) = \inf\{t > 0 : S_t \le -x\}$, and $L_t(x) = 0$ for $t < \tau_-(x)$. When the initial value x is unimportant, we write V_t instead of $V_t(x)$.

Define the additively shifted version $\{S_t^{(T)}\}_{t\ge 0}$ by $S_t^{(T)} = S_{t+T} - S_T$, and let $L_t^{(T)}(y)$, etc. be defined in terms of $\{S_t^{(T)}\}$ rather than $\{S_t\}$.

Proposition 2.1 $V_{t+T}(x) \;=\; V_T(x) + S_t^{(T)} + L_t^{(T)}\bigl(V_T(x)\bigr)$.

Proof. The stated expression for $V_{t+T} = V_{t+T}(x)$ is the same as

$$\begin{aligned} S_t^{(T)} + V_T \vee L_t^{(T)} &= S_{t+T} - S_T + (S_T + x \vee L_T) \vee \Bigl(S_T - \inf_{0\le v\le t} S_{v+T}\Bigr) \\ &= S_{t+T} + x \vee \Bigl(\sup_{0\le v\le T} -S_v\Bigr) \vee \Bigl(\sup_{0\le v\le t} -S_{v+T}\Bigr) \\ &= S_{t+T} + x \vee \Bigl(\sup_{0\le v\le t+T} -S_v\Bigr) \;=\; V_{t+T}. \end{aligned}$$

□

The result shows that V_{t+T} is constructed from $\{S_t^{(T)}\}$, V_T in the same way as $V_T(x)$ is constructed from $\{S_t\}$, x. If in discrete time we take $t = 1$, $\mathbb{T} = N$, Proposition 2.1 takes the form

$$V_{N+1} \;=\; V_N + X_{N+1} + (X_{N+1}^- - V_N)^+ \;=\; (V_N + X_{N+1})^+,$$

so that we are back to (2.1).

2b The Skorokhod Problem

An alternative characterization of the reflected process is as the solution to a so–called *Skorokhod problem*:

Proposition 2.2 *Let $\{L_t^*\}$ be any nondecreasing right–continuous process such that* (a) *the process $\{V_t^*\}$ given by $V_0^* = x$, $V_t^* = S_t + L_t^*$ satisfies $V_t^* \ge 0$ for all t,* (b) *$\{L_t^*\}$ can increase only when $V_t^* = 0$, i.e. $\int_0^T V_t^*\,\mathrm{d}L_t^* = 0$ for all T. Then $L_t^* = L_t(x)$, $V_t^* = V_t(x)$.*

Proof. Let $D_t = L_t - L_t^*$, $\Delta D_s = D_s - D_{s-}$. The integration–by–parts formula for a right–continuous process of bounded variation gives

$$\begin{aligned} D_t^2 &= 2\int_0^t D_s \mathrm{d}D_s - \sum_{s\le t}(\Delta D_s)^2 \\ &= 2\int_0^t (L_s - L_s^*)\mathrm{d}L_s - 2\int_0^t (L_s - L_s^*)\mathrm{d}L_s^* - \sum_{s\le t}(\Delta D_s)^2 \\ &= 2\int_0^t (V_s - V_s^*)\mathrm{d}L_s - 2\int_0^t (V_s - V_s^*)\mathrm{d}L_s^* - \sum_{s\le t}(\Delta D_s)^2 \end{aligned}$$

$$= \quad -2\int_0^t V_s^* dL_s - 2\int_0^t V_s \mathrm{d}L_s^* - \sum_{s\le t}(\Delta D_s)^2 .$$

Here the two first integrals are nonnegative since V_s^*, V_s are so, and also the sum is clearly so. Thus $D_t^2 \le 0$, which is only possible if $L_t \equiv L_t^*$. □

2c Stationarity and Loynes's Lemma

We will extend the framework slightly by including some supplementary variables. We call a $E \times \mathbb{R}$–valued process $\{(J_t, S_t)\}_{t\ge 0}$ *stationary marked additive* if

$$\bigl\{\bigl(J_t^{(T)}, S_t^{(T)}\bigr)\bigr\}_{t\ge 0} \overset{\mathscr{D}}{=} \{(J_t, S_t)\}_{t\ge 0}\,, \quad T > 0, \tag{2.3}$$

where $J_t^{(T)} = J_{t+T}$. A doubly infinite version is defined by the requirement that the time parameter is $-\infty < t < \infty$ and that (2.3) holds also for $T < 0$ (it is inherent in the definition that one must have $S_0 = 0$ also for a doubly infinite version).

Proposition 2.3 *A stationary marked additive process has always a doubly infinite version.*

Proof. For $T > 0$, $t \ge -T$ define $\widetilde{J}_t^{(-T)} = J_{t+T}$, $\widetilde{S}_t^{(-T)} = S_{t+T} - S_T$. A standard construction based upon Kolmogorov's consistency theorem then shows that there exists a unique probability measure μ on $D(\mathbb{R}, E\times\mathbb{R})$ such that the restriction of μ to $\mathscr{G}_T = \sigma(J_t, S_t : t \ge -T)$ is the same as the distribution of $\bigl\{\bigl(\widetilde{J}_t^{(-T)}, \widetilde{S}_t^{(-T)}\bigr)\bigr\}_{t\ge -T}$. Now just take the doubly infinite version as a process with distribution μ. □

Remark 2.4 In discrete time, the doubly infinite version can be constructed by taking a doubly infinite version of $\{(J_n, X_n)\}_{n\in\mathbb{Z}}$ (exists because obviously $\{(J_n, X_n)\}_{n\in\mathbb{N}}$ is stationary) and letting $S_0 = 0$,

$$S_n \;=\; \begin{cases} -X_{n+1} - \cdots - X_{-1} - X_0 & n < 0, \\ \quad X_1 + \cdots + X_n & n > 0. \end{cases}$$

To see this, just note that the process constructed in this way satisfies $S_n^{(T)} = X_{N+1} + \cdots + X_{n+N}$ (consider the cases $n + N < 0$, $= 0$ and > 0 separately). □

Theorem 2.5 *Let* $\bigl\{(J_t, S_t)\bigr\}_{t\in\mathbb{R}}$ *be a doubly infinite stationary marked additive process and define* $V_T^* = \sup_{-\infty<t\le T}(S_T - S_t)$. *Then* $\bigl\{(J_t, V_t^*)\bigr\}$ *is a doubly infinite stationary version of* $\bigl\{(J_t, V_t)\bigr\}$.

Proof. Stationarity is clear, so that by Proposition 2.1, all that needs to be checked is $V_{t+T} = V_T + S_t^{(T)} + L_t^{(T)}(V_T)$. But the r.h.s. is the same as

$S_t^{(T)} + V_T \vee L_t^{(T)}$ which in turn equals

$$\begin{aligned} & S_{t+T} - S_T + \Bigl(\sup_{-\infty < v \le T} (S_T - S_v) \Bigr) \vee \Bigl(S_T - \inf_{0 \le v \le t} S_{v+T} \Bigr) \\ = \; & S_{t+T} + \sup_{-\infty < v \le t+T} -S_v \; = \; V_{t+T}. \end{aligned}$$

□

In the doubly infinite situation, define $S_t^* = -S_{-t-}$, $J_t^* = -J_{-t-}$. We refer to $\{(J_t^*, S_t^*)\}$ as the *reversed* version of the marked additive process. Note that in discrete time, we have $S_n^* = X_0 + X_{-1} + \cdots + X_{-n+1}$, $n > 0$.

Corollary 2.6 *Assume $S_t \stackrel{\text{a.s.}}{\to} -\infty$, $t \to \infty$. Then $M^* = \sup_{t \ge 0} S_t^*$ is finite a.s., and the distribution of $(J_0^*, M^*) = (J_0, M^*)$ is the unique stationary distribution for $\bigl\{(J_t, V_t)\bigr\}$. Further, for all $x \ge 0$, $\bigl(J_t, V_t(x)\bigr) \to (J_0, M^*)$ in total variation.*

Proof. By coupling: let y be another (possibly random) initial value and $\tau = \inf\{t > 0 : S_t \le -x \vee y\}$. Then $\tau < \infty$ a.s. by assumption (as well as $M^* < \infty$). Since $L_\tau = -S_\tau \ge x \vee y$, we have

$$V_\tau(x) \; = \; S_\tau + x \vee L_\tau \; = \; S_\tau + L_\tau \; = \; 0$$

and similarly $V_\tau(y) = 0$. Proposition 2.1 then implies $V_t(x) = V_t(y)$ for all $t \ge \tau$. Now just consider the doubly infinite stationary situation and take $y = V_0$ which has the same distribution as M^*. □

The following alternative representation is often useful:

Corollary 2.7 *Define $\tau^*(x) = \inf\{t > 0 : S_t^* \ge x\}$. Then $\mathbb{P}_e(J_t \in A, V_t \ge x) = \mathbb{P}(J_0^* \in A, \tau^*(x) < \infty)$. In particular, if E is discrete and $\pi_i = \mathbb{P}_e(J_0^* = i) = \mathbb{P}_e(J_0 = i)$, $\mathbb{P}_i = \mathbb{P}_e(\cdot \mid J^*(0) = i)$, then $\mathbb{P}_e(J_t = i, V_t \ge x) = \pi_i \mathbb{P}_i(\tau^*(x) < \infty)$.*

2d Reflected Lévy Processes

We now assume that $\{S_t\}$ is a Lévy process as in Section 1.

Corollary 2.8 *A reflected Lévy process $\{V_t\}$ is strong Markov. Further, $\mathbb{P}_e(V_t \ge x) = \mathbb{P}(\sup_{t \ge 0} S_t \ge x) = \mathbb{P}(\tau(x) < \infty)$ where $\tau(x) = \inf\{t > 0 : S_t \ge x\}$.*

Proof. The Markov property follows immediately from Proposition 2.1. For the strong Markov property, combine with I.8.3 and an easy continuity argument, cf. Problem 1.1. For the last statement, just note that the time–reversion in Corollary 2.7 does not change the distribution. □

We next continue a study initiated in III.7: how does a reflected Lévy process leave 0? Throughout, $V_t = V_t(0)$.

Lemma 2.9 *Assume that $\{S_t\}$ is a pure jump Lévy process with Lévy measure ν satisfying $\int |x| \wedge 1\, \nu(\mathrm{d}x) < \infty$. Then $S_t/t \overset{\text{a.s.}}{\to} 0$ as $t \downarrow 0$.*

Proof. Denote by $Z = Z(\nu)$ the limit (if it exists) of S_t/t so that we must show $Z(\nu) = 0$ a.s. This is obvious if $\{S_t\}$ is compound Poisson, and since clearly $Z(\nu)$ is additive in ν, we may assume that ν is concentrated on $(0, 1]$.

Now note that $\{S_t/t\}_{0<t\leq 1}$ is a backward martingale (by the standard random walk analogue, this holds for $\big\{S_{k/n}/(k/n)\big\}_{k=1,\ldots,n}$; use an easy continuity argument or mimic the proof). Hence by the martingale convergence theorem in continuous time $Z(\nu)$ exists and satisfies $\mathbb{E}Z(\nu) \leq \mathbb{E}S_1$. Now if $\nu^{(n)}$ denotes the restriction of ν to $(0, 1/n]$ and $S_t^{(n)}$ denotes the sum of the jumps $\leq 1/n$ before t, we have $S_n = S_t^{(n)}$ for all small t, and hence $Z(\nu) = Z(\nu^{(n)})$,

$$\mathbb{E}Z(\nu) \;=\; \lim_{n\to\infty} \mathbb{E}Z\big(\nu^{(n)}\big) \;\leq\; \lim_{n\to\infty} \int_0^{1/n} y\, \nu(\mathrm{d}y) \;=\; 0,$$

so that $Z(\nu) = 0$. □

Example 2.10 Let $S_t = \theta t - Y_t$ where $\theta > 0$ and $\{Y_t\}$ is a pure jump Lévy process whose Lévy measure is infinite and concentrated on $(0, \infty)$. Then by Lemma 2.9, we have $S_v > 0$ for all small v, say $0 < v \leq v_0$, and hence $L_t = 0$ and $V_t = S_t$ for $t \leq v_0$. That is, $\{V_t\}$ *leaves state 0 instantaneously and in the same way as* $\{S_t\}$. □

Example 2.11 Let $\theta > 0$, say $\theta = 1$ and let $\{Y_t\}$ be as in Example 2.10, but take now $S_t = Y_t - \theta t$. We shall see that this case is much more complicated than Example 2.10 in the sense that $\tau_+ = \inf\{t > 0 : V_t > 0\}$ *is still zero but that for any* $\epsilon > 0$ *the Lebesgue measure of* $A_\epsilon = \{t \in [0, \epsilon] : V_t = 0\}$ *is nonzero.* To this end, we let $Y_t^{(n)}$, $V_t^{(n)}$ etc. refer to the case where all jumps $\leq 1/n$ have been neglected. Clearly, $Y_t^{(n)} - Y_v^{(n)} \uparrow Y_t - Y_v$ for $v < t$ and hence $V_t^{(n)} \uparrow V_t$. Now since $\big\{Y_t^{(n)}\big\}$ is compound Poisson, each $\big\{V_t^{(n)}\big\}$ is simply of $M/G/1$ workload form, i.e. decreases at a unit rate in states > 0 and has the same upward jumps as $\big\{Y_t^{(n)}\big\}$. From this it is obvious that

$$V_t^{(n)} \;=\; Y_t^{(n)} \;-\; \int_0^t I(V_t^{(n)} > 0)\, \mathrm{d}s,$$

and we may pass to the limit $n \to \infty$ to get

$$V_t(0) \;=\; Y_t \;-\; \int_0^t I(V_t > 0)\, \mathrm{d}s. \tag{2.4}$$

Now let $\epsilon > 0$ satisfy $Y_t \leq t/2$, $t \leq \epsilon$, and assume $|A_\epsilon| = 0$. Then (2.4) yields $V_\epsilon = S_\epsilon = Y_\epsilon - \epsilon < 0$, which is impossible. Also, $\tau_+ = 0$ a.s. follows, since if $\{Y_t\}$ jumps say δ at time t, then $V_t \geq \delta$; further, since ν is infinite, the jump times have 0 as accumulation point. □

Notes Theorem 2.5 and the many variants such as Corollaries 2.6 and 2.7, which are around, are often referred to as *Loynes's lemma* after Loynes (1962).

For Skorokhod problems, see e.g. example Rogers and Williams (1994), Revuz and Yor (1999) and Whitt (2002). The theory is more difficult in multidimensions. Tanaka (1979) is a classical reference for diffusions. A special problem on so–called *oblique reflection* comes up in connection with the heavy–traffic limits for queueing networks mentioned in the Notes to IV.5 and can be formulated as follows: for a given matrix $\boldsymbol{R}$ and a given D–function $\boldsymbol{x}(t)$ with values in $\mathbb{R}^K$ and $x_i(0) \geq 0$ for all $i = 1, \ldots, K$, find functions $\boldsymbol{v}(t)$ and $\boldsymbol{\ell}(t)$ (and show they are unique) such that $\boldsymbol{v}(0) = \boldsymbol{x}(0)$, $\boldsymbol{v}(0) \geq \boldsymbol{0}$, $\boldsymbol{v}(t) = \boldsymbol{x}(t) + \boldsymbol{R\ell}(t) \geq \boldsymbol{0}$, and each $\ell_i(t)$ in nondecreasing and can only increase when $v_i(t) = 0$. It is easily seen that conditions on $\boldsymbol{R}$ are required for this problem to be meaningful. See e.g. Harrison and Reiman (1981) and Chen and Yao (2001).

3 Martingales and Transforms for Reflected Lévy Processes

Let $\{S_t\}$ be a Lévy process with Lévy exponent $\kappa(\alpha)$. The Wald martingale is then $M_t = \mathrm{e}^{\alpha S_t - t\kappa(\alpha)}$; for some typical applications of this martingale, see III.8d and VIII.5.4. We now study a martingale obtained as a stochastic integral w.r.t. $\{M_t\}$ and which has a somewhat different range of applications; in particular, it allows for a more direct study of aspects of reflected Lévy processes.

Theorem 3.1 *Let $\{S_t\}$ be a Lévy process with Lévy exponent $\kappa(\alpha)$, let*

$$Y_t \;=\; \int_0^t \mathrm{d}Y_s^c \,+\, \sum_{0 \leq s \leq t} \Delta Y_s$$

be an adapted process of locally bounded variation with continuous part $\{Y_t^c\}$, D–paths and jumps $\Delta Y_s = Y_s - Y_{s-}$, and define $Z_t = x + S_t + Y_t$. For each t, let K_t be the r.v.

$$\kappa(\alpha)\int_0^t \mathrm{e}^{\alpha Z_s}\,\mathrm{d}s \,+\, \mathrm{e}^{\alpha x} \,-\, \mathrm{e}^{\alpha Z_t} \,+\, \alpha\int_0^t \mathrm{e}^{\alpha Z_s}\,\mathrm{d}Y_s^c \,+\, \sum_{0 \leq s \leq t} \mathrm{e}^{\alpha Z_s}(1 - \mathrm{e}^{-\alpha \Delta Y_s}).$$

Then $\{K_t\}$ is a local martingale whenever $\alpha \in \Theta$.

Proof. Let $B_t = \mathrm{e}^{\alpha Y_t + t\kappa(\alpha)}$. Then, by the general theory of stochastic integration, $K_t^* = \int_0^t B_{s-}\,\mathrm{d}M_s$ is a local martingale. Using the formula for integration by parts (see Protter, 1990, p. 60, for a version sufficiently general to deal with the present case) yields

$$M_tB_t - M_0B_0 \;=\; \int_0^t M_{s-}\,\mathrm{d}B_s \,+\, K_t^* \,+\, \sum_{0\le s\le t} \Delta M_s\Delta B_s.$$

Inserting

$$\sum_{0\le s\le t} \Delta M_s\Delta B_s \;=\; \int_0^t \Delta M_s\,\mathrm{d}B_s \;=\; \int_0^t (M_s - M_{s-})\,\mathrm{d}B_s,$$

it follows that

$$-K_t^* \;=\; \int_0^t M_s\,\mathrm{d}B_s \,+\, M_0B_0 \,-\, M_tB_t. \tag{3.1}$$

Using $M_sB_s = \mathrm{e}^{\alpha Z_s}$ and $\mathrm{d}B_s = B_s(\alpha\,\mathrm{d}Y_s^c + \kappa(\alpha)\mathrm{d}s + 1 - \mathrm{e}^{-\alpha\Delta Y_s})$ shows that the r.h.s. of (3.1) reduces to K_t. □

Let as in Section 2 $L_t = \inf_{0\le v\le t} S_v$ denote the local time, $L_t(x) = (L_t - x)^+$, and $V_t = V_t(x) = x + S_t + L_t(x)$ the reflected version of $\{S_t\}$ starting from $V_0 = x$. For simplicity, we will present most of the applications of Theorem 3.1 in the following setting:

Corollary 3.2 *Let $\{S_t\}$ be a Lévy process with no negative jumps, $\nu(-\infty,0) = 0$. Let $x \ge 0$ and consider $V_t = V_t(x)$. Then for $\kappa(\alpha) < \infty$, the process $\{K_t'\}$ defined by*

$$K_t' \;=\; \kappa(\alpha)\int_0^t \mathrm{e}^{\alpha V_s}\,\mathrm{d}s \,+\, \mathrm{e}^{\alpha x} \,-\, \mathrm{e}^{\alpha V_t} \,+\, \alpha L_t(x)$$

is a martingale. More generally, for any $\beta < 0$

$$K_t'' \;=\; \kappa(\alpha)\int_0^t \mathrm{e}^{\alpha V_s + \beta L_s}\,\mathrm{d}s \,-\, \mathrm{e}^{\alpha x} \,+\, \mathrm{e}^{\alpha V_t + \beta t} \,+\, (1 + \alpha/\beta)(\mathrm{e}^{\beta L_t} - 1)$$

defines a martingale. If τ is a stopping time, then a sufficient condition that either of K_τ', K_τ'' are integrable with mean 0 is that $\sup_{t\le\tau} \mathrm{e}^{\alpha V_t}$ and $\tau\sup_{t\le\tau} \mathrm{e}^{\alpha V_t}$ are both integrable.

For the proof, we need:

Lemma 3.3 (a) *If $\kappa(\alpha) < \infty$, then $\mathbb{E}_x \sup_{0\le s\le t} \mathrm{e}^{\alpha V_s} < \infty$ for all t;*
(b) *if $\mathbb{E}|S_1| < \infty$, then also $\mathbb{E}V_t < \infty$ and $\mathbb{E}L_t < \infty$ for all t.*

Proof. It is easy to see that $0 \le V_t(x) - V_t(0) \le x$ so we may assume $x = 0$. Then by Corollary 2.7, we have

$$\mathbb{P}_0(V_t \ge z) \;=\; \mathbb{P}(\tau(z) \le t) \tag{3.2}$$

where $\tau(z) = \inf\{t : S_t \ge z\}$. Choose $\epsilon, y > 0$ with $\mathbb{P}(L_t \le y) \ge \epsilon$ and let $\tau = \inf\{t : V_t > z + y\}$. Then

$$\mathbb{P}(S_t > z - y) \;\ge\; \mathbb{P}(\tau(z) \le t, S_t - S_{\tau(z)} > -y)$$

$$\begin{aligned} &\geq \epsilon\mathbb{P}(\tau(z) \leq t) = \epsilon\mathbb{P}_0(V_t \geq z) \geq \epsilon\mathbb{P}(\tau \leq t, S_t - S_\tau > -y) \\ &\geq \epsilon^2\mathbb{P}(\tau \leq t) = \epsilon^2\mathbb{P}_0\big(\sup_{0\leq s\leq t} V_s > z + y\big). \end{aligned}$$

That $\mathbb{E}_0 \sup_{0\leq s\leq t} \mathrm{e}^{\alpha V_s} < \infty$ now follows from $\mathbb{E}\mathrm{e}^{\alpha S_t} < \infty$ by using integration by parts which also yields $\mathbb{E}_0 V_t < \infty$ and hence $\mathbb{E}_0 L_t = \mathbb{E}_0 V_t - \mathbb{E}S_t < \infty$ when $\mathbb{E}|S_1| < \infty$. □

Proof of Corollary 3.2. Consider $\{K_t''\}$ (the case of $\{K_t'\}$ is similar but easier), where we take $Y_t = (1+\beta/\alpha)L_t(x)$. Note that $\{L_t\}$ has continuous paths in the case of no negative jumps so that the last term in the definition of K_t in Theorem 3.1 vanishes, and that $L_t(x)$ only increases when $V_t = 0$ so that the term involving $\mathrm{d}Y_s^c$ is

$$\begin{aligned} &\alpha \int_0^t \mathrm{e}^{\beta L_s(x)}(1+\beta/\alpha)\,\mathrm{d}L_s(x) \\ &\quad = (\alpha+\beta)\int_0^t \mathrm{e}^{\beta L_s(x)}\,\mathrm{d}L_s(x) = (1+\alpha/\beta)(\mathrm{e}^{\beta L_t(x)} - 1)\,. \end{aligned}$$

Therefore $K_t'' = K_t$. That $\{K_t''\}$ is a martingale and not just a local martingale follows since $\sup_{s\leq t} K_s''$ is integrable by Lemma 3.3 (see Protter, 1990, p. 35).

For the optional stopping problem, one has as always that $\mathbb{E}K''_{\tau\wedge t} = \mathbb{E}K_0''$ $= 0$. However, by dominated convergence

$$\mathbb{E}\int_0^{\tau\wedge t} \mathrm{e}^{\alpha V_s + \beta L_s}\,\mathrm{d}s \to \mathbb{E}\int_0^{\tau} \mathrm{e}^{\alpha V_s + \beta L_s}\,\mathrm{d}s, \quad \mathbb{E}\mathrm{e}^{\alpha V_{\tau\wedge t} + \beta L_{\tau\wedge t}} \to \mathbb{E}\mathrm{e}^{\alpha V_\tau + \beta L_\tau}$$

where both limits are finite. This implies $0 = \mathbb{E}K''_{\tau\wedge t} \to \mathbb{E}K''_\tau$. □

Recall the Pollaczeck–Khinchine formula in Problem VIII.5.2 for the m.g.f. of the steady–state $M/G/1$ workload. Here is a more general version:

Corollary 3.4 *Consider a Lévy process with no negative jumps and negative drift,* $-\infty < \mu = \kappa'(0) = \mathbb{E}S_1 < 0$. *Then the limit* V *in distribution of* V_t *exists and for* $\Re\alpha < 0$ *one has* $\mathbb{E}\mathrm{e}^{\alpha V} = \alpha\mu/\kappa(\alpha)$.

Proof. The existence of V was noted in Section 2. Take $\beta = 0$ and choose $V_0 = V^*$ as a r.v. distributed as V and independent of $\{S_t\}$. Then $\{V_t\}$ is stationary and we get

$$0 = \mathbb{E}K_1' = \kappa(\alpha)\mathbb{E}\mathrm{e}^{\alpha V} - \mathbb{E}\mathrm{e}^{\alpha V^*} + \mathbb{E}\mathrm{e}^{\alpha V_1} + \alpha\mathbb{E}L_1 = \kappa(\alpha)\mathbb{E}\mathrm{e}^{\alpha V} + \alpha\mathbb{E}L_1,$$

so that it only remains to show that in a stationary process $\mathbb{E}L_t = -t\mu$. However, if $\mathbb{E}V < \infty$, this follows from $V_t = V_0 + S_t + L_t$ and $\mathbb{E}V_t = \mathbb{E}V_0 = \mathbb{E}V$ in stationarity. The general case follows by an easy truncation argument; see Problem X.2.3. □

Example 3.5 Let $\{S_t\}$ be Brownian motion with drift $\mu < 0$ and variance 1 and $\{V_t\}$ the reflected version. We then get

$$\mathbb{E}e^{\alpha V} \;=\; \frac{\alpha\mu}{\alpha\mu+\alpha^2/2} \;=\; \frac{2|\mu|}{2|\mu|-\alpha},$$

which shows that V is exponential with intensity $2|\mu|$.

It is instructive to compare with an alternative proof using the Wald martingale and the limiting case $\mathbb{P}(V \ge y) = p$ where $p = \mathbb{P}(\tau(y) < \infty)$ of (3.2). We have to show $p = \mathrm{e}^{-2|\mu|y}$. Since $\gamma = -2\mu$ solves $\kappa(\gamma) = 0$, letting $\alpha = \gamma$ yields $M_t = \mathrm{e}^{-2\mu S_t}$ and

$$1 \;=\; \mathbb{E}M_{\tau(y)\wedge t} \;=\; \mathrm{e}^{-2\mu y}\mathbb{P}(\tau(y) \le t) \;+\; \mathbb{E}\big[\mathrm{e}^{-2\mu S_t};\, \tau(y) > t\big].$$

Since $S_t \overset{\text{a.s.}}{\to} -\infty$ and $-2\mu S_t \le 2|\mu|y$ on $\{\tau(y) > t\}$, dominated convergence yields $1 = \mathrm{e}^{-2\mu y}p + 0$ and the desired conclusion. □

Example 3.6 Let $\{S_t\}$ be Brownian motion with drift μ and variance 1 and $\{V_t\}$ the two–sided reflected version on $[0, K]$, $V_t = V_0 + S_t + L_t^{(0)} - L_t^{(K)}$ where $L_t^{(0)}, L_t^{(K)}$ are the local times at 0, resp. K as defined as the solution to an obvious generalization of the Skorokhod problem (see further XIV.3). Letting $Y_t = L_t^{(0)} - L_t^{(K)}$, we get

$$K_t' \;=\; \kappa(\alpha)\int_0^t \mathrm{e}^{\alpha V_s}\,\mathrm{d}s \;-\; \mathrm{e}^{\alpha x} \;+\; \mathrm{e}^{\alpha V_t} \;+\; \alpha L_t^{(0)} \;-\; \alpha\mathrm{e}^{\alpha K}L_t^{(K)}$$

and equating the expectation in stationarity to 0 ($x = V^*$ as above) yields

$$0 \;=\; (\alpha\mu + \alpha^2/2)\mathbb{E}\mathrm{e}^{\alpha V} \;+\; \alpha\ell^{(0)} \;-\; \alpha\mathrm{e}^{\alpha K}\ell^{(K)}$$

where $\ell^{(0)} = \mathbb{E}L_t^{(0)}/t$, $\ell^{(K)} = \mathbb{E}L_t^{(K)}/t$. In particular, taking $\alpha = -2\mu$ we get $\ell^{(K)} = \mathrm{e}^{2\mu K}\ell^{(0)}$. Also, $\mathbb{E}V_t = \mathbb{E}V_0 = \mathbb{E}V$ gives $\ell^{(0)} - \ell^{(K)} + \mu = 0$. It follows by easy algebra that $\ell^{(0)} = \mu/(\mathrm{e}^{2\mu K} - 1)$, $\ell^{(K)} = \mu\mathrm{e}^{2\mu K}/(\mathrm{e}^{2\mu K} - 1)$, and hence

$$\mathbb{E}\mathrm{e}^{\alpha V} \;=\; \frac{\mu\mathrm{e}^{\alpha K + 2\mu K} - \mu}{(\mathrm{e}^{2\mu K} - 1)(\mu + \alpha/2)},$$

which is readily seen to be the m.g.f. of a truncated exponential distribution with density $2\mu\mathrm{e}^{2\mu x}/(\mathrm{e}^{2\mu K} - 1)$, $0 \le x \le K$. □

Wald's identity $\mu\mathbb{E}\tau = \mathbb{E}S_\tau$ allows to find an expected stopping time when the distribution of S_τ is accessible. Here is a similar formula for the reflected case (by stopping time, we mean stopping time for $\{S_t\}$):

Corollary 3.7 *Consider a reflected Lévy process with no negative jumps and $V_0 = x$. Assume that a $\gamma \ne 0$ with $\kappa(\gamma) = 0$ exists. Let τ be a stopping time such that* $\sup_{t\le\tau}\mathrm{e}^{\gamma V_t}$ *and* $\tau\sup_{t\le\tau}\mathrm{e}^{\gamma V_t}$ *are both integrable. Then*

$$\mathbb{E}_x\tau \;=\; \frac{\gamma\mathbb{E}_x V_\tau - \mathbb{E}_x\mathrm{e}^{\gamma V_\tau} + \mathrm{e}^{\gamma x} - \gamma x}{\gamma\kappa'(0)}\,.$$

Proof. Taking $\alpha = \gamma$ and using optional stopping ($\mathbb{E}_x K'_\tau = 0$) in Corollary 3.2 yields $0 = \mathrm{e}^{\gamma x} - \mathbb{E}_x \mathrm{e}^{\gamma V_\tau} + \gamma \mathbb{E}_x L_\tau(x)$. On the other hand, from Wald's identity

$$\mathbb{E}_x \tau \;=\; \frac{\mathbb{E}_x S_\tau}{\kappa'(0)} \;=\; \frac{\mathbb{E}_x V_\tau - x - \mathbb{E}_x L_\tau(x)}{\kappa'(0)}\,. \qquad \Box$$

Example 3.8 Let $\{V_t\}$ be reflected Brownian motion with drift $\mu \neq 0$ and variance 1, and let $\tau = \inf\{t > 0 : V_t = 1\}$. Then $\kappa(\alpha) = \alpha\mu + \alpha^2/2$ which shows that $\gamma = -2\mu$ solves $\kappa(\gamma) = 0$. Hence $\mathbb{E}[\tau \,|\, V_0 = 0] = (2\mu + e^{-2\mu} - 1)/(2\mu^2)$; cf. VI.2c. □

Example 3.9 For an example involving negative jumps, consider the $M/M/1$ queue length process $\{Q_t\}$ with arrival intensity β and service intensity δ and the problem of evaluating $\mathbb{E}_x \tau$ where $\tau = \inf\{t > 0 : Q_t = n\}$. We will assume $x = Q_0 < n$ and then τ can be interpreted as the first buffer overflow time

Take $\{S_t\}$ as the independent difference between two independent Poisson processes with intensities β, resp. δ (then $\kappa(\alpha) = \beta(\mathrm{e}^{\alpha} - 1) + \delta(\mathrm{e}^{-\alpha} - 1)$), and $Y_t = L_t(x)$. Then $Z_t = Q_t$ in Theorem 3.1. Taking $\alpha = \gamma = \log\delta - \log\beta$, we have $\kappa(\gamma) = 0$, $\mathrm{e}^{-\gamma} = \rho$, and the martingale is

$$K_t \;=\; 0 - \mathrm{e}^{\gamma x} + \mathrm{e}^{\gamma Q_t} + 0 + L_t(x)(1 - \mathrm{e}^{-\gamma}) \;=\; -\rho^{-x} + \rho^{-Q_t} + L_t(x)(1-\rho)$$

(for the form of the last term note that $L_t(x)$ is the number of dummy service events in the idle state before t so that the jumps are 1 and only occurs at times where $Q_t = 0$). As in the proof of Corollary 3.7 we therefore get

$$\begin{aligned} 0 \;&=\; -\rho^{-x} + \rho^{-n} + \mathbb{E}_x L_\tau(x)(1-\rho), \\ \mathbb{E}_x \tau \;&=\; \frac{\mathbb{E}_x S_\tau}{\kappa'(0)} \;=\; \frac{\mathbb{E}_x Q_\tau - x - \mathbb{E}_x L_\tau(x)}{\beta - \delta} \;=\; \frac{(1-\rho)(n-x) - \rho^{-x} + \rho^{-n}}{(1-\rho)(\beta - \delta)}\,. \end{aligned}$$

□

Theorem 3.10 *Let $\{S_t\}$ be a Lévy process with no negative jumps, let $x \geq 0$ and consider $V_t = V_t(x)$. Then for each $\alpha, \beta < 0$ and $\delta > 0$,*

$$\int_0^\infty \mathrm{e}^{-\delta s} \mathbb{E}_x \mathrm{e}^{\alpha V_s + \beta L_s(x)}\, \mathrm{d}s \;=\; \frac{\mathrm{e}^{\alpha x}(\rho + \beta) - \mathrm{e}^{\rho x}(\alpha + \beta)}{(\delta - \kappa(\alpha))(\rho + \beta)} \tag{3.3}$$

where $\rho = \rho(\delta)$ is the negative root of $\kappa(\rho) = \delta$.

Proof. Let I denote the l.h.s. of (3.3) and let T be exponential with intensity δ and independent of $\{S_t\}$. Then

$$\mathbb{E}_x \int_0^T \mathrm{e}^{\alpha V_s + \beta L_s(x)}\, \mathrm{d}s \;=\; I \;=\; \delta^{-1} \mathbb{E}_x \mathrm{e}^{\alpha V_T + \beta L_T(x)}.$$

Taking $Y_t = (1+\beta/\alpha)L_t$ and using optional stopping at T in Theorem 3.1 (justified when $\alpha, \beta < 0$) therefore yields

$$0 \;=\; (\kappa(\alpha)-\delta)I \,+\, \mathrm{e}^{\alpha x} \,+\, (\alpha+\beta)\int_0^\infty \mathbb{E}\mathrm{e}^{\beta L_s(x)-\delta s}\,\mathrm{d}L_s.$$

Replacing α by ρ shows that the last integral is $-\mathrm{e}^{\rho x}/(\rho+\beta)$. □

Problems

3.1 Discuss the case $x > n$ in Example 3.9.
3.2 Find $\mathbb{E}\mathrm{e}^{\beta\tau}$ in Example 3.9. Do the same if instead of the $M/M/1$ queue length process one considers reflected Brownian motion or the $M/M/1$ workload process.
3.3 Let $\{S_t\}$ be Brownian motion with drift μ and variance 1 and T an independent exponential r.v. with intensity δ. Show that L_T and $V_T = S_T + L_T$ are independent exponential r.v.'s with intensities $\sqrt{\mu^2+2\delta}-\mu$, resp. $\sqrt{\mu^2+2\delta}+\mu$.
3.4 Show III.(8.13).

Notes Theorem 3.1 is from Kella and Whitt (1992); more or less related martingale techniques allowing to incorporate local time appear in Baccelli and Makowski (1989) and Revuz and Yor (1999), Ch. VI.2.4. Further references on martingale techniques in queueing theory include Robert (2000) and Rougham and Pearce (2002).

Corollary 3.7 is from Asmussen and Kella (2001) (their conditions for optional stopping are somewhat sharper than here). Corollary 3.4 and Theorem 3.10 are classical for Lévy processes, see e.g. Prabhu (1980, pp. 76–77) and Bertoin (1999).

A restriction of the approach of this section is that in many problems one needs to control the distribution of V_τ which often is only possible for processes that are skip–free or have only exponential jumps in one direction. The Markov additive extension in Asmussen and Kella (2000) allows, however, for phase–type jumps.

4 A More General Duality

We are concerned with extensions in two directions of the relations

$$\mathbb{P}(W_n \ge x) \;=\; \mathbb{P}(M_n \ge x) \;=\; \mathbb{P}(\tau(x) \le n), \tag{4.1}$$
$$\mathbb{P}(W \ge x) \;=\; \mathbb{P}(M \ge x) \;=\; \mathbb{P}(\tau(x) < \infty) \tag{4.2}$$

for the reflected version $\{W_n\}$ of a random walk $\{S_n\}$ (here $\tau(x) = \inf\{n \ge 1 : S_n \ge x\}$ is the ruin probability), cf. III.6 and the Loynes analogue in Section 2.

The first extension is to more general Markov processes. Let $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = [0,\infty)$, let $\{V_t\}_{t\in\mathbb{T}}$ be Markov with state space $E = [0,\infty)$ or $E = \mathbb{N}$, and let $V_t(x)$ be the version starting from $V_0 = x$. Then $\{V_t\}$ is *stochastically monotone* if $x \le y$ implies $V_t(x) \le_{\mathrm{so}} V_t(y)$ (stochastical ordering, cf. A4) for all $t \in \mathbb{T}$, i.e. if $\mathbb{P}_x(V_t \ge z) \le \mathbb{P}_y(V_t \ge z)$ for all t and z.

Proposition 4.1 *The existence of a Markov process* $\{R_t\}_{t\in\mathbb{T}}$ *on* $E\cup\{\infty\}$ *such that*

$$\mathbb{P}_x(V_t \geq y) \;=\; \mathbb{P}_y(R_t \leq x) \tag{4.3}$$

is equivalent to (i) $\{V_t\}$ *is stochastically monotone and* (ii) $\mathbb{P}_x(V_t \geq y)$ *is a right–continuous function of* x *for all* t *and* y.

Proof. If $\{R_t\}$ exists, the l.h.s. of (4.3) is nondecreasing and right–continuous in x and so necessity of (i), (ii) is clear. If conversely (i), (ii) hold, then the r.h.s. of (4.3) defines a probability measure $P^t(y,\cdot)$ (thus $P^t(y,\{\infty\} = 1-\lim_{x\to\infty}\mathbb{P}_x(V_t \geq y))$, and we shall show that the Chapman–Kolmogorov equations $P^{t+s} = P^tP^s$ hold; cf. I.8. This follows since

$$\begin{aligned} P^{t+s}\bigl(y,[0,x]\bigr) \;&=\; \mathbb{P}_x(V_{t+s}\geq y) \;=\; \int_E \mathbb{P}_x(V_t\in \mathrm{d}z)\mathbb{P}_z(V_s\geq y) \\ &=\; \int_E \mathbb{P}_x(V_t\in\mathrm{d}z)\int_0^z P^s(y,\mathrm{d}u) \;=\; \int_0^z P^s(y,\mathrm{d}u)\mathbb{P}_x(V_t\geq u) \\ &=\; \int_0^z P^s(y,\mathrm{d}u)P^t\bigl(u,[0,x]\bigr) \;=\; (P^tP^s)\bigl(y,[0,x]\bigr). \end{aligned}$$

□

Theorem 4.2 *The state* 0 *is absorbing for* $\{R_t\}$. *Furthermore, letting* $\tau = \inf\{t>0:\ R_t(x)\leq 0\} = \inf\{t>0:\ R_t(x)=0\}$, *one has*

$$\mathbb{P}_0(V_T\geq x) \;=\; \mathbb{P}_x(\tau\leq T), \tag{4.4}$$

and if V_t *converges in total variation, say to* V, *then*

$$\mathbb{P}_0(V\geq x) \;=\; \mathbb{P}_x(\tau<\infty), \tag{4.5}$$

Proof. Taking $x=y=0$ in (4.3) yields $\mathbb{P}_0(R_t\leq 0) = \mathbb{P}_0(V_t\geq 0) = 1$ so that indeed 0 is absorbing for $\{R_t\}$. We then get

$$\mathbb{P}_x(\tau\leq T) \;=\; \mathbb{P}_x(R_T\leq 0) \;=\; \mathbb{P}_0(V_T\geq x).$$

□

We turn to the second extension of (4.1) which does not require the Markov property but, however, works more easily when $\mathbb{T}=\mathbb{N}$ than when $\mathbb{T}=[0,\infty)$. We there assume that $\{V_n\}_{n\in\mathbb{N}}$ is generated by a recursion of the form

$$V_{n+1} \;=\; f(V_n,U_n), \tag{4.6}$$

where $\{U_n\}$ (the *driving sequence*) is a stationary sequence of random elements taking values in some arbitrary space F and $f:\ E\times F\to E$ is a function. The (time–homogeneous) Markov case arises when the U_n are i.i.d. (w.l.o.g., uniform on $F=(0,1)$), but also much more general examples are incorporated. We shall need the following easily proved lemma.

Lemma 4.3 *Assume that $f(x,u)$ is continuous and nondecreasing in x for each fixed $u \in F$ and define $g(x,u) = \inf\{y : f(y,u) \ge x\}$. Then for fixed u $g(x,u)$ is left–continuous in x, nondecreasing in x and strictly increasing on the interval $\{x : 0 < g(x,u) < \infty\}$. Further, $f(y,u) = \sup\{x : g(x,u) \le y\}$ and*

$$g(x,u) \le y \iff f(y,u) \ge x. \tag{4.7}$$

W.l.o.g., we can take $\{U_n\}$ with doubly infinite time, $n \in \mathbb{Z}$, and define the dual process $\{R_n\}_{n\in\mathbb{N}}$ by

$$R_{n+1} \;=\; g(R_n, U_{-n}), \quad n \in \mathbb{N}; \tag{4.8}$$

when the initial value $x = R_0$ is important, we write $R_n(x)$.

Theorem 4.4 *Equations* (4.3) *and* (4.5) *also hold in the set–up of* (4.6) *and* (4.8).

Proof. For $T \in \mathbb{N}$, define $V_0^{(T)}(y) = y$,

$$V_1^{(T)}(y) = f\bigl(V_0^{(T)}(y), U_{-(T-1)}\bigr), \;\ldots\; , V_T^{(T)}(y) = f\bigl(V_{T-1}^{(T)}(y), U_0\bigr).$$

We shall show by induction that

$$V_T^{(T)}(y) \ge x \iff R_T(x) \le y \tag{4.9}$$

(from this (4.3) follows by taking expectations and using the stationarity; since $g(0,u) = 0$, (4.4) then follows as above). The case $T = 0$ of (4.9) is the tautology $y \ge x \iff x \le y$. Assume (4.9) shown for T. Replacing y by $f(y, U_{-T})$ then yields

$$V_T^{(T)}\bigl(f(y,U_{-T})\bigr) \ge x \iff R_T(x) \le f(y,U_{-T}).$$

But $V_T^{(T)}\bigl(f(y,U_{-T})\bigr) = V_{T+1}^{(T+1)}(y)$ and by (4.7),

$$R_T(x) \le f(y,U_{-T}) \quad\iff\quad R_{T+1}(x) \;=\; g(R_T(x), U_{-T}) \;\le\; y.$$

Hence (4.9) holds for $T+1$. □

Example 4.5 Consider a reflected random walk $V_{n+1} = (V_n + X_n)^+$ with increments $X_0, X_1, \ldots$ which are i.i.d. or, more generally, stationary.

In the set–up of Proposition 4.1 and Theorem 4.2, we need (for the Markov property) to assume that $X_0, X_1, \ldots$ are i.i.d. We take $E = [0,\infty)$ and for $y > 0$, we then get

$$\mathbb{P}_y(R_1 \le x) \;=\; \mathbb{P}_x(V_1 \ge y) \;=\; \mathbb{P}(x + X_0 \ge y) \;=\; \mathbb{P}(y - X_0 \le x).$$

For $y = 0$, we have $\mathbb{P}_0(R_1 = 0) = 1$. These two formulas show that $\{R_n\}$ evolves as a random walk with increments $-X_0, -X_1, \ldots$ as long as $R_n > 0$, i.e. $R_n(x) = x - S_n$, $n < \tau$, $R_n(x) = 0$, $n \ge \tau$; when $(-\infty, 0]$ is hit, the value is instantaneously reset to 0 and $\{R_n\}$ then stays in 0 forever. We see further that in the setting of (4.1) we can identify τ and $\tau(x)$, and thus (4.5), (4.4) are the same as (4.1), (4.2).

Consider instead the approach via Theorem 4.4 (which allows for incremenst that are just stationary). We let again $E = [0,\infty)$, take $U_k = X_k$ and $f(x,u) = (x+u)^+$. It is easily seen that $g(y,u) = (y-u)^+$ and so $\{R_n\}$ evolves as a random walk with increments $-X_0, -X_{-1}, \ldots$ as long as $R_n > 0$, while 0 is absorbing. With $\breve{S}_n = -X_0 - X_{-1} - \cdots - X_{-n+1}$, $S_n^* = -\breve{S}_n$ it follows that $\tau = \inf\left\{n : x + \breve{S}_n \le 0\right\} = \inf\{n : S_n^* \ge x\}$, and this last expression shows that (4.4) is the same as in Loynes' lemma in the form of Corollary 2.7. □

Example 4.6 Consider again the setting of Example 4.5 but now with two reflecting barriers 0 and $B > 0$. That is,

$$V_{n+1} \;=\; \min\bigl[B, (V_n + X_n)^+\bigr]. \tag{4.10}$$

For Theorem 4.2, we take $X_0, X_1, \ldots$ i.i.d. and $E = [0,\infty)$ (not $[0,B]$!). For $y > B$, we then get

$$\mathbb{P}_y(R_1 \le x) \;=\; \mathbb{P}_x(V_1 \ge y) \;\le\; \mathbb{P}_x(V_1 > B) \;=\; 0$$

for all x, i.e. $\mathbb{P}_y(R_1 = \infty) = 1$. For $0 \le y \le B$, $\mathbb{P}_y(R_1 \le x) = \mathbb{P}_x(V_1 \ge y)$ becomes

$$\mathbb{P}\bigl((x+X_0)^+ \ge y\bigr) \;=\; \begin{cases} 1 & y = 0, \\ \mathbb{P}(y - X_0 \le x) & 0 < y \le B. \end{cases}$$

Combining these facts show that $\{R_n\}$ evolves as a random walk with increments $-X_0, -X_1, \ldots$ as long as $R_n \in (0,b]$. States 0 and ∞ are absorbing, and from $y > B$ $\{R_n\}$ is in the next step absorbed at ∞. Thus for $R_0 = x \in (0,B]$, absorbtion at 0 before N, i.e. $\tau \le N$, cannot occur if (B,∞) is entered and with $S_n = X_0 + \cdots + X_{n-1}$, $\tau[u,v) = \inf\bigl\{n \ge 0 : S_n \notin [u,v)\bigr\}$, $u \le 0 < v$, we get

$$\begin{aligned} \mathbb{P}_0(V_N \ge x) \;&=\; \mathbb{P}_x(\tau \le N) \\ &=\; \mathbb{P}\bigl(\tau[x-B,x) \le N, S_{\tau[x-B,x)} \ge x\bigr), \qquad (4.11) \\ \mathbb{P}(V \ge x) \;&=\; \mathbb{P}(S_{\tau[x-B,x)} \ge x) \qquad (4.12) \end{aligned}$$

(note that $\tau[x-B,x)$ is always finite). □

Example 4.7 Let $\{V_t\}$ be a diffusion on $[0,\infty)$ with differential generator $\mathscr{A}f(y) = a(y)f''(y)/2 + b(y)f'(y)$ (the domain will not be specified but certainly contains the class $\mathscr{K}$ of all C_2 functions with compact support contained in $(0,\infty)$). It seems reasonable to guess that $\{R_t\}$ is again a diffusion, and we shall verify that indeed the differential generator $\mathscr{B}$ is

$$\mathscr{B}f(x) \;=\; \frac{1}{2}a(x)f''(x) + \Bigl[\frac{1}{2}a'(x) - b(x)\Bigr]f'(x). \tag{4.13}$$

To this end, let $f, g \in \mathscr{K}$. Then

$$0 - 0 \;=\; [fg]_0^\infty \;=\; \int_0^\infty f'(x)\,\mathrm{d}x \;=\; \int_0^\infty g'(y)\,\mathrm{d}y, \tag{4.14}$$

$$0-0 \;=\; [fg]_0^\infty \;=\; \int_0^\infty (f'g+fg'), \tag{4.15}$$

$$\begin{aligned}\int_0^\infty f'(x)\mathbb{E}_x g(R_t)\,\mathrm{d}x \;&=\; \int_0^\infty f'(x)\,\mathrm{d}x \int_0^\infty g'(y)\mathbb{P}_x(R_t>y)\,\mathrm{d}y\\ &=\; \int_0^\infty f'(x)\,\mathrm{d}x \int_0^\infty g'(y)\bigl(1-\mathbb{P}_y(V_t\ge x)\bigr)\,\mathrm{d}y\\ &=\; -\int_0^\infty f'(x)\,\mathrm{d}x \int_0^\infty g'(y)\mathbb{P}_y(V_t\ge x)\,\mathrm{d}y \;=\; -\int_0^\infty g'(y)\mathbb{E}_y f(V_t)\,\mathrm{d}y.\end{aligned}$$

Inserting $\mathbb{E}_x g(R_t)=g(x)+t\mathscr{B}g(x)+\mathrm{o}(t)$, $\mathbb{E}_y f(V_t)=f(y)+t\mathscr{A}f(y)+\mathrm{o}(t)$, $t\downarrow 0$, and using (4.14), (4.15) yields

$$\begin{aligned}\int_0^\infty f'(x)\mathscr{B}g(x)\,\mathrm{d}x \;&=\; -\int_0^\infty g'(y)\mathscr{A}f(y)\,\mathrm{d}y\\ &=\; -\int_0^\infty g'(y)\Bigl[\frac{1}{2}a(y)f''(y)+b(y)f'(y)\Bigr]\,\mathrm{d}y\\ &=\; \int_0^\infty \Bigl\{-b(y)f'(y)g'(y)+\frac{1}{2}\bigl[a(y)g''(y)+a'(y)g'(y)\bigr]f'(y)\Bigr\}\,\mathrm{d}y\end{aligned}$$

(evaluating $\int (ag')f''$ by integration by parts in the last step and using $[ag'f']_0^\infty = 0-0$ as in (4.15)). This is the same as $\int f'\mathscr{B}^* g$, where $\mathscr{B}^* g$ denotes the r.h.s. of (4.13), and the truth of this for all $f\in\mathscr{K}$ implies (4.13). □

Problems

4.1 Assume that $\{V_n\}$ is a Markov chain on $\mathbb{N}$ with transition probabilities p_{ij}. Find the transition probabilities for $\{R_n\}$ [you should not expect a particularly simple answer!].

4.2 Derive (4.11), (4.12) via the recursive approach.

Notes We will see further important examples of duality in XIV.3 and XIV.5.

The Markov process approach of Theorem 4.2 is from Siegmund (1976a), and the theory is often referred to as *Siegmund duality*, whereas the recursive approach of Theorem 4.4 is from Asmussen and Sigman (1996). None of the approaches generalizes readily to higher dimension, as illustrated by Blaszczyszyn and Sigman (1999) in their study of many–server queues. For stochastic recursions in general, see Brandt *et al.* (1990) and Borovkov and Foss (1992).

The two–barrier formula (4.12) is implicit in Lindley (1959) and explicit in Siegmund (1976a), but has often been overlooked so that there are a number of alternative treatments of two–barrier models around. We treat such models in more detail in XIV.3.

When applying Siegmund duality when $\mathbb{T}=[0,\infty)$, it is often difficult to rigorously identify $\{R_t\}$, as illustrated by Example 4.7 which is somewhat at the heuristical level (but see Cox and Rössler, 1984). Asmussen (1995) gives a Markov–modulated generalization for $\mathbb{T}=[0,\infty)$, and there is some general theory for the recursive setting in Ryan and Sigman (2000).

Part C:
Special Models and Methods

X

Steady-State Properties of $GI/G/1$

1 Notation. The Actual Waiting Time

We consider the (FIFO) $GI/G/1$ queue in the notation of III.1b. That is, the customers are numbered $n = 0, 1, 2, \ldots$, U_n is the service time of n, T_n the time between the arrivals of n and $n+1$ and $A(x) = \mathbb{P}(T_n \le x)$ is the interarrival distribution, $B(x) = \mathbb{P}(U_n \le x)$ the service-time distribution (we assume $A(0) = \mathbb{P}(T_n = 0) = 0$, $B(0) = \mathbb{P}(U_n = 0) = 0$). We let $\mu_A = \mathbb{E}T_n$ denote the interarrival mean and $\mu_B = \mathbb{E}U_n$ the mean service time (μ_A, μ_B are assumed finite throughout). Then $\rho = \mu_B/\mu_A$ is the traffic intensity. Unless otherwise stated, *it is assumed that customer* 0 *has just arrived at time* $t = 0$ *to an empty queue.*

Some basic tools in the analysis of the system are: random walks that yield information on the waiting–time distribution; regenerative processes that permit conclusions to be made on the existence of limits of other functionals such as queue lengths; and rate conservation that will provide relations between the limits and in particular express the distributions of workload and queue size in terms of the waiting-time distribution.

Some of the basic facts on the waiting times have already been touched upon, but will now be put together. Define $X_n = U_n - T_n$, $\mu = \mathbb{E}X_n = \mu_B - \mu_A$, $S_0 = 0$, $S_n = X_0 + \cdots + X_{n-1}$, $M_n = \max_{0 \le k \le n} S_k$, $M = \max_{0 \le k < \infty} S_k$. Then the cases $\mu < 0$, $\mu = 0$ and $\mu > 0$ correspond to $\rho < 1$, $\rho = 1$, resp. $\rho > 1$, and III.6 yields:

Proposition 1.1 *The* (actual) *waiting time process* $\{W_n\}$ *is a Lindley process generated by* $\{S_n\}$*, i.e.* $W_{n+1} = (W_n + X_n)^+$*. In particular,*

$$W_n \;=\; \max\bigl(S_n, S_n - S_1, \ldots, S_n - S_{n-1}, 0\bigr) \tag{1.1}$$

$$\overset{\mathscr{D}}{=} \; M_n \tag{1.2}$$

and if $\rho < 1$*, then a limiting steady–state distribution exists and is given by* $\mathbb{P}_e(W_n \le x) = \mathbb{P}(M \le x)$*.*
[The formulas (1.1) and (1.2) require slight variants for $W_0 \neq 0$, cf. III.6. However, the limit result still holds true.]

Our interest in the following is centered around the so–called *stable case* $\rho < 1$ and we shall only briefly as a digression indicate the typical behaviour for $\rho \ge 1$.

Proposition 1.2 (i) *If* $\rho = 1$, $\sigma^2 = \mathbb{V}ar X_n < \infty$*, then the limiting distribution of* $W_n/\sqrt{n}$ *exists and is that of the absolute value of a normal r.v. with mean zero and variance* σ^2*;* (ii) *if* $\rho > 1$*, then* $W_n/n \overset{\text{a.s.}}{\to} \mu = \mu_A(\rho - 1)$*.*

Proof. In case (i), it is well known that $M_n/\sqrt{n}$ has the asserted limit properties (the easiest proof is presumably by Donsker's theorem, Billingsley, 1968, Ch. 2; for a direct proof, see Chung, 1974, pp. 217–222). In case (ii), we have $S_n/n \overset{\text{a.s.}}{\to} \mu > 0$. Hence by (1.1), $W_n > 0$ eventually. Hence if η is the last n with $W_n = 0$, we have $W_n = S_n - S_\eta$, $n \ge \eta$, from which we get $W_n/n \sim S_n/n \sim \mu$. □

Now define $\sigma(0) = 0$, $\sigma = \inf\{n \ge 1 : W_n = 0\}$, $\sigma(1) = \sigma$, $\sigma(k+1) = \inf\{n > \sigma(k) : W_n = 0\}$. Since $W_0 = 0$, we may interpret σ as the number of customers served in the first busy period and $\sigma(k)$ as the index of the customer initiating the kth busy cycle.

Proposition 1.3 *The* $\sigma(k)$ *are regeneration points for the waiting–time process. We have* $\mathbb{P}(\sigma < \infty) = 1$ *if and only if* $\rho \le 1$*. Hence for* $\rho \le 1$, $\{W_n\}$ *is aperiodic regenerative with imbedded renewal sequence* $\{\sigma(k)\}$*. Furthermore,* $\sigma = \sigma(1)$ *coincides with the weak descending ladder epoch,* $\sigma = \tau_- = \inf\{n \ge 1 : S_n \le 0\}$*. We have*

$$W_n \;=\; S_n \;=\; U_0 + \cdots + U_{n-1} - T_0 - \cdots - T_{n-1}, \quad n = 0, \ldots, \sigma - 1, \tag{1.3}$$

$$-S_\sigma = -S_{\tau_-} = I \tag{1.4}$$

where I *is the idle period corresponding to the first busy cycle, and furthermore* $\mathbb{E}\sigma < \infty$ *if and only if* $\rho < 1$*.*

Proof. By the Lindley process property, we have $W_n = S_n$, $n = 0, \ldots, \sigma - 1$, and this makes it clear that $\sigma = \tau_-$. Also I is the amount by which the last interarrival time exceeds the residual work at the time of the last arrival in the cycle,

$$I \;=\; T_{\sigma-1} - (W_{\sigma-1} + U_{\sigma-1}) \;=\; -S_\sigma \;=\; -S_{\tau_-}.$$

It is clear that the $\sigma(k)$ are regeneration points, and by general random walk results we have finally $\sigma = \tau_- < \infty$ a.s. if and only if $\mu = \mathbb{E}X_n \le 0$, i.e. $\rho \le 1$, and $\mathbb{E}\tau_- = \mathbb{E}S_{\tau_-}/\mu < \infty$ if and only if $\mu < 0$, i.e. $\rho < 1$. Finally aperiodicity follows from $\mathbb{P}(\sigma = 1) = \mathbb{P}(U \le T) > 0$. □

For the sake of easy reference, some of the main r.v.'s occurring in the rest of the chapter will now be introduced.

Definition 1.4 *Suppose $\rho < 1$. Then throughout this chapter:*
(i) *W will denote a random variable having the steady–state distribution H, say, of W_n, $H(x) = \mathbb{P}(W \le x) = \mathbb{P}_e(W_n \le x)$; similarly,*
(ii) *V, Q have the steady state distributions of the workload V_t, resp. the queue length Q_t* (which will be shown to exist if the interarrival distribution A is nonlattice)*;*
(iii) *Q_n^A, Q_n^D denote the queue length just prior to the nth arrival and just after the nth departure, and Q^A, Q^D the corresponding steady–state quantities;*
(iv) *U, T, X, $\{T^{(k)}\}_0^\infty$ have the distributions of U_n, T_n, $X_n = U_n - T_n$, $\{T_0 + \cdots + T_{k-1}\}_0^\infty$, respectively, and are mutually independent and independent of W, V, Q, Q^A, etc.; similar conventions apply for*
(v) *U^*, T^* having densities $\mathrm{d}B_0(x)/\mathrm{d}x = \overline{B}(x)/\mu_B$ and $\mathrm{d}A_0(x)/\mathrm{d}x = \overline{A}(x)/\mu_A$.*

The distributions B_0, A_0 are familiar from renewal theory, V.3. Also, from the independence of W_n and U_n, it is seen that we may identify $W + U$ by the sojourn time in the steady state.

A main problem for the study of the actual waiting time is obviously to study the distribution of W. Various expressions are available for $H(x) = \mathbb{P}(W \le x)$. From Proposition 1.1 and VIII.2.2 we have

$$H(x) \;=\; \big(1 - \|G_+\|\big)U_+(x) \;=\; \big(1 - \|G_+\|\big)\sum_{n=0}^{\infty} G_+^{*n}(x), \tag{1.5}$$

whereas Proposition 1.3 and VI.(1.5) yield

$$H(x) \;=\; \frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1} I(W_n \le x) \;=\; \frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x)\,. \tag{1.6}$$

These formulas are, however, not intrinsically different in view of VIII.2.3(b). A somewhat different characterization of H is as the unique solution to Lindley's integral equation III.(6.6) with

$$F(x) \;=\; \mathbb{P}(X_n \le x) \;=\; \mathbb{P}(U_n - T_n \le x) \;=\; \int_0^\infty B(x+y)\,A(\mathrm{d}y), \quad x \in \mathbb{R}.$$

Also, the characteristic function has been found in VIII.4 but is obviously quite complicated.

The representation (1.5) will turn out to be particularly useful when combined with VIII.3.1(b) stating that

$$G_+(A) \;=\; U_- * F(A) \;=\; \int_{-\infty}^{0} F(A-x)\,U_-(\mathrm{d}x), \quad A \subseteq (0,\infty). \tag{1.7}$$

Proposition 1.5 $W \overset{\mathscr{D}}{=} (W+X)^+$, *whereas the conditional distribution of* $(W+X)^-$ *given* $W+X \le 0$ *coincides with the common distribution of* $-S_{\tau_-}$ *and* I. *In particular, for* $f : [0,\infty) \to [0,\infty)$

$$\mathbb{E}f\big((W+X)^-\big) \;=\; \frac{\mathbb{E}f(-S_{\tau_-})}{\mathbb{E}\tau_-} \;=\; -\mathbb{E}X\frac{\mathbb{E}f(I)}{\mathbb{E}I}, \tag{1.8}$$

$$\mathbb{E}(W+X)^- \;=\; -\mathbb{E}X. \tag{1.9}$$

Proof. The first statement was noted previously in III.6.6 and yields in particular

$$\mathbb{P}(W+X \le 0) \;=\; \mathbb{P}\big((W+X)^+ = 0\big) = \mathbb{P}(W=0) \;=\; 1 - \|G_+\| \;=\; 1/\mathbb{E}\tau_-,$$

cf. VIII.2.3(c). Also, by VIII.3.2(b),

$$\begin{aligned}
&\mathbb{E}f\big((W+X)^-\big)\\
&= \int_{-\infty}^{0} f(-x)\,H * F(\mathrm{d}x) \;=\; \big(1 - \|G_+\|\big)\int_{-\infty}^{0} f(-x)\,U_+ * F(\mathrm{d}x)\\
&= \big(1 - \|G_+\|\big)\int_{-\infty}^{0} f(-x)\,G_-(\mathrm{d}x) \;=\; \frac{1}{\mathbb{E}\tau_-}\mathbb{E}f(-S_{\tau_-})\\
&= \mathbb{P}(W+X \le 0)\mathbb{E}f(-S_{\tau_-}).
\end{aligned}$$

Recalling $\mathbb{E}S_{\tau_-} = \mathbb{E}\tau_-\mathbb{E}X$ and (1.4), the proof is complete. □

Problems

1.1 Give a direct proof of (1.9) by using $W+X = (W+X)^+ - (W+X)^-$.

2 The Moments of the Waiting Time

The problem is to study conditions for the existence of $\mathbb{E}W^p$, $p>0$, and, as far as possible, to derive an explicit expression. In view of $W \overset{\mathscr{D}}{=} M$, this is really a random walk problem (as in the case for many other aspects of the behaviour of the waiting time, cf. e.g. Sections 6 and 7) and can therefore be formulated in that setting alone. The queueing interpretation may, however, require some slight reformulations: for example, in the following existence result, $\mathbb{E}(X^+)^{p+1} = \mathbb{E}\big((U-T)^+\big)^{p+1} < \infty$ is readily seen to be equivalent to $\mathbb{E}U^{p+1}$, whereas $\mathbb{E}X^- < \infty$ is automatic in view of $\mathbb{E}T = \mu_A < \infty$.

Theorem 2.1 *Consider a random walk with $\mu = \mathbb{E}X < 0$ and let $p > 0$. Then $\mathbb{E}M^p < \infty$ provided that $\mathbb{E}(X^+)^{p+1} < \infty$. Conversely, if $\mathbb{E}M^p < \infty$ and $\mathbb{E}\,X^- < \infty$, then $\mathbb{E}(X^+)^{p+1} < \infty$.*

Proof. We first note that the pth moment ν_n of a sum $Y_1 + \cdots + Y_n$ of nonnegative i.i.d. summands with $\mathbb{E}Y_1^p < \infty$ is $\mathrm{O}(n^p)$. Indeed, if $p \le 1$ Jensen's inequality gives $\nu_n \le (n\mathbb{E}Y)^p$, whereas for $p \ge 1$ we have

$$\nu_n^{1/p} \;=\; \bigl[\mathbb{E}(Y_1 + \cdots + Y_n)^p\bigr]^{1/p} \;=\; \|Y_1 + \cdots + Y_n\|_p \;\le\; n\|Y\|_p.$$

Hence if $\alpha = \mathbb{E}\bigl[S_{\tau_+}^p;\, \tau_+ < \infty\bigr] < \infty$,

$$\mathbb{E}M^p \;=\; \bigl(1 - \|G_+\|\bigr)\sum_{n=0}^{\infty}\int_0^\infty x^p G_+^{*n}(\mathrm{d}x) \;=\; \bigl(1 - \|G_+\|\bigr)\sum_{n=0}^{\infty}\|G_+\|^n \mathrm{O}(n^p)$$

will be finite in view of $\|G_+\| < 1$, whereas if $\alpha = \infty$ then the term corresponding to $n = 1$ in the sum is infinite and hence $\mathbb{E}M^p = \infty$.

Write $U(y) = U_-[-y, 0]$. Then by VIII.3.1(b)

$$\begin{aligned}\frac{\alpha}{p} \;&=\; \int_0^\infty x^{p-1}\mathbb{P}\bigl(S_{\tau_+} > x,\, \tau_+ < \infty\bigr)\,\mathrm{d}x \;=\; \int_0^\infty x^{p-1} U_- * F(x, \infty)\,\mathrm{d}x \\ &=\; \int_0^\infty F(\mathrm{d}y)\int_0^y x^{p-1}U(y-x)\,\mathrm{d}x\,. \end{aligned} \tag{2.1}$$

By the elementary renewal theorem (the proof is valid also if the interarrival distribution has an atom at 0 as G_-) we have for suitable c_1, c_2 that $U(z) \le c_1 + c_2 z$, and since for large y

$$\begin{aligned}&\int_0^y x^{p-1}\bigl[c_1 + c_2(y-x)\bigr]\,\mathrm{d}x \\ &\quad=\; \frac{1}{p}y^p c_1 + \frac{1}{p}y^{p+1}c_2 - \frac{1}{p+1}y^{p+1}c_2 \;\sim\; \frac{c_2}{p(p+1)}y^{p+1}\end{aligned} \tag{2.2}$$

it follows that $\mathbb{E}(X^+)^{p+1} < \infty$ implies $\alpha < \infty$ and hence $\mathbb{E}M^p < \infty$. Conversely, if $\mathbb{E}X^- < \infty$, then $\mathbb{E}S_{\tau_-} = \mathbb{E}\tau_-\mathbb{E}X > -\infty$ and hence $U(z) \ge d_1 + d_2 z$ with $d_2 > 0$. If $\mathbb{E}M^p < \infty$, then $\alpha < \infty$ and combining (2.1) and (2.2) yields $\mathbb{E}(X^+)^{p+1} < \infty$. □

Not even the moments of M (if they exist) can be found very explicitly. For example, VIII.4.5 and (1.5) yield the expressions

$$\mathbb{E}M \;=\; \sum_{n=1}^{\infty}\frac{1}{n}\mathbb{E}S_n^+ \;=\; \frac{\mathbb{E}[S_{\tau_+};\, \tau_+ < \infty]}{1 - \|G_+\|}\,. \tag{2.3}$$

A further important relation is the following:

Theorem 2.2 *If $\mathbb{E}|X|^{p+1} < \infty$ for some $p = 1, 2, \ldots$, then*

$$\sum_{q=0}^{p}\binom{p+1}{q}\mathbb{E}M^q\mathbb{E}X^{p+1-q} \;=\; \mathbb{E}\bigl[-(M+X)^-\bigr]^{p+1} \;=\; \frac{\mathbb{E}S_{\tau_-}^{p+1}}{\mathbb{E}\tau_-}\,. \tag{2.4}$$

(Note that in the queueing setting, we may rewrite the r.h.s. of (2.4) as $(-1)^p\mathbb{E}X\mathbb{E}I^{p+1}/\mathbb{E}I$; cf. (1.8).)

Proof. The last identity in (2.4) follows from (1.8). To show the remaining part of the theorem, first suppose $\mathbb{E}(X^+)^{p+2} < \infty$. Then

$$\mathbb{E}M^{p+1} < \infty, \quad \mathbb{E}\big[(M+X)^-\big]^{p+1} \le \mathbb{E}|X|^{p+1} < \infty,$$

and since $(M+X)^+(M+X)^- = 0$ we get

$$\begin{aligned}
(M+X)^{p+1} &= \big[(M+X)^+ - (M+X)^-\big]^{p+1} \\
&= \big[(M+X)^+\big]^{p+1} + \big[-(M+X)^-\big]^{p+1}, \\
\mathbb{E}(M+X)^{p+1} &= \sum_{q=0}^{p+1}\binom{p+1}{q}\mathbb{E}M^q\mathbb{E}X^{p+1-q} \\
&= \mathbb{E}\big[(M+X)^+\big]^{p+1} + \mathbb{E}\big[-(M+X)^-\big]^{p+1} \\
&= \mathbb{E}M^{p+1} + \mathbb{E}[-(M+X)^-]^{p+1}
\end{aligned}$$

and cancelling $\mathbb{E}M^{p+1}$, (2.4) follows. In the general case, replace X_n by $X_n^{(k)} = X_n \wedge k$ and let $M^{(k)}$ be defined in terms of the $X_n^{(k)}$ rather than the X_n. Then $\mathbb{E}(X^{(k)+})^{p+2} < \infty$, hence

$$\sum_{q=0}^{p}\binom{p+1}{q}\mathbb{E}M^{(k)q}\mathbb{E}X^{(k)p+1-q} \;=\; \mathbb{E}[-(M^{(k)}+X^{(k)})^-]^{p+1} \;<\; \infty\,.$$

But clearly, $M^{(k)} \le M$ and $M^{(k)} \uparrow M$ as $k \to \infty$. Hence the desired conclusion follows by monotone convergence as $k \to \infty$. □

Rewriting in queueing notation, we get in particular for the mean waiting time ($p = 1$) that

$$\begin{aligned}
2\mathbb{E}(-X)\mathbb{E}W &= \mathbb{E}X^2 - \mathbb{E}\big[(W+X)^-\big]^2 \;=\; \mathbb{V}ar X - \mathbb{V}ar(W+X)^- \\
&= \mathbb{E}X^2 - \frac{\mathbb{E}S_{\tau_-}^2}{\mathbb{E}\tau_-} \;=\; \mathbb{E}X^2 - \frac{\mathbb{E}(-X)\mathbb{E}I^2}{\mathbb{E}I} \qquad (2.5)
\end{aligned}$$

(here the second equality follows from (1.9)). Considerable effort has been put into converting these expressions into bounds or approximations that are more explicit in the sense that only the distribution of X (or U, T) is invoked, and preferably only even the first few moments. We return to the approximations in Section 7 and XIII.6, and here present only some of the roughest bounds,

$$\mathbb{E}U^2 - \mathbb{E}U\mathbb{E}T \;\le\; 2\mathbb{E}(-X)\mathbb{E}W \;\le\; \mathbb{V}ar\,X \;=\; \mathbb{V}ar\,U + \mathbb{V}ar\,T\,. \qquad (2.6)$$

(The lower bound may be negative and hence trivial. The upper bound is in fact sharp in an asymptotic sense; cf. Section 7.) Here the upper bound is obvious from $\mathbb{V}ar(W+X)^- \ge 0$. For the lower bound, rewrite (2.5) as

$$\mathbb{E}U^2 - 2\mathbb{E}U\mathbb{E}T + \mathbb{E}T^2 - \mathbb{E}[(W+X)^-]^2 \;=\; \mathbb{E}U^2 - 2\mathbb{E}U\mathbb{E}T + \mathbb{E}(CD)$$

where $C = T + (W+X)^-$, $D = T - (W+X)^-$. Here

$$D \;=\; T + W + X - (W+X)^+ \;=\; W + U - (W+X)^+$$

so that

$$\mathbb{E}(CD) \;=\; \mathbb{E}\bigl[T\bigl(W - (W+X)^+\bigr)\bigr] + \mathbb{E}T\mathbb{E}U + \mathbb{E}\bigl[(W+X)^-(W+U)\bigr] \quad (2.7)$$

The two last terms in (2.7) are obviously nonnegative, and thus it is sufficient to show that the first one is so too. But $f(T) = T$ and $g(T) = W - (W+U-T)^+$ are both nondecreasing in T for fixed W, U. Hence by a well–known inequality (Problem 2.2)

$$\begin{aligned}\mathbb{E}\bigl[f(T)g(T)\,\big|\,W,U\bigr] \;&\geq\; \mathbb{E}\bigl[f(T)\,\big|\,W,U\bigr]\cdot\mathbb{E}\bigl[g(T)\,\big|\,W,U\bigr]\\ &=\; \mathbb{E}T\cdot\mathbb{E}\bigl[W-(W+X)^+\,\big|\,W,U\bigr],\\ \mathbb{E}\bigl[T\bigl(W-(W+X)^+\bigr)\bigr] \;&\geq\; \mathbb{E}T\cdot\mathbb{E}\bigl[W-(W+X)^+\bigr] \;=\; \mathbb{E}T\cdot 0 \;=\; 0.\end{aligned}$$

□

Problems

2.1 Consider a random walk with $\|G_+\| = \|G_-\| = 1$. Show that $\mathbb{E}S_{\tau_+} < \infty$, $\mathbb{E}S_{\tau_-} > -\infty$ if and only if $\mathbb{E}X^2 < \infty$, $\mathbb{E}X = 0$, and that then $\mathbb{E}X^2 = -2\mathbb{E}S_{\tau_+}\mathbb{E}S_{\tau_-}$. [*Hint:* Necessity and the stated identity follows by Wiener–Hopf factorization of the ch.f.]

2.2 (CHEBYCHEFF'S COVARIANCE INEQUALITY) Let X be a r.v. and f, g nondecreasing functions. Show that $\mathbb{E}[f(X)g(X)] \geq \mathbb{E}f(X)\mathbb{E}g(X)$ [*Hint:* Reduce to the case $\mathbb{E}f(X) = 0$ and consider $\mathbb{E}[f(X)(g(X) - g(b))]$ where b is the point at which f changes sign.]

2.3 Carry out the last step in the proof of Corollary IX.3.4.

Notes Theorem 2.1 goes back to Kiefer & Wolfowitz (1956) and there are many proofs around. As one of many applications of (2.5), we mention in particular the observation by Minh and Sorli (1983) that when estimating $\mathbb{E}W$ by simulation, the only unknown quantities are $\mathbb{E}I$ and $\mathbb{E}I^2$, and that simulating these rather than $\mathbb{E}W$ increases precision. Bounds for $\mathbb{E}W$ and related quantities are surveyed in Stoyan (1983) and Daley *et al.* (1994). For Problem 2.2, see also Thorisson (2000), p. 2.

3 The Workload

In continuous time, there is a regenerative structure similar to the one in Proposition 1.3: the instants with a customer entering an empty queue are regeneration points. Letting C be the first such instant after $t = 0$ and recalling that we start with customer 0 having just arrived, it is seen that C is just the length of the first busy cycle. Furthermore, $C < \infty$ a.s. is equivalent to $\sigma < \infty$ a.s., i.e. to $\rho \leq 1$ (cf. Proposition 1.3). In fact, there is a close relation between σ, C and the first busy period G: since precisely the customers $0, 1, \ldots, \sigma - 1$ are served in the first busy period, we have $G = U_0 + \cdots + U_{\sigma-1}$ and the first busy cycle ends at the arrival time $C = T_0 + \cdots + T_{\sigma-1}$ of customer σ. One checks immediately that $\{\sigma \le n\}$ is independent of $T_n, T_{n+1}, \ldots, U_n, U_{n+1}$, and hence Wald's identity yields the first part of

Proposition 3.1 *Suppose $\rho \le 1$. Then the mean busy cycle is $\mathbb{E}C = \mu_B \mathbb{E}\sigma$, the mean busy period is $\mathbb{E}G = \mu_A \mathbb{E}\sigma$ and the mean idle period is $\mathbb{E}I = \mathbb{E}C - \mathbb{E}G = -\mu\mathbb{E}\sigma$. Furthermore the mean busy period is nonlattice if and only if the interarrival distribution A is so, and spread out if and only if A is so.*

The second part is often stated to be obvious, but some care is needed (cf. Problem 3.3), and we give the proof (when $\rho < 1$) in the form of the following more general result:

Proposition 3.2 *Let $T_0 > 0, T_1 > 0, \ldots$ be i.i.d. with common distribution A with $\mu_A < \infty$, and let $\sigma \ge 1$ be a random time such that $\mathbb{E}\sigma < \infty$ and $T_n, T_{n+1}, \ldots$ are independent of $\{\sigma \le n\}$ for all n. Then then distribution K of $C = T_0 + \cdots + T_{\sigma-1}$ is nonlattice if and only if A is so, and spread out if and only if A is so.*

Proof. By Wald's identity, we have $\mathbb{E}C < \infty$. Also by an obvious iterative procedure we may assume that random times $\sigma(1) = \sigma < \sigma(2) < \cdots$ have been constructed such that $\left\{T_0 + \cdots + T_{\sigma(k)-1}\right\}$ is a renewal process governed by K. Then, in the obvious notation, the renewal measures satisfy $U_A \ge U_K$. Suppose K was lattice, say aperiodic on $\mathbb{N}$, but A not. Then by Blackwell's renewal theorem,

$$\frac{h}{\mu_A} = \lim_{n\to\infty} \big[U_A(n) - U_A(n-h)\big] \ge \lim_{n\to\infty} \big[U_K(n) - U_K(n-h)\big] = \frac{1}{\mu_K}$$

for all $h < 1$, which is impossible. Similarly, assume that A is spread out but K not. Then U_K is concentrated on a Lebesgue null set N, and Stone's decomposition shows that the U_A–measure of N is finite, whereas the U_K–measure is infinite, contradicting $U_A \ge U_K$.

If, conversely, A is not spread out, then U_A is concentrated on a Lebesgue null set N. Hence U_K is concentrated on N, and K cannot be spread out. That K is lattice if A is so is even more trivial. □

The remaining part of Proposition 3.1 now follows immediately when $\rho < 1$. When $\rho = 1$, replace B by an equivalent (in the sense of null sets) and stochastically smaller distribution $\widetilde{B}$. Then the busy cycle distributions are equivalent, and since $\widetilde{\rho} < 1$, Proposition 3.2 applies to $\widetilde{C}$.

Corollary 3.3 *Suppose $\rho < 1$ and that A is nonlattice. Then a limiting steady–state distribution of the workload V_t exists and is given by*

$$\mathbb{E}f(V) \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C f(V_s)\,\mathrm{d}s. \tag{3.1}$$

If A is spread out, then $V_t \to V$ in total variation.

Proof. For $\rho < 1$, we have $\mathbb{E}\sigma < \infty$. Hence Proposition 3.1 ensures that the basic limit theorems for regenerative processes in VI.1 and VII.1 are applicable. □

As a first application of (3.1), note that the time spent by $\{V_t\}$ in state 0 in the time interval $[0, C)$ is the just the idle period. Thus combining with Proposition 3.1, we get

$$\begin{aligned}\mathbb{P}(V = 0) \;&=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C I(V_s = 0)\,\mathrm{d}s \\ &=\; \frac{\mathbb{E}I}{\mathbb{E}C} \;=\; \frac{(\mu_A - \mu_B)\mathbb{E}\sigma}{\mu_A\mathbb{E}\sigma} \;=\; 1 - \rho\,. \end{aligned}\tag{3.2}$$

[Note that this is always explicit in contrast to $\mathbb{P}(W = 0) = 1/\mathbb{E}\sigma$.]

We next express the distribution of V in terms of the steady–state waiting time distribution (for the meaning of U^*, T^*, see Definition 1.4):

Theorem 3.4 *The conditional distribution of V given $V > 0$ is the same as the distribution $H * B_0$ of $W + U^*$. Equivalently,*

$$\mathbb{P}(V \le x) \;=\; 1 - \rho + \rho\mathbb{P}(W + U^* \le x) \;=\; 1 - \rho + \rho H * B_0(x), \tag{3.3}$$

cf. (3.2). *An alternative characterization is $V \overset{\mathscr{D}}{=} (W + U - T^*)^+$.*

Proof. Let $X_t = (V_t - x)^+$. Then $\{X_t\}$ has derivative -1 when $V_t > x$ and 0 otherwise, whereas the jump at the arrival of customer n is $(W_n + U_n - x)^+ - (W_n - x)^+$. Hence the rate conservation law VII.6.6 applied to a stationary version yields

$$\mathbb{P}(V > x) \;=\; \frac{1}{\mu_A}\big[\mathbb{E}(W+U-x)^+ - \mathbb{E}(W-x)^+\big] \;=\; \frac{\mu_B}{\mu_A}\mathbb{P}\big(U^* > (W-x)^-\big)$$

where the last identity follows from

$$\mathbb{E}\big[(U+a)^+ - a^+\big] \;=\; \int_{a-}^{\infty}\overline{B}(u)\,\mathrm{d}u \;=\; \mu_B\mathbb{P}(U^* > a^-) \tag{3.4}$$

(integration by parts) by conditioning upon $a = W - x$. Hence $\mathbb{P}(V > x) = \rho\mathbb{P}\big(U^* > (x - W)^+\big)$ which (since $U^* > 0$) is the same as $\mathbb{P}(V > x) = \rho\mathbb{P}(U^* > x - W)$, $x \ge 0$, and (3.3) follows.

For $V \overset{\mathscr{D}}{=} (W + U - T^*)^+$, consider $X_t = \int_t^{M_t} I(V_s > x)\,\mathrm{d}x$ where M_t is the next arrival instant after t. The process $\{X_t\}$ decreases linearly at unit rate on intervals where $V_s > x$, is 0 at the nth arrival instant and then jumps to $\int_0^{T_n} I(W_n + U_n - s > x)\,\mathrm{d}s$. It follows by rate conservation that for $x \geq 0$

$$\begin{aligned}\mathbb{P}(V > x) &= \frac{1}{\mu_A}\int_0^\infty \overline{A}(s) I(W+U-s > x)\,\mathrm{d}s \;=\; \mathbb{P}(T^* < W + U - x) \\ &= \mathbb{P}(W + U - T^* > x) \;=\; \mathbb{P}\bigl((W+U-T^*)^+ > x\bigr).\end{aligned}$$

□

Note that in the $M/G/1$ case, we have $T^* \stackrel{\mathscr{D}}{=} T$ and hence

$$V \;\stackrel{\mathscr{D}}{=}\; (W+U-T^*)^+ \;\stackrel{\mathscr{D}}{=}\; (W+U-T)^+ \;\stackrel{\mathscr{D}}{=}\; W$$

so that we obtain another proof that $V \stackrel{\mathscr{D}}{=} W$ in $M/G/1$, as found already in III.9 and VII.6.

Since $\mathbb{E}U^* = \mathbb{E}U^2/2\mu_B$, it follows also by combining with (3.2) that:

Corollary 3.5 $\mathbb{E}V \;=\; \rho\Bigl\{\dfrac{\mathbb{E}U^2}{2\mu_B} + \mathbb{E}W\Bigr\}.$

It is instructive to consider the following two direct proofs of Corollary 3.5. The first uses rate conservation applied to $X_t = V_t^2$. Here in steady state, $\mathbb{E}X_t' = \mathbb{E}[2V;\, V > 0] = 2\mathbb{E}V$, so by rate conservation $\mathbb{E}V$ is

$$\frac{1}{2\mu_A}\mathbb{E}\bigl[(W+U)^2 - W^2\bigr] \;=\; \frac{1}{2\mu_A}\bigl[\mathbb{E}U^2 + 2\mu_B\mathbb{E}W\bigr] \;=\; \rho\Bigl\{\frac{\mathbb{E}U^2}{2\mu_B} + \mathbb{E}W\Bigr\}$$

(if $\mathbb{E}U^3 = \infty$ so that $\mathbb{E}W^2 = \infty$, use a truncation argument as in the proof of Theorem 2.2). The second proof uses a sample part decomposition of a regenerative cycle, cf. the partitioning of the subgraph of $\{V_t\}_{0 \leq t < C}$ into triangles and parallelograms in Fig. 3.1.

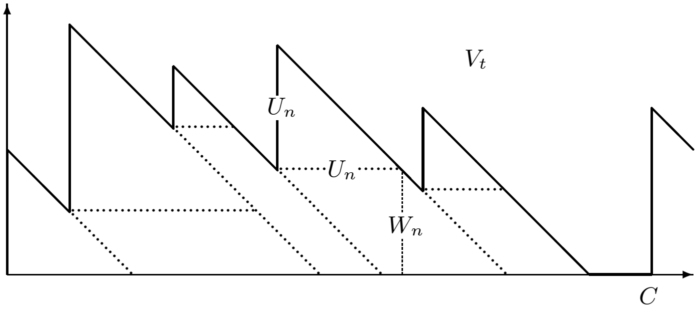

Figure 3.1

The area is $U_n^2/2$ of the nth triangle and W_nU_n of the nth parallelogram, hence

$$\mathbb{E}V \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C V_s\,\mathrm{d}s \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\sigma-1}\bigl[U_n^2/2 + W_nU_n\bigr]$$

$$
\begin{aligned}
&= \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\infty}\mathbb{E}\bigl[U_n^2/2+W_nU_n;\ \sigma>n\,\big|\,U_k,T_k,k=0,\ldots,n-1\bigr]\\
&= \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\infty}\mathbb{E}\bigl[U^2/2+W_nU;\ \sigma>n\bigr]\\
&= \frac{1}{\mu_A\mathbb{E}\sigma}\Bigl\{\frac{1}{2}\mathbb{E}U^2\mathbb{E}\sigma+\mu_B\mathbb{E}\sum_{n=0}^{\sigma-1}W_n\Bigr\}\\
&= \rho\Bigl\{\frac{\mathbb{E}U^2}{2\mu_B}+\frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1}W_n\Bigr\} \;=\; \rho\Bigl\{\frac{\mathbb{E}U^2}{2\mu_B}+\mathbb{E}W\Bigr\}.
\end{aligned}
$$

Problems

3.1 Define R_t as the residual service time of the customer being served at time t ($R_t = 0$ if the server is idle). Show that $\mathbb{P}(R \le x) = 1-\rho+\rho B_0(x)$.
3.2 Let $\{B_t\}$ be the forward recurrence time of the arrival process. Show that $\{(B_t,V_t)\}$ is strong Markov.
3.3 Show that the assumption $\mathbb{E}\sigma < \infty$ of Proposition 3.2 is indispensable. [*Hint:* $T_n = 1+\theta Z_n$ where $\theta \in (0,1)$ is irrational, $Z_n = \pm 1$ w.p. 1/2 and $\sigma = \inf\{n \ge 1 : Z_0+\cdots+Z_{n-1}=0\}$.]

Notes Since rate conservation holds in a more general stationary setting, it is clear that many results of the present section have parallels in such situations too, sometimes at a cost of a slightly more complicated formulation. See e.g. Sigman (1995).

4 Queue Length Processes

In the same way that the (actual) waiting time process is obtained by observing the virtual waiting time (workload) just before arrival times, it is sometimes of interest to look at the queue length (number in system) at certain random times. In particular, seen from the point of view of the arriving customer, the queue length at the time of arrival is a basic quantity and motivates the study of $\{Q_n^A\}_{n\in\mathbb{N}}$; cf. Definition 1.4. To distinguish from $\{Q_n^A\}$, $\{Q_n^D\}$, we use the terminology "at an arbitrary point of time" when considering $\{Q_t\}_{t\ge 0}$ in the steady state, and Q in Definition 1.4 refers to this case (i.e. $\mathbb{P}(Q=k)=\lim_{t\to\infty}\mathbb{P}(Q_t=k)$). We start by an elementary but celebrated result:

Theorem 4.1 (LITTLE'S LAW) *Suppose $\rho < 1$ and that A is nonlattice. Then the arrival rate $\lambda = \mu_A^{-1}$, the mean steady–state queue length $\ell = \mathbb{E}Q$ at an arbitrary point of time and the mean steady–state sojourn time $w = \mathbb{E}(W+U)$ are related by $\ell = \lambda w$.*

Proof. By reference to Proposition 3.1, regenerative processes apply to $\{Q_t\}_{t\geq 0}$ exactly as to the workload to show that a limiting steady–state r.v. Q exists (the convergence is always in t.v. since the state space $\mathbb{N}$ is discrete) and has distribution given by

$$\mathbb{E}f(Q) \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C f(Q_s)\,\mathrm{d}s. \tag{4.1}$$

Letting $f(x) = x$, it is seen that each of the customers $n = 0, 1, \ldots, \sigma - 1$ provide a contribution $W_n + U_n$ to $\int_0^C Q_s\,\mathrm{d}s$. Hence

$$\ell \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C Q_s\,\mathrm{d}s \;=\; \frac{1}{\mu_A\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1}(W_n+U_n) \;=\; \lambda\mathbb{E}(W+U) \;=\; \lambda w.$$

□

Theorem 4.2 (DISTRIBUTIONAL LITTLE'S LAW) *Let $\{N^*(t)\}$ be a time–stationary version of the renewal arrival process that is independent of W, U, etc. Then $Q \overset{\mathscr{D}}{=} N^*(W+U)$, i.e. $\mathbb{P}(Q=0)=\rho$ and*

$$\mathbb{P}(Q\geq k) \;=\; \mathbb{P}(N^*(W+U)\geq k) \;=\; \mathbb{P}\bigl(W+U > T^* + T^{(k-1)}\bigr)$$

for $k = 1, 2, \ldots$. An alternative characterization is

$$\begin{aligned}\mathbb{P}(Q\geq k) \;&=\; \mathbb{P}\bigl(V > T^{(k-1)}\bigr) \;=\; \rho\mathbb{P}\bigl(W+U^* > T^{(k-1)}\bigr), \qquad (4.2)\\ \mathbb{P}(Q=k) \;&=\; \int_0^\infty \bigl[A^{*(k-1)}(t) - A^{*k}(t)\bigr]\,H*B_0(\mathrm{d}t).\end{aligned}$$

Proof. Let $\tau_n = T_0 + \cdots + T_{n-1}$ be the arrival time of customer n. Then his service interval is $[\tau_n + W_n, \tau_n + W_n + U_n)$, and the time during this interval where $Q_t \geq k$ is the intersection with $[\tau_{n+k-1}, \infty)$ which has length

$$r_n \;=\; (\tau_n + W_n + U_n - \tau_{n+k-1})^+ - (\tau_n + W_n - \tau_{n+k-1})^+.$$

Since $\tau_{n+k-1} - \tau_n$ is independent of $W_n, U_n, \{\sigma > n\}$ and distributed as $T^{(k-1)}$, it follows that

$$\begin{aligned}\mathbb{P}(Q\geq k) \;&=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C I(Q_s\geq k)\,\mathrm{d}s\\ &=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\int_0^C I(Q_s\geq k)\,\mathrm{d}s \;=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\sum_{n=0}^{\sigma-1} r_n\\ &=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\sum_{n=0}^{\sigma-1}\Bigl[(W_n+U_n-T^{(k-1)})^+ - (W_n - T^{(k-1)})^+\Bigr].\end{aligned}$$

But conditionally upon $T^{(k-1)} = x$, this last expression is of the same form as (3.4) and so becomes

$$\begin{aligned}\rho\mathbb{P}\bigl(U^* > T^{(k-1)} - W\bigr) \;&=\; \rho\mathbb{P}\bigl(V > T^{(k-1)}\,\big|\,V>0\bigr) \;=\; \mathbb{P}\bigl(V > T^{(k-1)}\bigr) \qquad (4.3)\\ &=\; \mathbb{P}\bigl(W+U-T^* > T^{(k-1)}\bigr) \;=\; \mathbb{P}\bigl(N^*(W+U)\geq k\bigr), \qquad (4.4)\end{aligned}$$

where we used the first part of Theorem 3.4 for (4.3) and the last for (4.4). The first part of the present theorem now follows from (4.4) and the last from (4.3). □

Theorem 4.3 *Suppose $\rho < 1$. Then Q^A, Q^D are welldefined and have the same distribution given by*

$$\mathbb{P}(Q^A \geq k) \;=\; \mathbb{P}(Q^D \geq k) \;=\; \mathbb{P}\bigl(W + U \geq T^{(k)}\bigr). \tag{4.5}$$

If either A or B is continuous, this may be rewritten as

$$\begin{aligned}
\mathbb{P}(Q^A = 0) &\;=\; \mathbb{P}(Q^D = 0) \;=\; \mathbb{P}(W = 0) \;=\; H(0), && (4.6)\\
\mathbb{P}(Q^A \geq k) &\;=\; \mathbb{P}(Q^D \geq k) \;=\; \mathbb{P}\bigl(W > T^{(k-1)}\bigr), \quad k = 1, 2, \ldots, && (4.7)\\
\mathbb{P}(Q^A = k) &\;=\; \mathbb{P}(Q^D = k) \;=\; \mathbb{P}\bigl(T^{(k-1)} < W < T^{(k)}\bigr), && (4.8)\\
\mathbb{P}(Q^A = k) &\;=\; \mathbb{P}(Q^D{=}k) \;=\; \int_{0+}^{\infty} \bigl[A^{*(k-1)}(t) - A^{*k}(t)\bigr]\, H(\mathrm{d}t). && (4.9)
\end{aligned}$$

Proof. Clearly, $\{Q_n^A\}$, $\{Q_n^D\}$ are regenerative w.r.t. the renewal sequence $\{\sigma(k)\}$ and hence the existence of the limiting distribution is immediate. That the distributions are equal follow by rate conservation applied to $X_t = I(Q_t \geq k+1)$, since upward jumps occur at arrival epochs with $Q_n^A = k$ and downward at departure epochs with $Q_n^D = k$.

As in the proof of Theorem 4.2, we have

$$\bigl\{Q_{n+k}^A \geq k\bigr\} \;=\; \bigl\{\tau_{n+k} \leq \tau_n + W_n + U_n\bigr\}.$$

From this (4.5) follows by taking probabilities and letting $n \to \infty$. If either A or B is continuous, then so is the distribution of $W + U - T$ so that for $k \geq 1$ (4.5) becomes

$$\mathbb{P}\bigl(W + U > T^{(k)}\bigr) \;=\; \mathbb{P}\bigl(W + U - T > T^{(k-1)}\bigr) \;=\; \mathbb{P}\bigl(W > T^{(k-1)}\bigr)$$

(using $W \overset{\mathscr{D}}{=} (W + U - T)^+$). From this (4.6)–(4.9) follow by easy manipulations (in (4.9), 0 must be excluded from the domain of integration to deal with the case $k = 1$ where $T^{(k-1)} = 0$). □

Problems

4.1 Consider the set–up of Theorem 4.3. Explain that if $\mathbb{P}(U = T) > 0$, it may happen that $\mathbb{P}(Q^A = 0)$ is effectively smaller than $\mathbb{P}(W = 0)$.
4.2 Derive the distribution of Q^D by a direct argument similar to the one used for Q^A.

Notes It is clear as in Section 3 that much of the analysis carries over beyond the independence assumptions in $GI/G/1$. In particular, this is the case for Little's law which basically does not require anything more than the existence of limits of the Cesaro averages of the number of customers in continuous time and of the sojourn times in discrete time. The literature is extensive; see e.g.

Sigman (1995) and El–Taha and Stidham (1999). A series of papers by Glynn and Whitt, e.g. Glynn and Whitt (1986, 1989), deal with broader interpretations of $\ell = \lambda w$. For an extension $H = \lambda G$ that has been studied extensively, see e.g. Sigman (1995); it is essentially equivalent to the rate conservation law and applies typically to the same kind of problems.

5 $M/G/1$ and $GI/M/1$

Most of the steady–state characteristics of $M/G/1$ and $GI/M/1$ have already been found at various places; see in particular Sections 3, 4 and VIII.5. We collect here some of the main facts, give some complements and sketch some alternative approaches.

Theorem 5.1 *Consider the $GI/M/1$ queue with interarrival distribution A, service intensity δ and $\rho = (\delta\mu_A)^{-1} < 1$. Then in the steady state:*
(a) *The distribution of the waiting time W is a mixture of an atom at 0 and an exponential distribution with intensity η on $(0,\infty)$ with weights $1-\theta$, resp. θ. Here $\theta = \mathbb{E}\mathrm{e}^{-\eta T} = 1 - \eta/\delta$, where η is the solution > 0 of*

$$1 \;=\; \mathbb{E}\mathrm{e}^{\eta(U_n - T_n)} \;=\; \frac{\delta}{\delta - \eta}\int_0^\infty \mathrm{e}^{-\eta x}\, A(\mathrm{d}x)\,. \tag{5.1}$$

(b) *The distribution of the workload V is a mixture of an atom at 0 and an exponential distribution with intensity η on $(0,\infty)$ with weights $1-\rho$, resp. ρ.*
(c) *The distribution of the queue length Q at an arbitrary point of time is modified geometric and given by $\mathbb{P}(Q = 0) = 1 - \rho$, $\mathbb{P}(Q \geq k) = \rho\theta^{k-1}$, $k = 1, 2, \ldots$.*
(d) *The common distribution of the queue lengths Q^A, Q^D just before arrivals, resp. just after departures, is geometric with parameter θ, i.e. with point probabilities $\pi_n = (1-\theta)\theta^n$.*

Proof. (a) was shown in VIII.5.8. When U is exponential, we have $U^* \overset{\mathscr{D}}{=} U$, and we get the Laplace transform of $W + U^*$ as

$$\Bigl[1 - \theta + \theta\frac{\eta}{\eta + s}\Bigr]\frac{\delta}{\delta + s} \;=\; \frac{[\eta + (1-\theta)s]\delta}{(\eta + s)(\delta + s)} \;=\; \frac{\eta\delta + \eta s}{(\eta + s)(\delta + s)} \;=\; \frac{\eta}{\eta + s}$$

which proves (b); cf. Theorem 3.4. For (c) and (d), note first that conditioning upon $T^{(k-1)}$ in (4.7) yields

$$\mathbb{P}\bigl(W > T^{(k-1)}\bigr) \;=\; \theta\mathbb{E}\mathrm{e}^{-\eta T^{(k-1)}} \;=\; \theta\,\bigl[\mathbb{E}\mathrm{e}^{-\eta T}\bigr]^{k-1} \;=\; \theta^k,$$

and (d) follows. (c) is obtained similarly from (4.2) and $\mathbb{P}\bigl(V > T^{(k-1)}\bigr) = \rho\theta^{k-1}$. □

Imbedded Markov chain analysis plays an important historical role in the proof of results like (c) and (d) and is also applicable to a number of further models. We therefore next present the main steps of this approach, though it is certainly neither the shortest nor the most elegant one for simple queues such as $GI/M/1$ (and $M/G/1$ below). It was found in III.6.2 that the Markov chain $\{Q_n^A\}$ has transition matrix

$$\boldsymbol{P} \;=\; \begin{pmatrix} r_0 & q_0 & 0 & 0 & \dots \\ r_1 & q_1 & q_0 & 0 & \\ r_2 & q_2 & q_1 & q_0 & \\ \vdots & & & & \ddots \end{pmatrix},$$

where $q_k = \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^k}{k!}A(\mathrm{d}t)$ and $r_n = q_{n+1}+q_{n+2}+\cdots$. By direct insertion it is now seen that $\pi_n = (1-\theta)\theta^n$ solves $\boldsymbol{\pi P} = \boldsymbol{\pi}$, provided that θ satisfies (i) $\sum_0^\infty r_n\theta^n = 1$, (ii) $\sum_0^\infty q_n\theta^n = \theta$. An elementary calculation shows that (i) follows from (ii). If η, θ are connected by $\eta = \delta(1-\theta)$, $\theta = 1-\eta/\delta$, we may rewrite (ii) as

$$1-\frac{\eta}{\delta} \;=\; \sum_{n=0}^\infty q_n\theta^n \;=\; \int_0^\infty \mathrm{e}^{-\eta t}\,A(\mathrm{d}t),$$

which is the same as (5.1). Alternatively, $\boldsymbol{\pi}$ can be derived by remarking that $\{Q_n^A\}$ is a Lindley process governed by $f_1 = q_0$, $f_0 = q_1$, $f_{-1} = q_2, \ldots$, hence the stationary distributions is that of the random walk maximum M which was found in VIII.5.5(b).

To proceed from Q^A to Q, we use semi–regeneration; cf. VII.5. The cycle length is an interarrival time T and we let $\mathbb{E}_k$ refer to the case where k customers were present just before the start of the interarrival interval. The imbedded Markov chain in VII.5 is just $\{Q_n^A\}$ with stationary distribution $\boldsymbol{\pi}$, and thus by VII.(5.1), we have

$$\mathbb{P}(Q=j) \;=\; \frac{1}{m}\sum_{k=0}^\infty \pi_k\,\mathbb{E}_k\int_0^T I(Q_t=j)\,\mathrm{d}t,$$

where $m = \sum_0^\infty \pi_k\mathbb{E}_kT = \mu_A$. If $\{N_s\}$ is a Poisson process with intensity δ and $j \le k+1$, integration by parts yields

$$\begin{aligned}\mathbb{E}_k\int_0^T I(Q_t=j)\,\mathrm{d}t \;&=\; \int_0^\infty \mathbb{P}(N_t = k+1-j)\overline{A}(t)\,\mathrm{d}t \\ =\; \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^{k+1-j}}{(k+1-j)!}\overline{A}(t)\,\mathrm{d}t \;&=\; \int_0^\infty \delta^{-1}\sum_{\ell=k+2-j}^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^\ell}{\ell!}A(\mathrm{d}t)\end{aligned}$$

which equals $\delta^{-1}r_{k+1-j}$. For $j > k+1$, we get 0 and hence

$$\begin{aligned}\mathbb{P}(Q=j) \;&=\; \frac{1}{\mu_A}\sum_{k=j-1}^\infty \pi_k\delta^{-1}r_{k+1-j} \;=\; \rho\sum_{i=0}^\infty (1-\theta)\theta^{j-1}\theta^i r_i \\ &=\; \rho(1-\theta)\theta^{j-1}\end{aligned}$$

(using (i) above) and (c) is shown. For an alternative proof using rate conservation, let $X_t = I(Q_t \geq j+1)$. The rate of upward jumps is $\mu_A^{-1}\pi_j$ and the rate of downward jumps is $\delta\mathbb{P}(Q = j+1)$. Equating these two quantities yields $\mathbb{P}(Q = j+1) = \rho\pi_j$ which shows that (c) follows immediately from (d).

Theorem 5.2 *Consider the $M/G/1$ queue with interarrival intensity β, service time distribution B, and $\rho = \beta\mu_B < 1$. Then in the steady state:*
(a) *The distributions of the waiting time W and the workload V are the same and given as $H = (1-\rho)\sum_0^\infty \rho^n B_0^{*n}$, where $B_0(x) = \mu_B^{-1}\int_0^x \overline{B}(y)\,\mathrm{d}y$ is the stationary excess distribution.*
(b *The distributions of the queue lengths Q, Q^A, Q^D at an arbitrary point, just before arrivals, resp. just after departures, are the same, say $\boldsymbol{\pi}$, which can be expressed in terms of H and the Poisson distribution by $\pi_0 = 1-\rho$,*

$$\pi_k \;=\; \int_{0+}^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{k-1}}{(k-1)!}\,H(\mathrm{d}t) \;=\; \rho\int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{k-1}}{(k-1)!}\,H * B_0(\mathrm{d}t)\,, \quad (5.2)$$

$k = 1, 2, \ldots$. *In particular,*

$$\mathbb{E}Q \;=\; \rho\bigl[1 + \beta(\mathbb{E}W + \mathbb{E}U^*)\bigr] \;=\; \rho + \beta\mathbb{E}W \;=\; \rho + \frac{\rho^2\mu_B^{(2)}}{2(1-\rho)\mu_B^2}\,. \quad (5.3)$$

Proof. For $W \overset{\mathscr{D}}{=} V$, see III.9.2, VII.6.7 and Section 3. Further, $H = (1-\rho)\sum_0^\infty B_0^{*n}$ is just the Pollaczeck–Khinchine formula VIII.(5.5).

In (b), $Q \overset{\mathscr{D}}{=} Q^A \overset{\mathscr{D}}{=} Q^D$ follows from $W \overset{\mathscr{D}}{=} V$ and Theorems 4.3 and 4.2. Also, Theorem 4.3 yields $\mathbb{P}(Q = 0) = 1-\rho$ and

$$\mathbb{P}(Q \geq k) \;=\; \mathbb{P}\bigl(W > T^{(k-1)}\bigr) \;=\; \int_{0+}^\infty \sum_{\ell=k-1}^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^\ell}{\ell!}\,H(\mathrm{d}t)\,,$$

from which the first part of (5.2) follows; the second follows since (a) shows that $\rho H * B_0$ coincides with H on $(0,\infty)$. The proof of (5.3) is now easy. □

The alternative approach of imbedded Markov chain analysis for $M/G/1$ starts by noting that

$$Q_{n+1}^D \;=\; (Q_n^D - 1)^+ + K_n \quad (5.4)$$

where K_n is the number of customers arriving while customer n is being served. Clearly, $\{Q_n^D\}$ is a Markov chain with transition matrix

$$\boldsymbol{P} \;=\; \begin{pmatrix} q_0 & q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ 0 & q_0 & q_1 & q_2 & \cdots \\ 0 & 0 & q_0 & q_1 & \cdots \\ 0 & 0 & 0 & q_0 & \cdots \\ \vdots & & & & \ddots \end{pmatrix},$$

where $q_k = \mathbb{P}(K_n = k) = \int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^k}{k!}\,B(\mathrm{d}t)$. Irreducibility is obvious since all $q_k > 0$, and also $\mathbb{E}K_n$ is the expected number $\beta\mu_B = \rho$ of arriving customers in a service interval. Thus $\mathbb{E}_i Q_1^D = \rho + i - 1$ for $i \geq 1$, and it is a matter of routine to check from Foster's criteria in I.5 that we have recurrence when $\rho \leq 1$ and ergodicity when $\rho < 1$ (when $\rho = 1$, there is in fact null recurrence, and when $\rho > 1$ there is transience; cf. Problem 5.4).

Assume in the following that $\rho < 1$. Then the equation $\boldsymbol{\pi P} = \boldsymbol{\pi}$ becomes

$$\begin{aligned} \pi_0 &= \pi_0 q_0 + \pi_1 q_0, \\ \pi_1 &= \pi_0 q_1 + \pi_1 q_1 + \pi_2 q_0, \\ \pi_2 &= \pi_0 q_2 + \pi_1 q_2 + \pi_2 q_1 + \pi_3 q_0 \\ &\vdots \end{aligned} \tag{5.5}$$

Letting $r_n = q_{n+1} + q_{n+2} + \cdots$, it follows by adding equations $0, \ldots, n$ and solving for $\pi_{n+1} q_0$ that

$$\begin{aligned} \pi_1 q_0 &= \pi_0 r_0, \\ \pi_2 q_0 &= \pi_0 r_1 + \pi_1 r_1, \\ \pi_3 q_0 &= \pi_0 r_2 + \pi_1 r_2 + \pi_2 r_1 \\ &\vdots \end{aligned} \tag{5.6}$$

If we sum these equations and note that $\sum_0^\infty r_n = \rho$, we get

$$(1 - \pi_0) q_0 \;=\; \pi_0 \rho + (1 - \pi_0)(\rho - r_0),$$

from which it easily follows that $\pi_0 = 1 - \rho$. The remaining π_n are then recursively determined by (5.6), but cannot be found in closed formulas.

However, many properties of $\boldsymbol{\pi}$ can be derived directly from equations (5.4)–(5.6). Let us look at (5.4) which in the limit becomes

$$Q^D \overset{\mathscr{D}}{=} (Q^D - 1)^+ + K \;=\; Q^D - I(Q^D > 0) + K \tag{5.7}$$

(in obvious notation). Taking squared expectations yields

$$\mathbb{E}{Q^D}^2 \;=\; \mathbb{E}{Q^D}^2 + \mathbb{P}(Q^D > 0) + \mathbb{E}K^2 - 2\mathbb{E}Q^D + 2\mathbb{E}Q^D\,\mathbb{E}K - 2\mathbb{P}(Q^D > 0)\mathbb{E}K.$$

Eliminating $\mathbb{E}{Q^D}^2$ and solving for $\mathbb{E}Q^D$ using $\mathbb{E}K = \rho$, $\mathbb{P}(Q^D = 0) = 1 - \pi_0 = \rho$ and

$$\mathbb{E}K^2 \;=\; \int_0^\infty \sum_{k=0}^\infty k^2 \mathrm{e}^{-\beta t}\frac{(\beta t)^k}{k!}\,B(\mathrm{d}t) \;=\; \int_0^\infty [\beta t + (\beta t)^2]\,B(\mathrm{d}t) \;=\; \rho + \beta^2 \mu_B^{(2)}$$

then easily yields the same expression as in (5.3) ($Q^D \overset{\mathscr{D}}{=} Q$ will be shown in a moment). Also the generating function $\widehat{\pi}[s] = \sum_0^\infty s^n \pi_n = \mathbb{E}s^{Q^D}$ can be found in the same way. In fact, (5.7) yields

$$\begin{aligned}
\widehat{\pi}[s] &= \mathbb{E}s^{Q^D - I(Q^D>0)}\mathbb{E}s^K \;=\; (\pi_0 + \pi_1 + s\pi_2 + s^2\pi_3)\sum_{n=0}^{\infty} s^n q_n, \\
s\widehat{\pi}[s] &= \bigl[\widehat{\pi}[s] + \pi_0(s-1)\bigr]\widehat{q}[s] \;=\; \bigl[\widehat{\pi}[s] + (1-\rho)(s-1)\bigr]\widehat{q}[s], \\
\widehat{\pi}[s] &= \frac{(1-\rho)(1-s)\widehat{q}[s]}{\widehat{q}[s]-s}
\end{aligned} \tag{5.8}$$

where, letting $\widehat{B}[\cdot]$ denote the Laplace transform of B,

$$\widehat{q}[s] \;=\; \int_0^\infty \sum_{k=0}^{\infty} \mathrm{e}^{-\beta t}\frac{(s\beta t)^k}{k!}\,B(\mathrm{d}t) \;=\; \int_0^\infty \mathrm{e}^{-\beta t(1-s)}\,B(\mathrm{d}t) \;=\; \widehat{B}\bigl[\beta(1-s)\bigr]\,.$$

To proceed from Q^D to Q, we use again semi–regeneration. The imbedded Markov chain is $\{Q_n^D\}$ with stationary distribution $\boldsymbol{\pi}$, and a cycle C started by $Q_n^D = k \geq 1$ is just a service interval of length U; for $k=0$ we have to add the idle period of expected length $1/\beta$. It follows that for $j \geq 1$ we have

$$\mathbb{P}(Q=j) \;=\; \frac{1}{m}\sum_{k=0}^{\infty}\pi_k\,\mathbb{E}_k\!\int_0^C I(Q_t=j)\,\mathrm{d}t,$$

where

$$m \;=\; \pi_0\,(1/\beta + \mu_B) + \sum_{i=1}^{\infty}\pi_i\mu_B \;=\; \frac{1-\rho}{\beta} + \mu_B \;=\; \frac{1}{\beta}.$$

For $j \geq 1$ fixed, write $\alpha_k = \mathbb{E}_k \int_0^C I(Q_t=j)\mathrm{d}t$. Then $\alpha_0 = \alpha_1$. For $k > j$ we have $\alpha_k = 0$, whereas for $1 \leq k \leq j$ we get

$$\begin{aligned}
\alpha_k &= \int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{j-k}}{(j-k)!}\overline{B}(t)\,\mathrm{d}t \;=\; \int_0^\infty \beta^{-1}\sum_{\ell=j-k+1}^{\infty}\mathrm{e}^{-\beta t}\frac{(\beta t)^\ell}{\ell!}\,B(\mathrm{d}t) \\
&= \beta^{-1}r_{j-k}.
\end{aligned}$$

It follows that

$$\mathbb{P}(Q=j) \;=\; \beta\sum_{k=0}^{\infty}\pi_k\alpha_k \;=\; \pi_0 r_{j-1} + \sum_{k=1}^{j}\pi_k r_{j-k} \;=\; \pi_j,$$

where the last equality follows from (5.6). The truth of this for all $j \geq 1$ implies $Q \overset{\mathscr{D}}{=} Q^D$.

For an alternative proof using rate conservation, let $X_t = I(Q_t \geq j+1)$. The rate of upward jumps is $\beta\mathbb{P}(Q=j)$ and the rate of downward jumps is $\beta\mathbb{P}(Q^D=j)$ (interpret β as the departure rate). Equating these two quantities yields $Q \overset{\mathscr{D}}{=} Q^D$.

Problems

5.1 Derive the steady–state characteristics of the $GI/G/1$ queue where $U-1$ is exponential with rate say δ and $T \geq 1$.
5.2 Check that the formulas for $\mathbb{E}W$ (see VIII.5.7) and $\mathbb{E}Q$ in $M/G/1$ are in agreement with Little's formula, and that $\mathbb{E}W = \mathbb{E}V$ in agreement with Corollary 3.5.
5.3 Show that $\mathbb{E}W$ and $\mathbb{E}Q$ in $M/G/1$ are minimized by $M/D/1$ subject to the constraints that β and μ_B are fixed. See further XI.5.
5.4 Show by a modification of the derivation of $\pi_0 = 1-\rho$ from (5.6) that the stationary measure is infinite for $\rho = 1$ and that therefore $\{Q_n^D\}$ is null recurrent. Show also that there is transience for $\rho > 1$. [*Hint:* $Y_{n+1} \geq Y_n - 1 + K_n$.]
5.5 Give a direct derivation of (5.8) by multiplying equation n in (5.6) by s^{n+1} and summing over n. Check the formula for the mean by differentiation.
5.6 Let R_t denote the attained service of the customer in service at time t (if any) and define $D_n(x) = \mathbb{P}(R \leq x, Q = n)$, $n = 1, 2, \ldots$ (thus $\|D_n\| = \pi_n$). Show that D_n has density

$$\beta\overline{B}(x)\mathrm{e}^{-\beta t}\left\{(\pi_0+\pi_1)\frac{(\beta t)^{n-1}}{(n-1)!} + \sum_{k=0}^{n-2}\pi_{n-k}\frac{(\beta t)^k}{k!}\right\}.$$

Notes A further classical topic for the $M/G/1$ queue is the connection of the busy period to branching processes. This is most readily understood in the preemptive LCFS setting (where the busy period distribution is the same as for FCFS). Here one defines the children of a particular customer as the customers who arrived while he was in service. A simple example of the connection is then that the number of customers served in the busy period is the same as the total number of progeny of the customer initiating the busy period. A fairly general formulation is in Shalmon (1988) who also gives references to earlier work (to which we add Neuts, 1969). A recent generalization goes from the compound Poisson $M/G/1$ case to Lévy processes, see LeGall and Le Yan (1998).

6 Continuity of the Waiting Time

We consider here and in the next two sections a family of $GI/G/1$ queueing systems indexed by $k = 0, 1, 2, \ldots$ with service time distribution $B^{(k)}$, interarrival distribution $A^{(k)}$ and $U_n^{(k)}$, $T_n^{(k)}$, $X_n^{(k)}$, $S_n^{(k)}$, $W_n^{(k)}$, $W^{(k)}$, etc. defined the obvious way. The problem, stated in a rough form, is to study the limiting behaviour of $W^{(k)}$ as $k \to \infty$ under appropriate conditions, assuming that $\rho_k < 1$ for $k = 1, 2, \ldots$ and that $A^{(k)} \xrightarrow{w} A^{(0)}$, $B^{(k)} \xrightarrow{w} B^{(0)}$ (weak convergence). In Sections 7, 8 we consider the extreme cases where the limit has traffic intensity $\rho_0 = 1$ or $\rho_0 = 0$, whereas the situation here is $0 < \rho_0 < 1$. It is then reasonable to ask for conditions under which $W^{(k)} \xrightarrow{\mathscr{D}} W^{(0)}$. This is denoted as a *continuity* (or *stability* or *robustness*) property of the waiting time, and is of importance for example to justify the approximation of a queueing system with given $A^{(0)}$, $B^{(0)}$ by systems with $A^{(k)}$, $B^{(k)}$ of phase type (cf. III.4).

To facilitate notation, we suppress from now on indices n and $k = 1, 2, \ldots$ whenever convenient (thus, e.g. $\mathbb{E}U \to \mathbb{E}U^{(0)}$ or $\lim_{k\to\infty} \mathbb{E}U = \mathbb{E}U^{(0)}$ means $\mathbb{E}U_n^{(k)} \to \mathbb{E}U_n^{(0)}$).

We shall first state and prove the main result in random walk terms, and thereafter reformulate in terms more natural for queues.

Theorem 6.1 *Consider random walks* $\{S_n\}_{n\in\mathbb{N}}$, $\{S_n^{(0)}\}_{n\in\mathbb{N}}$ *with* $\mu = \mathbb{E}X < 0$, $F \xrightarrow{w} F^{(0)}$, $k \to \infty$, $\mu_0 = \mathbb{E}X_n^{(0)} < 0$. *Then* $M \xrightarrow{\mathscr{D}} M^{(0)}$ *provided that the* X^+ *are uniformly integrable or equivalently that* $\mathbb{E}X^+ \to \mathbb{E}X^{(0)+}$.

The key step of the proof is

Lemma 6.2 *Define* $K_n = \max_{r\geq n} S_r$. *Then* $\overline{\lim}_{n\to\infty}\ \overline{\lim}_{k\to\infty} \mathbb{P}(K_n > 0) = 0$.

Proof. By general results on weak convergence, $\mathbb{E}X^+ \to \mathbb{E}X^{(0)+}$ is equivalent to the uniform integrability of the X^+ since $X^+ \xrightarrow{\mathscr{D}} X^{(0)+}$. Choose $c < 0$ such that $\mathbb{E}\big[X^{(0)+} \vee c\big] < 0$ and define $\breve{X}_n = X_n \vee c$. Then $\breve{X}_n \xrightarrow{\mathscr{D}} \breve{X}^{(0)}$, $\breve{S}_n \geq S_n$, $\breve{K}_n \geq K_n$. Hence for the proof it is no restriction to assume that the X are uniformly bounded below, say by c. Then the X themselves are uniformly integrable, hence $\mu = \mathbb{E}X \to \mu_0 < 0$. Now for $\mu < 0$,

$$\begin{aligned}\mathbb{P}(K_n > 0) &= \mathbb{P}\bigg(\max_{r\geq n} \frac{S_r}{r} > 0\bigg) = \mathbb{P}\bigg(\max_{r\geq n}\bigg\{\frac{S_r}{r} - \mu\bigg\} > -\mu\bigg) \\ &\leq \frac{1}{|\mu|}\mathbb{E}\bigg|\frac{S_n}{n} - \mu\bigg|,\end{aligned}$$

using the fact that $\{S_r/r - \mu\}_{r=n,n+1,\ldots}$ is a backward martingale and Kolmogorov's inequality. Decompose $S_n - n\mu$ as $\widetilde{S}_n + \widetilde{\widetilde{S}}_n$, where

$$\widetilde{X}_n = X_n I(X_n \leq d) - \mathbb{E}[X_n; X_n \leq d],\ \widetilde{\widetilde{X}}_n = X_n I(X_n > d) - \mathbb{E}[X_n; X_n > d]$$

with d satisfying $\mathbb{P}(X_n^{(0)} = d) = 0$, $\mathbb{E}[X_n; X_n > d] < \epsilon$ for all k. Then $\widetilde{\sigma}^2 = \mathbb{V}ar\widetilde{X}_n \to \widetilde{\sigma}_0^2 = \mathbb{V}ar\widetilde{X}_n^{(0)}$ since the $\widetilde{X}$ are bounded uniformly in k, so that

$$\mathbb{E}|S_n/n - \mu| \ \leq\ \mathbb{E}|\widetilde{S}_n/n| + \mathbb{E}|\widetilde{\widetilde{S}}_n/n| \ \leq\ \widetilde{\sigma}/\sqrt{n} + 2\epsilon,$$

using the Cauchy–Schwarz inequality. Hence

$$\overline{\lim_{n\to\infty}}\ \overline{\lim_{k\to\infty}}\ \mathbb{P}(K_n > 0) \ \leq\ \overline{\lim_{n\to\infty}} \frac{1}{|\mu_0|}\big[\widetilde{\sigma}_0/\sqrt{n} + 2\epsilon\big] \ =\ \frac{2\epsilon}{|\mu_0|},$$

and since ϵ is arbitrary, the proof is complete. □

Proof of Theorem 6.1. From $X \xrightarrow{\mathscr{D}} X^{(0)}$ it follows that $\{X_r\}_{r=0}^n \xrightarrow{\mathscr{D}} \{X_r^{(0)}\}_{r=0}^n$ and hence by the continuous mapping theorem $M_n \xrightarrow{\mathscr{D}} M_n^{(0)}$.

Now let $x > 0$ satisfy $\mathbb{P}(M^{(0)} = x) = 0$. Then also $\mathbb{P}(M_n^{(0)} = x) = 0$ for each n and hence

$$\begin{aligned}\limsup_{k\to\infty} \mathbb{P}(M^{(k)} > x) &\le \limsup_{k\to\infty} \bigl[\mathbb{P}(M_n^{(k)} > x) + \mathbb{P}(K_n^{(k)} > 0)\bigr] \\ &= \mathbb{P}(M_n^{(0)} > x) + \limsup_{k\to\infty} \mathbb{P}(K_n^{(k)} > 0), \\ \liminf_{k\to\infty} \mathbb{P}(M^{(k)} > x) &\ge \liminf_{k\to\infty} \mathbb{P}(M_n^{(k)} > x) \;=\; \mathbb{P}(M_n^{(0)} > x).\end{aligned}$$

Letting $n \to \infty$ yields $\mathbb{P}(M > x) \to \mathbb{P}(M^{(0)} > x)$. Hence $M \xrightarrow{\mathscr{D}} M^{(0)}$ (note that $x = 0$ is not a continuity point of M). □

Apparently the point mass of M at zero is of particular interest, but $\mathbb{P}(M = 0) \to \mathbb{P}(M^{(0)} = 0)$ does not follow alone from $M \xrightarrow{\mathscr{D}} M^{(0)}$. However:

Proposition 6.3 *If in addition to the assumptions of Theorem 6.1 the distribution $F^{(0)}$ of $X^{(0)}$ is continuous, then $\mathbb{P}(M = 0) \to \mathbb{P}(M^{(0)} = 0)$.*

Proof. The assumptions ensure that $\mathbb{P}(S_n^{(0)} = 0) = 0$ for each $n \ge 1$ and hence $\mathbb{P}(M_n = 0) \to \mathbb{P}(M_n^{(0)} = 0)$. Now argue exactly as above. □

Corollary 6.4 *Consider for $k = 0, 1, 2, \ldots$ $GI/G/1$ queues with $A \xrightarrow{w} A^{(0)}$, $B \xrightarrow{w} B^{(0)}$, $\rho_0 < 1$. Then $W \xrightarrow{\mathscr{D}} W^{(0)}$ provided that the U are uniformly integrable, or equivalently, that $\mathbb{E}U \to \mathbb{E}U^{(0)}$. If in addition either $A^{(0)}$ or $B^{(0)}$ is continuous, then also $\mathbb{P}(W = 0) \to \mathbb{P}(W^{(0)} = 0)$.*

Proof. Appealing to the interpretation $X = U - T$, $W \overset{\mathscr{D}}{=} M$, it is straightforward to check the assumptions of Theorem 6.1 and Proposition 6.3 (the uniform integrability of the X^+ follows from $X^+ \le U$ and the uniform integrability of the U). □

Problems

6.1 Let $F^{(k)}$ be concentrated at $-1, k$ with point masses $1 - 1/2k$, $1/2k$ and let $F^{(0)} = \lim F^{(k)}$. Show that

$$\mathbb{P}(M^{(k)} \ge 1) \;\ge\; \mathbb{P}(X_n = k \text{ for some } n = 1, \ldots, k) \;\to\; \mathrm{e}^{-1/2}$$

and deduce that $M^{(k)} \xrightarrow{\mathscr{D}} M^{(0)}$ does not hold.

Notes Continuity problems are treated e.g. in Borovkov (1976), Stoyan (1983), Brandt *et al.* (1990) and Kalashnikov (1994). A classical reference for Markov chains is Karr (1975).

7 Heavy Traffic Limit Theorems

If, in the set–up of Section 6, the limiting traffic intensity ρ_0 is 1 rather than < 1, we are in the situation of *heavy traffic* where all queueing systems are heavily congested. We expect again $W = W^{(k)} \xrightarrow{\mathscr{D}} W^{(0)}$, but now $W^{(0)} = \infty$ a.s. It will turn out that a more precise result can be obtained, namely that under weak conditions $|\mu|W$ is approximately exponentially distributed. We start again by formulating this for the random walk setting (σ^2 denotes $\mathbb{V}arX$).

Theorem 7.1 *Consider random walks* $\{S_n\}_{n\in\mathbb{N}}$, $\{S_n^{(0)}\}_{n\in\mathbb{N}}$ *with* $\mu < 0$, $\mu \to 0$, $\underline{\lim}_{k\to\infty} \sigma^2 > 0$, *and the* X^2 *uniformly integrable. Then* $Y = |\mu|M/\sigma^2$ *is approximately exponentially distributed with intensity* 2, *i.e.* $\mathbb{P}(Y > y) \to \mathrm{e}^{-2y}$. *Furthermore,* $\mathbb{E}Y \to 1/2$.

Remark 7.2 The conditions of Theorem 7.1 are not intrinsically different from the apparently stronger

$$X \xrightarrow{\mathscr{D}} X^{(0)}, \quad \sigma^2 \to \sigma_0^2 > 0, \quad \mu_0 = 0. \tag{7.1}$$

Indeed, the uniform integrability ensures that $\{F\} = \{F^{(k)}\}$ is tight. Thus every subsequence $\{k'\}$ has a weakly convergent subsequence $\{k''\}$, i.e. $X^{(k'')} \xrightarrow{\mathscr{D}} X^{(0)}$ for some $X^{(0)}$. But then by uniform integrability, $\mu_0 = \lim \mu_{k''} = 0$, $\sigma_0^2 = \lim \sigma_{k''}^2 > 0$. Furthermore, a standard analytical argument shows that if we can show the asymptotic exponentiality for $\{k''\}$, then it will hold for $\{k'\}$ as well. Hence *for the proof we can* (and shall) *assume that* (7.1) *holds.* Also, by rescaling, we may take $\sigma^2 = 1$; then $Y = |\mu|M = -\mu M$. □

Two approaches to Theorem 7.1 will be considered, the first being based on characteristic functions $\varphi_Y(y) = \mathbb{E}\mathrm{e}^{\mathrm{i}yY}$. Thus we have to show $\varphi_Y(y) = \varphi_M(-\mu y) \to (1 - \mathrm{i}y/2)^{-1}$. In the proof, we let $\mu_2 = \mathbb{E}X^2$ (thus $\mu_2 \to 1$ since $\mu \to 0$, $\sigma^2 \to 1$).

Lemma 7.3 *For each* y, *it holds as* $k \to \infty$ *that*

$$\varphi_X(-\mu y) \;=\; 1 - \mathrm{i}\mu^2 y - \frac{\mu^2 y^2}{2} + \mathrm{o}(\mu^2). \tag{7.2}$$

Proof. Define $g(z) = \mathrm{e}^{\mathrm{i}yz} - 1 - \mathrm{i}yz + y^2z^2/2$. Then for each $\epsilon > 0$, we can bound $|g(z)|$ by $c_\epsilon|z^3|$ for $|z| \le \epsilon$ and by $d_\epsilon|z^2|$ for $|z| > \epsilon$. Hence

$$\begin{aligned}\big|\mathbb{E}g(-\mu X)\big| \;&\le\; c_\epsilon\mathbb{E}\big[|-\mu X\big|^3;\ \big|-\mu X\big| \le \epsilon\big] + d_\epsilon\mathbb{E}\big[(\mu X)^2;\ \big|-\mu X\big| > \epsilon\big] \\ &\le\; \mu^2\big\{\epsilon c_\epsilon \mathbb{E}X^2 + d_\epsilon\mathbb{E}\big[(\mu X)^2;\ |-\mu X| > \epsilon\big]\big\}\end{aligned}$$

and therefore $\limsup_{k\to\infty} \mu^{-2}|\mathbb{E}g(-\mu X)| \le \epsilon c_\epsilon$ by uniform integrability. Since c_ϵ remains bounded as $\epsilon \downarrow 0$, it follows that

$$\varphi_X(-\mu y) - \Big(1 - \mathrm{i}\mu^2 y - \frac{\mu^2 y^2}{2}\mu_2\Big) \;=\; \mathbb{E}g(-\mu X) \;=\; \mathrm{o}(\mu^2),$$

and the lemma follows since $\mu_2 \to 1$. □

Proof of Theorem 7.1. We first note that as in Section 6 we have $M \xrightarrow{\mathscr{D}} \infty$. But

$$\mathbb{E}\big[(M+X)^-\big]^2 \;\le\; \mathbb{E}(X^-)^2\mathbb{P}(M \le c) + \mathbb{E}[X^2;\, X < -c].$$

Letting first $k \to \infty$ and next $c \to \infty$ yields

$$\mathbb{E}\big[(M+X)^-\big]^2 \to 0 \;\;\text{(hence } \mathbb{E}(M+X)^- \to 0). \tag{7.3}$$

From this $\mathbb{E}Y \to 1/2$ is clear from (2.5). Now for each z, $\mathrm{e}^{\mathrm{i}yz^+} = \mathrm{e}^{\mathrm{i}yz} + 1 - \mathrm{e}^{-\mathrm{i}yz^-}$. Letting $Z = M + X$ and taking expectations we get

$$\varphi_M(y) \;=\; \varphi_M(y)\varphi_X(y) + 1 - \varphi_{-(M+X)^-}(y) \;=\; \frac{1-\varphi_{-(M+X)^-}(y)}{1-\varphi_X(y)}. \tag{7.4}$$

Since $\mathrm{e}^{\mathrm{i}z} - 1 - \mathrm{i}z \;=\; z^2\mathrm{O}(1)$ for z real, we get

$$\begin{aligned}\varphi_{-(M+X)^-}(-\mu y) \;&=\; 1 + \mathrm{i}\mu y\mathbb{E}(M+X)^- + \mathrm{O}(1)\mu^2y^2\mathbb{E}\big[(M+X)^-\big]^2 \\ &=\; 1 - \mathrm{i}\mu^2 y + o(\mu^2),\end{aligned}$$

using (1.9) and (7.3). Hence by Lemma 7.3 and (7.4),

$$\varphi_M(-\mu y) \;=\; \frac{\mathrm{i}\mu^2 y + \mathrm{o}(\mu^2)}{\mathrm{i}\mu^2 y + \mu^2y^2/2 + \mathrm{o}(\mu^2)} \;\to\; \frac{1}{1-\mathrm{i}y/2}. \qquad \square$$

The second proof of Theorem 7.1 involves more advanced tools (weak convergence in function space) but is perhaps more illuminating and yields additional information, namely asymptotics of the $M_n^{(k)}$. We let $\{B_\xi(t)\}_{t\ge 0}$ denote Brownian motion with unit variance and drift ξ. The *inverse Gaussian distribution function* $G(t;\,\xi,c)$ with parameters $\xi \in \mathbb{R}$, $c > 0$ is the c.d.f. of the first passage time $\tau(\xi,c) = \inf\{t > 0:\, B_\xi(t) \ge c\}$,

$$G(T;\,\xi,c) \;=\; \mathbb{P}(\tau(\xi,c) \le T) \;=\; \mathbb{P}\Big(\max_{0\le t\le T} B_\xi(t) \ge c\Big). \tag{7.5}$$

This distribution (defective for $\xi < 0$) can in fact be found explicitly. We defer the derivation to XIII.4 and here use only the formula

$$\big\|G(\cdot;\,\xi,c)\big\| \;=\; \mathbb{P}\Big(\max_{0\le t<\infty} B_\xi(t) \ge c\Big) \;=\; \mathrm{e}^{2\xi c}, \quad \xi < 0. \tag{7.6}$$

Proposition 7.4 *Under the conditions of Theorem 7.1, it holds for any $T < \infty$ that*

$$\frac{|\mu|}{\sigma^2}M_{\lfloor T\sigma^2/\mu^2\rfloor} \xrightarrow{\mathscr{D}} \max_{0\le t\le T} B_{-1}(t), \quad \mathbb{P}\left(\frac{|\mu|}{\sigma^2}M_{\lfloor T\sigma^2/\mu^2\rfloor} > y\right) \to G(T;\,-1,y).$$

Proof. We may again assume that (7.1) holds with $\sigma_0^2 = 1$. Let $\{c\} = \{c^{(m)}\}$ be any sequence with $c^{(m)} \to \infty$ and define

$$B(t) \;=\; B^{(m)}(t) \;=\; \frac{1}{\sqrt{c}}\big[S_{\lfloor ct\rfloor} - \lfloor ct\rfloor\mu\big].$$

It then follows from the invariance principle (Donsker's theorem) in its standard form (e.g. Billingsley, 1968, Ch. 3) that $B \xrightarrow{\mathscr{D}} B_0$ in D. Taking $c = \mu^{-2}$ we have $\lfloor ct \rfloor \mu/\sqrt{c} \to -t$, i.e.

$$\begin{aligned}\big\{|\mu| S_{\lfloor t/\mu^2 \rfloor}\big\}_{0 \le t < \infty} &= \big\{B(t) + \lfloor ct \rfloor \mu/\sqrt{c}\big\}_{0 \le t < \infty} \\ &\overset{\mathscr{D}}{\to} \big\{B_0(t) - t\big\}_{0 \le t < \infty} \overset{\mathscr{D}}{=} B_{-1}.\end{aligned}$$

Hence, since $f \to \sup_{0 \le t \le T} f(t)$ is continuous a.e. on D w.r.t. any probability distribution concentrated on the continuous functions, it follows from the continuity of B_{-1} that

$$|\mu| M_{\lfloor T/\mu^2 \rfloor} = \sup_{0 \le t \le T} |\mu| S_{\lfloor t/\mu^2 \rfloor} \overset{\mathscr{D}}{\to} \max_{0 \le t \le T} B_{-1}(t)$$

which yields the desired conclusion in view of $\sigma^2 \to 1$. □

Proof of Theorem 7.1. We assume again $\sigma^2 \to 1$ and write

$$Y = Y_1 \vee Y_2 = \big(|\mu| M_{\lfloor T/\mu^2 \rfloor}\big) \vee \Big(\sup_{n > T/\mu^2} |\mu| S_n\Big).$$

Here by (7.6) and Proposition 7.4,

$$\lim_{T \to \infty} \lim_{k \to \infty} \mathbb{P}(Y_1 > y) = \lim_{T \to \infty} G(T; -1, y) = \mathrm{e}^{-2y}, \tag{7.7}$$

whereas $\big\{(S_n - n\mu)^2\big\}$ is a backward submartingale, hence

$$\begin{aligned}\mathbb{P}(Y_2 > 0) &= \mathbb{P}\Big(\max_{n > T/\mu^2} (S_n/n - \mu) > -\mu\Big) \\ &\le \frac{1}{\mu^2} \mathbb{E}\big[S_{\lfloor T/\mu^2 \rfloor}/_{\lfloor T/\mu^2 \rfloor} - \mu\big]^2 = \frac{\sigma^2}{\mu^2 \lfloor T/\mu^2 \rfloor},\end{aligned}$$

$$\lim_{T \to \infty} \lim_{k \to \infty} \mathbb{P}(Y_2 > 0) \le \lim_{T \to \infty} \frac{1}{T} = 0. \tag{7.8}$$

Combining (7.7) and (7.8), the desired conclusion is obtained exactly as in the proof of Theorem 6.1. □

Corollary 7.5 *Consider* $GI/G/1$ *queueing systems with* $A \xrightarrow{w} A^{(0)}$, $B \xrightarrow{w} B^{(0)}$, *where* $A^{(0)}$, $B^{(0)}$ *are not both degenerate,* $\rho < 1$, $\rho \to \rho_0 = 1$ *and the* U^2, T^2 *uniformly integrable. Then* $Y = |\mu| W/\sigma^2$ *is approximately exponentially distributed with intensity* 2 *and* $\mathbb{E}Y \to 1/2$. *Here* $\mu = \mathbb{E}X = \mathbb{E}U - \mathbb{E}T$, $\sigma^2 = \mathbb{V}arX$. *Furthermore, for each* T

$$\mathbb{P}\left(\frac{|\mu|}{\sigma^2} W_{\lfloor T\sigma 2/\mu^2 \rfloor} > y\right) \to G(T; -1, y).$$

The proof is a routine application of Theorem 7.1 and is omitted.

Results of the type in Corollary 7.5 are of high potential relevance, since the heavy traffic situation occurs widely in practice (when designing a service facility, one usually avoids for economical reasons to keep the server idle for a large proportion of the time). Given a queue with ρ smaller than but close to 1, we may imbed the system in the set–up of Corollary 7.5, writing $A = A^{(k)}$, $B = B^{(k)}$ for some large k. It is then suggested that the following approximations may be used:

$$\mathbb{E}W \approx \mathbb{V}arX/2\mathbb{E}(-X), \quad \mathbb{P}(W > y) \approx \exp\bigl\{-2\mathbb{E}(-X)y/\mathbb{V}arX\bigr\}. \qquad (7.9)$$

Note that when $\mathbb{E}X \approx 0$, we have $\mathbb{E}X^2 \approx \mathbb{V}arX$ and one may thus replace $\mathbb{V}arX$ by $\mathbb{E}X^2$ in (7.9). However, inspection of (2.5) shows that $\mathbb{E}X^2/2\mathbb{E}(-X)$ and $\mathbb{V}arX/2\mathbb{E}(-X)$ are both upper bounds for $\mathbb{E}W$ and hence $\mathbb{V}arX/2\mathbb{E}(-X)$ is the best approximation.

We return to a special aspect of heavy traffic approximations in XIII.6, but finally we mention that in view of the formulas of Sections 3 and 4, it is straightforward to derive analogues of Corollary 7.5 for workload, queue length and so on (cf. Problem 7.1).

Problems

7.1 Show that under the conditions of Corollary 7.5 the steady–state workload V has the same limiting distribution as W. Show similarly, using the results of Section 4, that $|\mu|Q/\sigma^2$, $|\mu|Q^A/\sigma^2$ have limiting exponential distributions with intensities $2\mu_A$.

7.2 Show (7.6) by optional stopping of the martingale $\{\mathrm{e}^{-2\xi B_\xi(t)}\}$ at $\tau(\xi,c)\wedge T$.

Notes Heavy traffic limit theory was largely initiated by Kingman in the 1960s, with the functional CLT point of view being developed by Iglehart and Whitt. For surveys, see Glynn (1990) and Whitt (2002).

Without second moments, one often gets a stable rather than a Brownian limit. See e.g. Furrer *et al.* (1997) and Heath *et al.* (1999) for recent papers in the area, and Whitt (2002) for a survey and references.

A notable recent development is heavy traffic limit theory for queueing networks, where the limit is reflected Brownian motion in an orthant. See further the Notes to IV.6 and IX.2.

8 Light Traffic

Intuitively, light traffic means that the generic interarrival time T is much larger than the generic service time U, implying that typically the system is idle in the steady state. When considering the $GI/G/1$ queue at an arbitrary point of time, the idleness probability is $1-\rho = \mathbb{P}(Q=0) = \mathbb{P}(V=0)$, so that light traffic certainly requires ρ to be close to 0. A more refined question is to study the behaviour of Q, V given $\{Q>0\} = \{V>0\}$. To this end, we consider a sequence of $GI/G/1$ queues in the notation of Sections 6 and 7, assuming throughout that the interarrival time $T = T^{(k)}$ satisfies $T \xrightarrow{\mathscr{D}} \infty$, $k \to \infty$, and that the service time distribution B is fixed, i.e. does not depend on $k = 1, 2, \ldots$ (this certainly implies $\rho \to 0$).

Proposition 8.1 *As $k \to \infty$, it holds without further conditions that* (a) $W \to 0$ *in t.v.,* (b) V *conditionally upon* $V > 0$ *converges to the equilibrium service time* U^* *in t.v.,* (c) Q *conditionally upon* $Q > 0$ *converges to* 1 *in t.v.*
[For basic facts about total variation convergence, see A8.]

Proof. Let Z_ϵ denote a r.v. that is 0 w.p. ϵ and ϵ^{-1} w.p. $1-\epsilon$. Then $T^{(k)}$ is stochastically larger than Z_ϵ for all large k so that $W \leq_{\text{so}} M_\epsilon$ (stochastical order), the maximum of a random walk with increments distributed as $U - Z_\epsilon$. Since $M_\epsilon \leq_{\text{so}} M_\delta \stackrel{\mathscr{D}}{=}$ when $\epsilon < \delta$, we get

$$M_\epsilon \stackrel{\mathscr{D}}{=} (M_\epsilon + U - Z_\epsilon)^+ \leq_{\text{so}} (M_\delta + U - Z_\epsilon)^+$$

which converges in t.v. to 0 as $\epsilon \downarrow 0$. Hence $\mathbb{P}(M_\epsilon > 0) \to 0$ and therefore $P(W > 0) \to 0$, proving (a). It follows by Theorem 3.4 that

$$\mathbb{P}(V \in A \,|\, V > 0) \;=\; \mathbb{P}(W + U^* \in A) \;\sim\; \mathbb{P}(U^* \in A)$$

uniformly in $A \subseteq (0, \infty)$, showing (b). For (c), (4.2) then yields

$$\mathbb{P}(Q \geq 2) \;=\; \rho\mathbb{P}(W + U^* > T) \;\sim\; \rho\mathbb{P}(U^* > T) \;=\; \mathrm{o}(\rho),$$

$$\mathbb{P}(Q = 1 \,|\, Q > 0) \;=\; 1 - \mathbb{P}(Q \geq 2 \,|\, Q > 0) \;=\; 1 - \frac{\mathrm{o}(\rho)}{\rho} \;\to\; 1. \qquad \square$$

The intuitive content of Proposition 8.1(b),(c) is that a busy cycle in light traffic with high probability only contains one customer and that if we observe the system at an arbitrary point of time and see it busy, it is because we sample the single service time in the cycle rather than the following idle period. The situation at arrival instants is different: if a customer has to wait ($W > 0$), we expect him to be customer $n = 1$ in the cycle, not $n = 0$ who does not have to wait, so that W given $W > 0$ should most often be the residual service time $U_0 - T_0$ of the previous customer given it is positive. To rigorously verify this intuition as well as to derive precise asymptotics of $\mathbb{P}(W > 0)$ is, however, more difficult than in the case of V and will occupy the rest of this section.

We start again in a triangular array random walk setting, where we are given random walks $\{S_n\} = \{S_n^{(k)}\}$ with increments $X_0, X_1, \ldots$, increment distributions $F(x) = \mathbb{P}(X \leq x)$, maxima $M = \max_{n=0,1,\ldots} S_n$ etc. (indexed by $k = 1, 2, \ldots$). Call two families $\{R\} = \{R^{(k)}\}$, $\{S\} = \{S^{(k)}\}$ of r.v.'s with values in $[0, \infty)$ *light traffic equivalent* if

$$\mathbb{P}(R > 0) \to 0,\ \mathbb{P}(S > 0) \to 0,\ \frac{\mathbb{P}(R > 0)}{\mathbb{P}(S > 0)} \to 1 \tag{8.1}$$

as $k \to \infty$ and the conditional t.v. distance converges to 0,

$$\big\|\mathbb{P}(R \in \cdot \,|R > 0) - \mathbb{P}(S \in \cdot \,|S > 0)\big\| \;\to\; 0. \tag{8.2}$$

Theorem 8.2 *Assume that* $X \xrightarrow{\mathscr{D}} -\infty$ *as* $k \to \infty$*, and that*

$$0 \;=\; \lim_{a\uparrow\infty}\overline{\lim_{k\to\infty}}\;\frac{\mathbb{E}[X;X>a]}{p_+} \;=\; \lim_{a\uparrow\infty}\overline{\lim_{k\to\infty}}\;\frac{\int_a^\infty x\,F(\mathrm{d}x)}{p_+} \tag{8.3}$$

where $p_+ = \mathbb{P}(X>0) = \int_0^\infty F(\mathrm{d}x)$*. Then* M *and* X^+ *are light traffic equivalent.*

Remark 8.3 Let F_+ denote the conditional distribution of X given $X>0$. Then (8.3) means that the family $\{F_+\}$ is uniformly integrable. This should be compared with the unconditional uniform integrability conditions for heavy traffic in Section 7. □

The key step in the proof is (take $S_{\tau_+}=0$ when $\tau_+=\infty$):

Lemma 8.4 *The ascending ladder heights* S_{τ_+} *and the* X^+ *are light traffic equivalent.*

Proof. By (1.7), we can write $G_+ = L+K$ where L,K are the restriction to $(0,\infty)$ of F, resp. $F * \sum_1^\infty G_-^{*n}$ (L is the contribution from the atom of U_- at zero). For simplicity of notation, let $R=R^{(k)}$ be the measure $R(\mathrm{d}x) = \sum_1^\infty G_-^{*n}\big(\mathrm{d}(-x)\big)$ on $(0,\infty)$. Then for $z\ge 0$,

$$\begin{aligned}\overline{K}(z) \;&=\; \sum_{n=1}^\infty \int_{-\infty}^0 \overline{F}(z-x)\,G_-^{*n}(\mathrm{d}x) \;=\; \int_0^\infty \overline{F}(z+x)R(\mathrm{d}x)\\ &=\; \int_z^\infty R(y-z)F(\mathrm{d}y) \;\le\; \int_z^\infty R(y)F(\mathrm{d}y).\end{aligned} \tag{8.4}$$

To proceed from (8.4), we will need the estimate

$$R(t) \;\le\; \varphi(t)(1+t), \tag{8.5}$$

where $\varphi(t)$ is bounded uniformly in k, nondecreasing and tends to 0 for any fixed t as $k\to\infty$. First $X \xrightarrow{\mathscr{D}} -\infty$ implies $S_{\tau_-} \xrightarrow{\mathscr{D}} -\infty$ (with high probability S_{τ_-} coincides with X_0). In particular, $\overline{G}_-(-t)\to 0$ for all $t>0$. Since $R(1) \le \overline{G}_-(-1)\big(1+R(1)\big)$, this implies that $R(1)$ is bounded. Similarly, $R(n-1,n] \le \overline{G}(-n)\big(1+R(1)\big)$ so that $R(n) \le \overline{G}_-(-n)n\big(1+R(1)\big)$, and from these estimates (8.5) follows.

Letting $z=0$ in (8.4), we get

$$\begin{aligned}\limsup_{k\to\infty}\frac{\mathbb{P}\left(S_{\tau_+}>0, S_{\tau_+}\ne X_0^+\right)}{p_+} \;&=\; \limsup_{k\to\infty}\frac{K(0,\infty)}{p_+}\\ \le\; \limsup_{k\to\infty}&\frac{\int_0^\infty \varphi(y)(1+y)F(\mathrm{d}y)}{p_+}\\ \le\; \limsup_{k\to\infty}&\Big\{(1+a)\varphi(a) + 2\varphi(\infty)\frac{\int_a^\infty yF(\mathrm{d}y)}{p_+}\Big\}\\ =\; \limsup_{k\to\infty}&\,2\varphi(\infty)\frac{\int_a^\infty yF(\mathrm{d}y)}{p_+}.\end{aligned}$$

Letting $a \to \infty$, this converges to 0 according to (8.3), which easily yields the assertion. □

Proof of Theorem 8.2: Just note that $\mathbb{P}(M > 0) \;=\; \mathbb{P}(S_{\tau_+} > 0)$,

$$\mathbb{P}(M \neq S_{\tau_+} \,|\, M > 0) \;=\; \mathbb{P}\bigl(\tau_+(2) < \infty \,\big|\, \tau_+ < \infty\bigr) \;=\; \mathbb{P}(S_{\tau_+} > 0) \;\to\; 0$$

and appeal to Lemma 8.4. □

We next consider $GI/G/1$ queues. In view of $W \overset{\mathscr{D}}{=} M$, Theorem 8.2 states that W and $(U-T)^+$ are light traffic equivalent provided $X = U-T$ satisfies (8.3). It remains to carry out the relevant translation to conditions in terms of A, B, and to give some examples.

The first example is thinning of the arrival process where the results are in terms of $\Gamma(t) = \sum_1^\infty A^{*n}(t)$ (the renewal function except that the $n = 0$ term is not included).

Corollary 8.5 *Given a $GI/G/1$ queueing system specified in terms of U, T, define for each $k = 1, 2, \ldots$ another GI/G/1 system by thinning of the arrival process with retention probability $1/k$. That is, $T = T_0 + \cdots + T_{N-1}$ where N is independent of $T_0, T_1, \ldots$ with $\mathbb{P}(N = \ell) = (1 - 1/k)^{\ell-1}/k$, $\ell = 1, 2, \ldots$ Then W and $(U-T)^+$ are light traffic equivalent provided that $\mathbb{E}U^2 < \infty$. Writing $\Gamma = \sum_1^\infty A^{*n}$, one then has*

$$\mathbb{P}(W > 0) \;\sim\; p_+ \;=\; \mathbb{P}(U - T > 0) \;\sim\; \frac{1}{k}\mathbb{E}\Gamma(U). \tag{8.6}$$

Proof. Obviously,

$$\mathbb{P}(U - T > y) \;=\; \int_y^\infty \sum_{\ell=1}^\infty \frac{1}{k}(1 - 1/k)^{\ell-1} A^{*\ell}(u - y) B(\mathrm{d}u)$$

for $y > 0$ so that

$$k\mathbb{P}(U - T > y) \;\uparrow\; \int_y^\infty \Gamma(u - y) B(\mathrm{d}u),\; k \to \infty. \tag{8.7}$$

Taking $y = 0$ gives $p_+ \sim \mathbb{E}\Gamma(U)/k$. Further, by integration by parts we have

$$\int_a^\infty xF(\mathrm{d}x) \;=\; a\mathbb{P}(U - T > a) + \int_a^\infty \mathbb{P}(U - T > x)\,\mathrm{d}x. \tag{8.8}$$

We can bound $\Gamma(y)$ by $c(1 + y)$, and therefore an upper bound for (8.8) is

$$\begin{aligned} & \frac{c}{k}\Bigl[a\overline{B}(a) + a\mathbb{E}(U - a)^+ + \int_a^\infty \bigl\{\overline{B}(x) + \mathbb{E}(U - x)^+\bigr\}\mathrm{d}x\Bigr] \\ = \;& \frac{c}{k}\Bigl[a\overline{B}(a) + (1 + a)\mathbb{E}(U - a)^+ + \frac{1}{2}\mathbb{E}(U - a)^{+2}\Bigr]. \end{aligned}$$

Here $[\cdots] \to 0$ as $a \to \infty$ because of $\mathbb{E}U^2 < \infty$, and combining with $p_+ \sim \mathbb{E}\Gamma(U)/k$ shows that (8.3) holds. □

Next consider the scaling case $T = kT_*$ with $A_*(t) = \mathbb{P}(T_* \le t)$ independent of k.

Corollary 8.6 *Assume $T = kT_*$ where $\mathbb{P}(T_* \le t) \sim ct^\alpha$, $t \downarrow 0$. Then W and $(U-T)^+$ are light traffic equivalent provided that $\mathbb{E}U^{\alpha+1} < \infty$, and then*

$$\mathbb{P}(W>0) \ \sim\ p_+ \ =\ \mathbb{P}(U-T>0) \sim \frac{c}{k^\alpha}\mathbb{E}U^\alpha \tag{8.9}$$

Proof. For $y > 0$,

$$\mathbb{P}(U-kT_*>y) \ =\ \int_y^\infty A_*\Bigl(\frac{u-y}{k}-\Bigr)B(\mathrm{d}u).$$

Letting $y = 0$, we get

$$k^\alpha p_+ \ = \int_0^\infty k^\alpha A_*\Bigl(\frac{u}{k}-\Bigr)\,B(\mathrm{d}u) \ \to c\int_0^\infty u^\alpha B(\mathrm{d}u)$$

(using dominated convergence and $c_1 = \sup_t A_*(t-)/t^\alpha < \infty$). These estimates show also that an upper bound for (8.8) is

$$\begin{aligned}&\frac{c_1}{k^\alpha}\Bigl[a\mathbb{E}(U-a)^{+\alpha} + \int_a^\infty \mathbb{E}(U-x)^{+\alpha}\mathrm{d}x\Bigr]\\ &\qquad =\ \frac{c}{k}\Bigl[a\mathbb{E}(U-a)^{+\alpha} + \frac{1}{\alpha+1}\mathbb{E}(U-a)^{+\alpha+1}\Bigr].\end{aligned}$$

Here $[\cdots] \to 0$ as $a \to \infty$ because of $\mathbb{E}U^{\alpha+1} < \infty$, and combining with $p_+ \sim c\mathbb{E}U^\alpha/k^\alpha$ shows that (8.3) holds. □

Let $B^{(x)}$ denote the overshoot distribution, $\overline{B}^{(x)}(y) = \overline{B}(x+y)/\overline{B}(x)$.

Corollary 8.7 *Assume that there exists a distribution G with finite mean such that $B^{(x)}$ is stochastically dominated by G for all x. Then W and $(U-T)^+$ are light traffic equivalent.*

Proof. For $y > 0$,

$$\mathbb{P}(U-T>y) \ =\ p_+\mathbb{P}(U>T+y\,|\,U>T) \ \le\ p_+\overline{G}(y).$$

Hence an upper bound for (8.8) is

$$p_+\left[a\overline{G}(a) + \int_a^\infty \overline{G}(x)\,\mathrm{d}x\right].$$

Here $[\cdots] \to 0$ as $a \to \infty$ when $\mu_G = \int_0^\infty \overline{G}(x)\,\mathrm{d}x < \infty$, and therefore (8.3) holds. □

Remark 8.8 Intuitively, what are the reasons that delay occurs in light traffic? Two reasons come immediately to mind: short interarrival times (clustering) or long service times. To make such a study more rigorous, one way is to describe the conditional distribution of U, T given $X = U-T > 0$ For example, in the scaling case $T = kT_*$ in Corollary 8.6, one has in

$M/D/1$ that $U \equiv 1$ (being constant) is unchanged in this distribution, whereas the conditional distribution of T_* is that of T_* given $T_* \le 1/k$ (which is asymptotically the uniform distribution on $(0, 1/k)$) so that delay is caused by short interarrival times. If instead one considers $D/M/1$, $T = k$ is unchanged in the conditional distribution, whereas the conditional distribution of U is that of $U + k$ so that delay is caused by long service times. See further the Problems, which also contain an example (Problem 8.3) where it is necessary to have *both* long service times and short interarrival times if delay is to occur in light traffic. □

Problems

8.1 Show that in Corollary 8.6, one has

$$\mathbb{P}\left(U \le u, T_* \le t/k \mid U - kT_* > 0\right) \;\to\; \frac{\int_0^u (y \wedge t)^\alpha B(\mathrm{d}y)}{\int_0^\infty y^\alpha B(\mathrm{d}y)}, \quad 0 < t < u.$$

8.2 Show that if the service time U has a nondecreasing failure rate, then (8.3) holds.

8.3 Take $\mathbb{P}(U > u) = \mathrm{e}^{-u^2}$, $T = kT_*$, $\mathbb{P}(T_* \le t) = \mathrm{e}^{-1/\sqrt{t}}$. Show using Problem 8.2 that (8.3) holds, and that conditionally upon $U - kT_* > 0$, $U/k^{1/5} \xrightarrow{\mathbb{P}} K$, $k^{4/5}T_* \xrightarrow{\mathbb{P}} K$, $U - kT_* \xrightarrow{\mathbb{P}} 0$, where $K = 4^{-2/5}$ is the unique point where $\varphi(z) = z^{-1/2} + z^2$ attains it minimum.

Notes The study of light traffic goes back to Bloomfield and Cox (1972), but the first mathematically more substantial results are those of Daley and Rolski (1984, 1991). The present exposition follows Asmussen (1992b), who also gives further examples and conditions for $\mathbb{E}W^{(k)^p} \sim \mathbb{E}(U - T)^p$, $p > 0$, together with the corresponding asymptotics. See also Sigman (1992) for workloads.

Whitt (1989) suggests approximations in the whole range $\rho \in (0, 1)$ using interpolating between heavy traffic ($\rho \uparrow 1$) and light traffic ($\rho \downarrow 0$); the details involve the explicit solution of $M/M/1$. Further frequently studied topics in light traffic limit theory are Taylor expansions such as $\mathbb{E}W \approx a_1\rho + \cdots + a_n\rho^n$ and, of course, models beyond $GI/G/1$ such as networks. See e.g. Kovalenko (1995) and Baccelli and Schmidt (1996) for these and further subjects.

9 Heavy–Tailed Asymptotics

We now assume that the service time distribution B is heavy–tailed, more precisely that B is long–tailed (for all y, $\overline{B}(x - y)/\overline{B}(x) \to 1$ as $x \to \infty$) and that its stationary excess (integrated tail) distribution $B_0(x) = \int_0^x \overline{B}(y)\,\mathrm{d}y \,/\, \mu_B$ is in the class $\mathscr{S}$ of subexponential distributions (see A5 for these concepts). We will derive tail asymptotics first for the steady–state waiting time W and later, under the added regularity condition $B \in \mathscr{S}^*$ (see (A.5.3)), for the maximal waiting time in a busy cycle (the parallel results for light tails are given in XIII.5 and state that both tails decay with the same exponential rate). We assume throughout $\rho < 1$. The result on W is as follows:

Theorem 9.1 (a) *Consider a random walk such that* $\mu = \mathbb{E}X < 0$ *and that* $\overline{F}(x) \sim B(x)$, $x \to \infty$, *for some distribution* B *on* $(0, \infty)$ *which is long–tailed and satisfies* $B_0 \in \mathscr{S}$. *Then, writing* $\overline{F}_I(x) = \int_x^\infty \overline{F}(y)\,\mathrm{d}y$, *it holds that*

$$\mathbb{P}(M > x) \;\sim\; \frac{1}{|\mu|}\overline{F}_I(x), \quad x \to \infty; \tag{9.1}$$

(b) *for a* $GI/G/1$ *queue with* $\rho < 1$ *and the service time distribution* B *satisfying the assumptions of* (a),

$$\mathbb{P}(W > x) \;\sim\; \frac{\rho}{1-\rho}\overline{B}_0(x), \quad x \to \infty. \tag{9.2}$$

The proof uses the following lemma:

Lemma 9.2 *Let* $Y_1, Y_2, \ldots$ *be i.i.d. with common distribution* $G \in \mathscr{S}$ *and let* N *be an independent integer–valued r.v. with* $\mathbb{E}z^N < \infty$ *for some* $z > 1$. *Then* $\mathbb{P}(Y_1 + \cdots + Y_N > u) \sim \mathbb{E}N\,\overline{G}(u)$.

Proof. Recall from A5 that $\overline{G^{*n}}(u) \sim n\overline{G}(u)$, $u \to \infty$, and that for each $z > 1$ there is a $D < \infty$ such that $\overline{G^{*n}}(u) \le \overline{G}(u)Dz^n$ for all u. Therefore we can use dominated convergence with $\sum \mathbb{P}(N = n)\,Dz^n$ as majorant to obtain

$$\frac{\mathbb{P}(Y_1 + \cdots + Y_N > u)}{\overline{G}(u)} \;=\; \sum_{n=0}^{\infty} \mathbb{P}(N = n)\frac{\overline{G^{*n}}(u)}{\overline{G}(u)} \;\to\; \sum_{n=0}^{\infty} \mathbb{P}(N = n)\cdot n \;=\; \mathbb{E}N.$$

□

For the proof of Theorem 9.1, it is instructive to first consider the $M/G/1$ case where A is exponential with rate β. The Pollaczeck–Khinchine formula states that $W \overset{\mathscr{D}}{=} Y_1 + \cdots + Y_K$ where the Y_i have distribution B_0 and K is geometric with parameter ρ, $\mathbb{P}(K = k) = (1-\rho)\rho^k$. Since $\mathbb{E}K = \rho/(1-\rho)$ and $\mathbb{E}z^K < \infty$ whenever $\rho z < 1$, the result follows immediately from Lemma 9.2. The argument for the general random walk or $GI/G/1$ case is similar. In fact, we have a similar representation $M = Y_1 + \cdots + Y_K$ where K is the number of ladder steps and $Y_1, Y_2, \ldots$ are i.i.d. with common distribution $G = G_+/\|G_+\|$. The difficulty is that whereas K is still geometric, then the parameter $\theta = \|G_+\|$ is not explicit as for $M/G/1$, and also it is not a priori clear that the tail behaviour of G is the same as that of B_0.

Write $\overline{G}_+(x) = G_+(x, \infty) = \mathbb{P}(S_{\tau_+} > x,\, \tau_+ < \infty)$ and let μ_{G_-} be the mean of G_-, $U_- = \sum_0^\infty G_-^{*n}$.

Lemma 9.3 $\overline{G}_+(x) \sim \overline{F}_I(x)/|\mu_{G_-}|$, $x \to \infty$.

Proof. By (1.7),

$$\overline{G}_+(x) \;=\; \int_{-\infty}^{0} \overline{F}(x-y)\, U_-(dy).$$

The heuristics is now that the contribution from the interval $(-N,0]$ to the integral is $\mathrm{O}(\overline{F}(x))$ which by long–tailedness is $\mathrm{o}(\overline{F}_I(x))$, whereas for large y, $U_-(\mathrm{d}y)$ is close to Lebesgue measure on $(-\infty,0]$ normalized by $|\mu_{G_-}|$ so that we should have

$$\overline{G}_+(x) \;\sim\; \frac{1}{|\mu_{G_-}|}\int_{-\infty}^{0} \overline{F}(x-y)\,\mathrm{d}y \;=\; \frac{1}{|\mu_{G_-}|}\overline{F}_I(x)\,.$$

We now make this precise. If G_- is nonlattice, then by Blackwell's renewal theorem $U_-(-n-1,-n] \to 1/|\mu_{G_-}|$. In the lattice case, we can assume that the span is 1 and then the same conclusion holds since then $U_-(-n-1,-n]$ is just the probability of a renewal at $-n$.

Given ϵ, choose N such that $\overline{F}(n-1)/\overline{F}(n) \le 1+\epsilon$ for $n \ge N$ (this is possible since B is long–tailed, cf. A5.1(a)), and that $U_-(-n-1,-n] \le (1+\epsilon)/|\mu_{G_-}|$ for $n \ge N$. We then get

$$\begin{aligned}
&\overline{\lim_{x\to\infty}}\, \frac{\overline{G}_+(x)}{\overline{F}_I(x)} \\
&\quad\le\; \overline{\lim_{x\to\infty}} \int_{-N}^{0} \frac{\overline{F}(x-y)}{\overline{F}_I(x)}\, U_-(\mathrm{d}y) \;+\; \overline{\lim_{x\to\infty}} \int_{-\infty}^{-N} \frac{\overline{F}(x-y)}{\overline{F}_I(x)}\, U_-(\mathrm{d}y) \\
&\quad\le\; \overline{\lim_{x\to\infty}}\, \frac{\overline{F}(x)}{\overline{F}_I(x)}\, U_-(-N,0] \;+\; \overline{\lim_{x\to\infty}}\, \frac{1}{\overline{F}_I(x)} \sum_{n=N}^{\infty} \overline{F}(x+n) U_-(-n-1,-n] \\
&\quad\le\; 0 \;+\; \overline{\lim_{x\to\infty}}\, \frac{1}{\overline{F}_I(x)}\, \frac{1+\epsilon}{|\mu_{G_-}|} \sum_{n=N}^{\infty} \overline{F}(x+n) \\
&\quad\le\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\, \overline{\lim_{x\to\infty}}\, \frac{1}{\overline{F}_I(x)} \int_{N}^{\infty} \overline{F}(x+y)\,\mathrm{d}y \\
&\quad=\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\, \overline{\lim_{x\to\infty}}\, \frac{\overline{F}_I(x+N)}{\overline{F}_I(x)} \;=\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\,.
\end{aligned}$$

Here in the third step we used that $\overline{B}(x)/\overline{B}_0(x) \to 0$ (since B is long–tailed) and hence $\overline{F}(x)/\overline{F}_I(x) \to 0$, and in the last that $\overline{F}_I$ is asymptotically proportional to $B_0 \in \mathscr{S}$. Similarly,

$$\underline{\lim_{x\to\infty}}\, \frac{\overline{G}_+(x)}{\overline{F}_I(x)} \;\ge\; \frac{(1-\epsilon)^2}{|\mu_{G_-}|}\,.$$

Letting $\epsilon \downarrow 0$, the proof is complete. □

Proof of Theorem 9.1. We first show part (a). By Lemma 9.3, $\mathbb{P}(Y_i > x) \sim \overline{F}_I(x)/(\theta|\mu_{G_-}|)$. Hence using dominated convergence precisely as for

$M/G/1$, $M = Y_1 + \cdots + Y_K$ yields

$$\mathbb{P}(M > u) \ \sim \ \sum_{k=1}^{\infty}(1-\theta)\theta^k \, k \, \frac{\overline{F}_I(u)}{\theta|\mu_{G_-}|} \ = \ \frac{\overline{F}_I(u)}{(1-\theta)|\mu_{G_-}|} \, .$$

Now just observe that $(1-\theta)|\mu_{G_-}| = (1 - \|G_+\|)|\mu_{G_-}| = |\mu|$ by VIII.(2.1).

To get (b) from (a), just observe that

$$\frac{\overline{F}(x)}{\overline{B}(x)} \ = \ \int_0^{\infty} \frac{\overline{B}(x+y)}{\overline{B}(x)} \, A(\mathrm{d}y) \ \to \ \int_0^{\infty} 1 \cdot A(\mathrm{d}y) \ = \ 1$$

by dominated convergence. This implies $\overline{F}_I(x) \sim \mu_B \overline{B}_0(x)$ and, using $|\mu| = \mu_A - \mu_B$, that

$$\mathbb{P}(W > x) \ = \ \mathbb{P}(M > x) \ \sim \ \frac{\mu_B}{|\mu|}\overline{B}_0(x) \ = \ \frac{\rho}{1-\rho}\overline{B}_0(x) \, . \qquad \square$$

Now consider the cycle maximum. In the random walk case, we consider a reflected version (Lindley process) $\{W_n\}$ starting from $W_0 = 0$ and define the cycle σ as for $GI/G/1$,

$$\sigma \ = \ \inf\{n \geq 1 : \, W_n = 0\} \ = \ \tau_- \ = \ \inf\{n \geq 1 : \, S_n \leq 0\} \, .$$

The cycle maximum is

$$M_\sigma \ = \ \max_{0 \leq n < \sigma} S_n \ = \ \max_{0 \leq n < \sigma} W_n .$$

Its relevance for extreme value theory has been explained in VI.4, and in fact, VI.4.10 and the following result immediately show that $\max_{0 \leq k \leq n} W_n$ after a suitable normalization has a Fréchet limit distribution as $n \to \infty$ when B is regularly varying (analogously Problem VI.4.1 gives a Gumbel limit when B is heavy–tailed Weibull; it is straighforward to adapt the argument to see that the same is the case for the log–normal distribution).

Theorem 9.4 *Consider a reflected random walk (Lindley process) $\{W_n\}$ such that $\mu = \mathbb{E}X < 0$ and that $\overline{F}(x) \sim B(x)$, $x \to \infty$, for some $B \in \mathscr{S}^*$. Then*

$$\mathbb{P}(M_\sigma > x) \ \sim \ \mathbb{E}\sigma\overline{F}(x), \quad x \to \infty. \tag{9.3}$$

The same conclusion holds for the $GI/G/1$ waiting time when the service time distribution B satisfies $B \in \mathscr{S}^*$.

For the proof, we first introduce some notation. Define

$$\begin{aligned} N_1(x, x_0) \ &= \ \#\big\{n < \sigma : \, S_n \leq x_0, S_{n+1} > x\big\}, \\ p_1(x, x_0) \ &= \ \mathbb{P}\big(S_{n+1} > x \text{ for some } n < \sigma \text{ with } S_n \leq x_0\big), \\ p_2(x, x_0) \ &= \ \mathbb{P}\big(\tau(x) < \sigma, x_0 \leq S_{\tau(x)-1} \leq x\big) \ . \end{aligned}$$

where $\tau(x) = \inf\{n \geq 1 : \, S_n > x\}$ (note that the definitions of $p_1(x, x_0)$ and $p_2(x, x_0)$ are not symmetric in the sets $[0, x_0]$ and (x_0, ∞)). Then

$$p_1(x, x_0) \ \leq \ \mathbb{P}(M_\sigma > x) \ \leq \ p_1(x, x_0) + p_2(x, x_0) \ . \tag{9.4}$$

Lemma 9.5 $\mathbb{E}N_1(x,x_0) \;\sim\; \mathbb{E}\sigma\mathbb{P}(M \le x_0)\overline{F}(x)$.

Proof. Define $C(A) = \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \in A) = \mathbb{E}\sigma\mathbb{P}(W \in A)$. We get

$$\begin{aligned}\mathbb{E}N_1(x,x_0) &= \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x_0, S_{n+1} > x) \;=\; \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x_0)\overline{F}(x - S_n) \\ &= \int_0^{x_0} \overline{F}(x-y)C(\mathrm{d}y) \;=\; \mathbb{E}\sigma\int_0^{x_0} \overline{F}(x-y)\mathbb{P}(W \in \mathrm{d}y)\;.\end{aligned}$$

Now just divide by $\overline{F}(x)$ and use $\overline{F}(x-y)/\overline{F}(x) \to 1$ uniformly in $0 \le y \le x_0$, as follows from $1 \le \overline{F}(x-y)/\overline{F}(x) \le \overline{F}(x-x_0)/\overline{F}(x) \to 1$. □

Lemma 9.6 $p_1(x,x_0) \;\sim\; \mathbb{E}\sigma\mathbb{P}(W \le x_0)\overline{F}(x)$.

Proof. After $\tau(x)$, the expected time $\{S_n\}$ spends in $(0,x_0)$ before hitting $(-\infty,0]$ is bounded by $a_1 + a_2x_0$. Hence with $\alpha(x,x_0) = (a_1+a_2x_0)\overline{F}(x-x_0)$, we have

$$\begin{aligned}\mathbb{P}\bigl(N_1(x,x_0) \ge k+1 \,\big|\, N_1(x,x_0) \ge k\bigr) &\le \alpha(x,x_0), \\ \mathbb{P}\bigl(N_1(x,x_0) \ge k+1\bigr) &\le p_1(x,x_0)\alpha(x,x_0)^k, \\ \mathbb{E}\bigl[N_1(x,x_0);\, N_1(x,x_0) \ge 2\bigr] &\le \frac{p_1(x,x_0)\alpha(x,x_0)}{1-\alpha(x,x_0)}\;, \\ p_1(x,x_0) \;\le\; \mathbb{E}N_1(x,x_0) &\le p_1(x,x_0) + \frac{p_1(x,x_0)\alpha(x,x_0)}{1-\alpha(x,x_0)}.\end{aligned}$$

Now just note that $\alpha(x,x_0) \to 0$ and use Lemma 9.5. □

Letting first $x \to \infty$ and next $x_0 \to \infty$ in (9.4), the following estimate will complete the proof of Theorem 9.4:

Lemma 9.7 $\displaystyle\lim_{x_0\to\infty}\limsup_{x\to\infty}\frac{p_2(x,x_0)}{\overline{F}(x)} \;=\; 0.$

The proof is based upon a downcrossing argument. Define $m_+ = \mathbb{E}X_+$, $m_- = \mathbb{E}X_-$ (thus $m = -\mu = m_- - m_+$) and

$$\begin{aligned}D_\sigma(x) &= \mathbb{E}\sum_{n=0}^{\sigma-1} I\bigl(S_n > x, S_{n+1} \le x\bigr), \\ D(x) &= \mathbb{E}\sum_{n=0}^{\infty} I\bigl(S_n > -x, S_{n+1} \le -x\bigr).\end{aligned}$$

Lemma 9.8 $\displaystyle\lim_{x\to\infty} D(x) \;=\; \frac{m_-}{m}.$

Proof. Let U denote the occupation (renewal) measure of the random walk, $U(A) = \sum_0^\infty \mathbb{P}(S_n \in A)$. Then (Problem VIII.3.4) $U[x,x+z] \le a_1 + a_2z$ for all x,z and has limit z/m as $z \to -\infty$ in the nonlattice case (that is, $U(\mathrm{d}z - x)$ converges vaguely to Lebesgue measure normalized by m).

Similar estimates as in the proof of the key renewal theorem then yield

$$\begin{aligned} D(x) &= \int_{-x}^{\infty} U(\mathrm{d}y)F(-x-y) \;=\; \int_{0}^{\infty} U(\mathrm{d}z-x)F(-z) \\ &\to \frac{1}{m}\int_0^{\infty} F(-z)\mathrm{d}z \;=\; \frac{m_-}{m}\,. \end{aligned}$$

The lattice case is similar though easier. □

Proof of Lemma 9.7. By regenerative process theory,

$$\frac{D_\sigma(x)}{\mathbb{E}\sigma} \;=\; \lim_{n\to\infty} \mathbb{P}(W_n > x, W_{n+1} \le x) \;=\; \int_x^{\infty} \mathbb{P}(W \in \mathrm{d}y)F(x-y).$$

Theorem 9.1 makes it plausible that we can replace $\mathbb{P}(W \in \mathrm{d}y)$ by $m^{-1}\overline{F}(y)\,\mathrm{d}y$; for the rigorous proof which indeed uses $B \in \mathscr{S}^*$ in an essential way, see Asmussen *et al.* (2002). We then get

$$\begin{aligned} \frac{D_\sigma(x)}{\mathbb{E}\sigma} &\sim \frac{1}{m}\int_x^{\infty} \overline{F}(y)\mathrm{d}y \int_{-\infty}^{x-y} F(\mathrm{d}z) \\ &= \frac{\overline{F}(x)}{m}\int_{-\infty}^{0} F(\mathrm{d}z)\int_x^{x-z} \frac{\overline{F}(y)}{\overline{F}(x)}\mathrm{d}y \\ &\sim \frac{\overline{F}(x)}{m}\int_{-\infty}^{0} |z|F(\mathrm{d}z) \;=\; \overline{F}(x)\frac{m_-}{m}\,, \end{aligned}$$

where the third step is an easy consequence of long–tailedness.

On the other hand, the overshoot over x after an upcrossing from a level $\le x_0$ converges in distribution to ∞ by long–tailedness, so that the expected subsequent number of downcrossings of level x before $[0, x_0]$ is hit is approximately m_-/m by Lemma 9.8. Hence we get

$$\begin{aligned} \mathbb{E}\sigma\overline{F}(x)\frac{m_-}{m} &\sim D_\sigma(x) \;\ge\; \mathbb{E}N_1(x,x_0)\frac{m_-}{m} + p_2(x,x_0) \\ &\sim \mathbb{E}\sigma\overline{F}(x)\mathbb{P}(M \le x_0)\frac{m_-}{m} + p_2(x,x_0)\,, \\ \limsup_{x\to\infty} \frac{p_2(x,x_0)}{\overline{F}(x)} &\le \mathbb{E}\sigma\mathbb{P}(M > x_0)\frac{m_-}{m}\,. \end{aligned}$$

Let $x_0 \uparrow \infty$. □

Notes Theorem 9.1 has a long history associated with the names of (in alphabetical order) von Bahr, Borovkov, Cohen, Pakes and Veraverbeke. These contributions are given a final form in Embrechts and Veraverbeke (1982). There are numerous recent analogues for more general models, e.g. Whitt (2001) and Boxma *et al.* (2002) for many–server queues, Heath *et al.* (1999), Jelenkovic and Momcilovic (2001) and Zwart *et al.* (2003) for fluid queues, and Baccelli *et al.* (1999) and Baccelli and Foss (2003) for (feed-forward) networks. Also tail asymptotics for the busy period has received considerable attention, see Baltrunas *et* *al.* (2002) and references therein. For other queue disciplines than FIFO, see the Notes to III.9.

For some remarkable explicit waiting–time distributions in $M/G/1$ with heavy tails, see Abate and Whitt (1999).

Theorem 9.4 was given independently by Samorodnitsky *et al.* (1997), assuming regular variation, and Asmussen (1998a); the latter paper used the "plausible" step in the proof of Lemma 9.7, which was only recently justified by Asmussen *et al.* (2002; in connection with the results of that paper, see also Bertoin and Doney, 1994b, and Asmussen *et al.*, 2003).

A current trend in the literature related to stressing the importance of heavy tails is the study of *long–range dependence* (LRD). In a stationary process setting, this means that the dependence between X_0 and X_t decays slowly; a common precise definition is that $|\mathbb{C}ov(X_0, X_t)|$ is not integrable (note that this is a necessary condition for a CLT for $\int_0^T X_t \, \mathrm{d}t$ with variance constant proportional to $\sqrt{T}$; cf. the Notes to VI.3). Again, statistical studies are taken as the main motivation, but they are far from uncontroversial; see Mikosch and Stărică (2003). LRD is related to *self–similarity*, i.e. the existence of a constant H (the *Hurst parameter*) such that $\left\{c^{-H} X_{tc}\right\}_{t \geq 0} \overset{\mathscr{D}}{=} \{X_t\}_{t \geq 0}$. The volumes edited by Park and Willinger (2000) and Taqqu *et al.* (2002) may be taken as a starting point for the area. A main example is *fractional Brownian motion* (FBM), a certain Gaussian process with stationary long–range dependent increments; see e.g. Massoulie and Simonian (1999), Norros (2000) and Piterbarg (2001).

The simplest result pointing to the connection between heavy tails and LRD is covariance asymptotics for renewal processes (Daley, 1999). Another simple case is alternating renewal processes where in the notation of VI.2b one of F_0, F_1 is heavy–tailed; this is in turn relevant for fluid models involving on–off sources with heavy–tailed on periods. See Heath *et al.* (1998, 1999).

XI
Markov Additive Models

1 Some Basic Examples

1a. Markovian Point Processes
1b. Markovian Fluids
1c. The $MArP/PH/1$ Workload

1a Markovian Point Processes

In applied probability, the Poisson process has served as the main point process model for many years and generalizations have concentrated on the renewal process. However, it is very seldom that the renewal process can be given a similar intuitive interpretation as the Poisson process (as a binomial limit). Further, the class of renewal processes is not particularly flexible, and in particular, arrivals that tend to occur in bursts cannot be modelled in this way.

A popular model for bursty arrivals is the Markov–modulated Poisson process. This is defined in terms of a background Markov process $\{J_t\}$ with $p < \infty$ states, such that the arrival intensity is β_i on time intervals where $J_t = i$. See Fig. 1.1 for an illustration (here we have $p = 2$ states with β_1 much smaller than β_2).

We shall here study a generalization of the Markov–modulated Poisson process, commonly denoted as the *Markovian arrival process* ($MarP$);[1]

[1] The literature on this process uses the abbreviation MAP. However, this is unfortunate since it is also standard for Markov additive process.

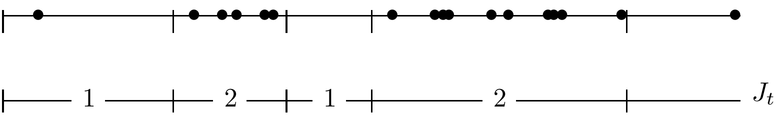

Figure 1.1

it incorporates further examples such as phase–type renewal processes and semi–Markov point processes with phase–type interarrival times. The definition is as for the Markov–modulated Poisson process, except that there may be some additional arrivals when $\{J_t\}$ changes state: w.p. q_{ij} an arrival occurs at a jump from i to $j \neq i$.

The *MArP* extends the Poisson process in much the same way as phase–type distributions extend the exponential distribution. In particular, the *MArP* has the feature of making many analytic properties explicit or at least computationally tractable, and *MArP*'s are dense (w.r.t. the standard topology for weak convergence) in the space of point processes on $(0, \infty)$.

For a more detailed treatment, we use a slightly different formalism, corresponding to the decomposition $\boldsymbol{\Lambda} = \boldsymbol{C} + \boldsymbol{D}$ of the intensity matrix $\boldsymbol{\Lambda}$ of the background Markov process $\{J_t\}$, where $\boldsymbol{D}$ gives the "intensities of state changes with arrivals", and $\boldsymbol{C}$ those of "state changes without arrivals". That is,

$$d_{ij} = \begin{cases} \beta_i & i = j \\ \lambda_{ij} q_{ij} & i \neq j \end{cases}, \quad c_{ij} = \begin{cases} -\sum_{k \neq i} c_{ik} - \sum_{k=1}^{p} d_{ik} & i = j \\ \lambda_{ij}(1 - q_{ij}) & i \neq j \end{cases}.$$

Note that a "state change with an arrival" may be dummy, corresponding to a transition $i \to i$; e.g. this is the case for all arrivals in the Markov–modulated Poisson process that corresponds to $\boldsymbol{D} = (\beta_i)_{\text{diag}}$, $\boldsymbol{C} = \boldsymbol{\Lambda} - (\beta_i)_{\text{diag}}$.

In addition to the matrices $\boldsymbol{C}, \boldsymbol{D}$, the complete specification of a *MArP* also requires specification of the distribution of J_0. We do this in terms of the row vector $\boldsymbol{\alpha}$ with ith element $\alpha_i = \mathbb{P}(J_0 = i)$. The counting process of the *MArP* is denoted by $\{N_t\}$ (N_t is the number of arrivals in $[0, t]$).

Example 1.1 (PHT RENEWAL PROCESSES) Consider a renewal process with interarrival distribution F that is of phase type with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ (the corresponding exit rate vector is $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{1}$). Piecing the phase processes for individual interarrival times together as in III.5, we obtain a background Markov process $\{J_t\}$ with intensity matrix $\boldsymbol{\Lambda} = \boldsymbol{T} + \boldsymbol{t}\boldsymbol{\alpha}$. Obviously, $\boldsymbol{T}$ correspond to state changes without arrivals and $\boldsymbol{t}\boldsymbol{\alpha}$ to state changes with arrivals. Thus, we have a *MArP* with the same $\boldsymbol{\alpha}$ and $\boldsymbol{C} = \boldsymbol{T}$, $\boldsymbol{D} = \boldsymbol{t}\boldsymbol{\alpha}$. □

We now turn to the general theory of the *MArP*, first Palm theory (cf. VII.6; as the accompanying process $\{X_t\}$ there we may just use $\{J_t\}$). From the Markovian interpretation, we immediately have:

Proposition 1.2 *A MArP with parameters $(\boldsymbol{\alpha}, \boldsymbol{C}, \boldsymbol{D})$ is time–stationary provided $\boldsymbol{\alpha} = \boldsymbol{\pi}$ where $\boldsymbol{\pi}$ is a stationary vector for $\boldsymbol{\Lambda} = \boldsymbol{C} + \boldsymbol{D}$, i.e. $\boldsymbol{\pi\Lambda} = \boldsymbol{0}$, $\boldsymbol{\pi 1} = 1$. If $\{J_t\}$ is ergodic, then $\boldsymbol{\alpha} = \boldsymbol{\pi}$ is also necessary for time–stationarity.*

Corollary 1.3 *For a time–stationary MArP, the intensity $\lambda = \mathbb{E}_{\boldsymbol{\pi}} N_t/t$ is given by $\lambda = \boldsymbol{\pi D 1}$.*

Proof. The intensity of an arrival when $J_t = i$ is $\sum_{j=1}^{p} d_{ij}$ so that $\lambda = \sum_{i=1}^{p} \pi_i \sum_{j=1}^{p} d_{ij} = \boldsymbol{\pi D 1}$. □

Proposition 1.4 *A MArP with parameters $(\boldsymbol{\alpha}, \boldsymbol{C}, \boldsymbol{D})$ and having an arrival at $t = 0$ is event–stationary provided $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ where $\boldsymbol{\alpha}_0 = \boldsymbol{\pi D}/\boldsymbol{\pi D 1}$.*

Proof. Let Y_n be the value of $\{J_t\}$ just after the nth arrival. Since $\{J_t\}$ moves according to $\boldsymbol{C}$ between arrivals and generates arrivals (and states just after) according to $\boldsymbol{D}$, the transition matrix of $\{Y_n\}$ is

$$\boldsymbol{P} \;=\; \int_0^\infty \mathrm{e}^{\boldsymbol{C}t} \boldsymbol{D}\, \mathrm{d}t \;=\; -\boldsymbol{C}^{-1}\boldsymbol{D}.$$

From $\boldsymbol{\pi}(\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{0}$ we have $\boldsymbol{\pi D} = -\boldsymbol{\pi C}$ and hence

$$\boldsymbol{\alpha}_0 \boldsymbol{P} \;=\; \frac{1}{\boldsymbol{\pi D 1}}(-\boldsymbol{\pi C})(-\boldsymbol{C}^{-1}\boldsymbol{D}) \;=\; \frac{\boldsymbol{\pi D}}{\boldsymbol{\pi D 1}} \;=\; \boldsymbol{\alpha}_0.$$

Since clearly $\boldsymbol{\alpha}_0 \boldsymbol{1} = 1$, $\boldsymbol{\alpha}_0$ is therefore stationary for $\boldsymbol{P}$, showing the assertion. □

Proposition 1.5 *For a MArP with parameters $(\boldsymbol{\alpha}, \boldsymbol{C}, \boldsymbol{D})$, the joint density at $x_1, \ldots, x_n$ of the first n interarrival times is*

$$\boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{C}x_1} \boldsymbol{D} \mathrm{e}^{\boldsymbol{C}x_2} \boldsymbol{D} \cdots \boldsymbol{D} \mathrm{e}^{\boldsymbol{C}x_n} \boldsymbol{D 1}\,.$$

Proof. Consider the event that $J_0 = i$, the first interarrival time is x_1, that $J_{x_1-} = j$, $J_{x_1} = k$, and that the second interarrival time is x_2 and $J_{x_2-} = \ell$, $J_{x_2} = m$. The probability of this (in the density sense) is

$$\alpha_i \boldsymbol{1}_i' \mathrm{e}^{\boldsymbol{C}x_1} \boldsymbol{1}_j \, d_{jk} \, \boldsymbol{1}_k' \mathrm{e}^{\boldsymbol{C}x_2} \boldsymbol{1}_\ell \, d_{\ell m}$$

Summing over i, j, k, ℓ, m gives the result for $n = 2$, and the case $n > 2$ is similar. □

Proposition 1.6 *The matrix $\widehat{\boldsymbol{F}}_t[z]$ with ijth element $\mathbb{E}_i[z^{N_t}; \, J_t = j]$ is given by $\widehat{\boldsymbol{F}}_t[z] = \mathrm{e}^{t(\boldsymbol{C}+z\boldsymbol{D})}$. In particular, $\mathbb{E}_{\boldsymbol{\alpha}} z^{N(t)} = \boldsymbol{\alpha} \mathrm{e}^{t(\boldsymbol{C}+z\boldsymbol{D})} \boldsymbol{1}$.*

Proof. See Proposition 2.2 of the next section. □

Now consider more detailed moment expansions than in Corollary 1.3. Define $\boldsymbol{M}(t)$ as the matrix with ijth element $\mathbb{E}_i[N_t; \, J_t = j]$ and $\boldsymbol{M}_2(t)$ as the matrix with ijth element $\mathbb{E}_i[N_t(N_t - 1); \, J_t = j]$.

Proposition 1.7

$$\boldsymbol{M}(t) \;=\; \int_0^t \mathrm{e}^{\boldsymbol{\Lambda}x}\boldsymbol{D}\mathrm{e}^{\boldsymbol{\Lambda}(t-x)}\,\mathrm{d}x\,, \tag{1.1}$$

$$\boldsymbol{M}_2(t) \;=\; 2\int_0^t\int_0^x \mathrm{e}^{\boldsymbol{\Lambda}x}\boldsymbol{D}\mathrm{e}^{\boldsymbol{\Lambda}(x-y)}\,\mathrm{d}y\;\boldsymbol{D}\mathrm{e}^{\boldsymbol{\Lambda}(t-x)}\,\mathrm{d}x\,. \tag{1.2}$$

Proof. Obviously $\boldsymbol{M}(0) = \boldsymbol{M}_2(0) = \boldsymbol{0}$. Further,

$$\begin{aligned}\boldsymbol{M}'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{F}}_t[z]\Big|_{z=1} \;=\; \frac{\mathrm{d}}{\mathrm{d}z}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{e}^{t(\boldsymbol{C}+z\boldsymbol{D})}\Big|_{z=1}\\ &= \frac{\mathrm{d}}{\mathrm{d}z}\mathrm{e}^{t(\boldsymbol{C}+z\boldsymbol{D})}(\boldsymbol{C}+z\boldsymbol{D})\Big|_{z=1} \;=\; \boldsymbol{M}(t)\boldsymbol{\Lambda} + \mathrm{e}^{\boldsymbol{\Lambda}t}\boldsymbol{D},\\ \boldsymbol{M}_2'(t) &= \frac{\mathrm{d}^2}{\mathrm{d}z^2}e^{t(\boldsymbol{C}+z\boldsymbol{D})}(\boldsymbol{C}+z\boldsymbol{D})\Big|_{z=1} \;=\; \boldsymbol{M}_2(t)\boldsymbol{\Lambda} + 2\boldsymbol{M}(t)\boldsymbol{D}\,.\end{aligned}$$

Now just check that the r.h.s.'s of (1.1), (1.2) satisfies the same boundary conditions (obvious) and the same differential equations; to this end, use repeatedly the rule

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t f(t,x)\,\mathrm{d}x \;=\; f(t,t) + \int_0^t \frac{\mathrm{d}}{\mathrm{d}t}f(t,x)\,\mathrm{d}x\,. \qquad \square$$

We next show that $\boldsymbol{M}(t)$ has a linear asymptote. That is, $\boldsymbol{M}(t) = t\boldsymbol{A} + \boldsymbol{B} + \mathrm{o}(1)$, and we identify $\boldsymbol{A}, \boldsymbol{B}$ (not surprisingly, $\boldsymbol{A}$ involves λ, cf. Proposition 1.3). Define $\boldsymbol{\Lambda}^- = (\boldsymbol{\Lambda} - \boldsymbol{1\pi})^{-1}$; $\boldsymbol{\Lambda}^-$ is a generalized inverse of $\boldsymbol{Q}$ in the sense that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^-\boldsymbol{\Lambda} = \boldsymbol{\Lambda}$ and we have (cf. II.4.9)

$$\mathrm{e}^{\boldsymbol{\Lambda}t} \;=\; \boldsymbol{1\pi} + \mathrm{O}(t^r\mathrm{e}^{-bt}) \tag{1.3}$$

for some $b > 0$ and some integer r (in many examples, $r = 0$), and

$$\int_0^t \mathrm{e}^{\boldsymbol{\Lambda}x}\,\mathrm{d}x \;=\; t\boldsymbol{1\pi} + \boldsymbol{\Lambda}^-(\mathrm{e}^{\boldsymbol{\Lambda}t} - \boldsymbol{I}) \;=\; t\boldsymbol{1\pi} - \boldsymbol{\Lambda}^- + \mathrm{O}(\mathrm{e}^{-bt}), \tag{1.4}$$

$$\int_0^t x\mathrm{e}^{\boldsymbol{\Lambda}x}\,\mathrm{d}x \;=\; \frac{t^2}{2}\boldsymbol{1\pi} + t\left[\boldsymbol{1\pi} + \boldsymbol{\Lambda}^-(\mathrm{e}^{\boldsymbol{\Lambda}t} - \boldsymbol{I})\right] - \boldsymbol{\Lambda}^{-^2}(\mathrm{e}^{\boldsymbol{\Lambda}t} - \boldsymbol{I}) \tag{1.5}$$

$$\;=\; \frac{t^2}{2}\boldsymbol{1\pi} + t\left[\boldsymbol{1\pi} - \boldsymbol{\Lambda}^-\right] + \boldsymbol{\Lambda}^{-^2} + \mathrm{O}(t^{r+1}\mathrm{e}^{-bt}). \tag{1.6}$$

Proposition 1.8 $\boldsymbol{M}(t) \;=\; t\lambda\boldsymbol{1\pi} + \boldsymbol{d\pi} + \boldsymbol{1c} - 2\lambda\boldsymbol{1\pi} + \mathrm{O}(t^{2r+1}\mathrm{e}^{-bt})$ *where* $\boldsymbol{d} = -\boldsymbol{\Lambda}^-\boldsymbol{D1}$, $\boldsymbol{c} = -\boldsymbol{\pi D}\boldsymbol{\Lambda}^-$.

Proof. Write $\boldsymbol{M}(t) = t\lambda\boldsymbol{1\pi} + I_1(t) + I_2(t) + I_3(t)$ where

$$\begin{aligned}I_1(t) &= \int_0^t (\mathrm{e}^{\boldsymbol{\Lambda}x} - \boldsymbol{1\pi})\boldsymbol{D}\big(\mathrm{e}^{\boldsymbol{\Lambda}(t-x)} - \boldsymbol{1\pi}\big)\,\mathrm{d}x\,,\\ I_2(t) &= \int_0^t \boldsymbol{1\pi D}\big(\mathrm{e}^{\boldsymbol{\Lambda}(x-t)} - \boldsymbol{1\pi}\big)\,\mathrm{d}x\,,\\ I_3(t) &= \int_0^t \big(\mathrm{e}^{\boldsymbol{\Lambda}x} - \boldsymbol{1\pi}\big)\boldsymbol{D1\pi}\,\mathrm{d}x\,.\end{aligned}$$

By (1.3),

$$I_1(t) \;=\; \int_0^t \mathrm{O}(x^r \mathrm{e}^{-bx})\, \boldsymbol{D}\, \mathrm{O}\bigl((t-x)^r \mathrm{e}^{-b(t-x)}\bigr)\, \mathrm{d}x \;=\; \mathrm{O}(t^{2r+1}\mathrm{e}^{-bt})\,.$$

Using (1.4) yields

$$\begin{aligned} I_2(t) &= \boldsymbol{1\pi D}(\mathrm{e}^{\boldsymbol{\Lambda} t} - \boldsymbol{I})\boldsymbol{\Lambda}^- \;=\; \boldsymbol{1\pi D}(\boldsymbol{1\pi} - \boldsymbol{I})\boldsymbol{\Lambda}^- + \mathrm{O}(t^r\mathrm{e}^{-bt}) \\ &= \lambda\boldsymbol{1\pi\Lambda}^- - \boldsymbol{1\pi D\Lambda}^- + \mathrm{O}(t^r\mathrm{e}^{-bt}) \;=\; -\lambda\boldsymbol{1\pi} + \boldsymbol{1c} + \mathrm{O}(t^r\mathrm{e}^{-bt}), \\ I_3(t) &= \boldsymbol{\Lambda}^-(\mathrm{e}^{\boldsymbol{\Lambda} t} - \boldsymbol{I})\boldsymbol{D1\pi} \;=\; \boldsymbol{\Lambda}^-(\boldsymbol{1\pi} - \boldsymbol{I})\boldsymbol{D1\pi} + \mathrm{O}(t^r\mathrm{e}^{-bt}) \\ &= \lambda\boldsymbol{\Lambda}^-\boldsymbol{1\pi} - \boldsymbol{\Lambda}^-\boldsymbol{D1\pi} + \mathrm{O}(t^r\mathrm{e}^{-bt}) \;=\; -\lambda\boldsymbol{1\pi} + \boldsymbol{d\pi} + \mathrm{O}(t^r\mathrm{e}^{-bt}). \end{aligned}$$

Collecting terms, the result follows. □

By similar but more lengthy calculations, one can show that $\boldsymbol{M}_2(t) = \lambda^2 t^2 \boldsymbol{1\pi} + t\boldsymbol{A}_2 + \boldsymbol{B}_2 + \mathrm{o}(1)$. The expressions for $\boldsymbol{A}_2, \boldsymbol{B}_2$ are quite complicated. We shall quote the details only for the stationary situation.

Proposition 1.9 $\mathbb{V}ar_{\boldsymbol{\pi}}\, N_t \;=\; t\bigl\{\lambda - 2\lambda^2 + 2\boldsymbol{cD1}\bigr\} \;+\; 2\boldsymbol{c}\left(\mathrm{e}^{\boldsymbol{Q}x} - \boldsymbol{I}\right)\boldsymbol{d} = t\bigl\{\lambda - 2\lambda^2 + 2\boldsymbol{cD1}\bigr\} \;+\; 2\lambda^2 - 2\boldsymbol{cd} + \mathrm{O}(t^{3r+2}\mathrm{e}^{-bt})\,.$

Problems

1.1 Show that $\mathbb{V}ar\, N_t \ge \mathbb{E}N_t$ in the Markov–modulated Poisson processes but not necessarily for a *MArP*.

1.2 Show that the counting process $\{N_t\}$ may be stationary even if $\{J_t\}$ is not so. [*Hint:* Take the row sums of $\boldsymbol{D}$ equal.]

1.3 (PHASE–TYPE SEMI–MARKOV POINT PROCESSES) Let $F_1, \ldots, F_q$ be phase–type distributions with representation $(\boldsymbol{\alpha}^{(j)}, \boldsymbol{T}^{(j)})$ for F_j. Assume further that we have a background Markov chain $\{Y_n\}$ with q states and transition probabilities p_{jk}, such that the semi–Markov point process has nth interarrival time governed by F_j when $Y_n = j$. Show by piecing together the phase processes that the counting process is a *MArP* and write up $\boldsymbol{C}$ and $\boldsymbol{D}$.

Notes The seminal paper on the *MArP* is Neuts (1979), though certainly the model has roots in earlier work such as Hermann (1965) and Rudemo (1973). Neuts (1992) contains an extensive bibliography.

For denseness, see Hermann (1965) and Asmussen and Koole (1993). For moment calculations and more detail concerning Proposition 1.9, see Naryana and Neuts (1992). Statistical aspects are surveyed in Rydén (2000).

1b Markovian Fluids

Let $\{J_t\}_{t\ge 0}$ be a Markov process with a finite state space E and $r = \bigl(r(i)\bigr)_{i\in E}$ a function on E. We define $\{S_t\}_{t\ge 0}$ by $S_0 = 0$, $\dot{S}_t = r(J_t)$, or, equivalently, $S_t = \int_0^t r(J_s)\,\mathrm{d}s$; $\{S_t\}_{t\ge 0}$ is called a (unrestricted) *Markovian fluid model.* See Fig. 1.2 for an example of a sample path in the case $E = \{1,2,3\}$, $r(2) < r(3) < 0 < r(1)$. The restricted process obtained by modifying $\{S_t\}$ to have reflection at 0 (cf. IX.2) is denoted by $\{V_t\}$ and is of considerable interest in many applications.

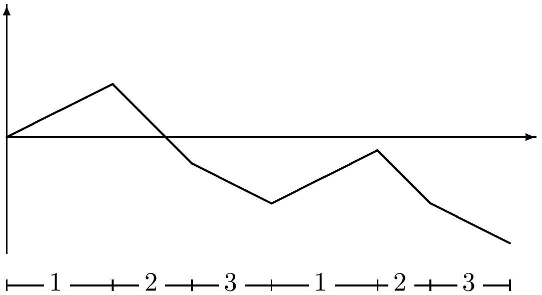

Figure 1.2

Example 1.10 In recent ATM (Asynchronous Transfer Mode) technology, K channels (sources) feed digital signals in an ATM buffer, which then outputs the signals in a shared single channel (the sink). Assume that each channel can be either in an ON or an OFF mode with exponential holding times μ, λ and transmits at a rate of a bits per unit time in the ON mode, and that the maximal capacity of the sink is b bits per unit time. Let J_t denote the number of channels in the ON mode at time t. Then $\{J_t\}$ is a birth–death process on $E = \{0, \ldots, K\}$ with birth intensities $\beta_n = (K-n)\lambda$, $n < K$, and death intensities $\delta_n = n\mu$, $n > 0$. The netput (input minus maximal output) then changes the buffer content by $na - b$ bits per unit time on intervals where $J_t = n$, and since a, b are usually enormous compared to the time between jumps of $\{J_t\}$, we can model the netput as an unrestricted fluid $\{S_t\}$ with $r(n) = na - b$. The restricted process $\{V_t\}$ is then the buffer content (assuming an infinite buffer).

The model can be considerably generalized without violating the Markov property. For example, by suitably extending $\{J_t\}$, one can allow the ON–OFF times to be phase type rather than exponential, the sources to be different, a Markov–modulated environment (λ depending on J_t), etc. □

Example 1.11 Let J_t denote the number of the K channels that are busy in Erlang's loss system III.3e. Then $\{J_t\}$ is a birth–death process on $E = \{0, \ldots, K\}$ with birth intensities $\beta_n = \lambda$, $n < K$, and death intensities $\delta_n = n\mu$, $n > 0$. Assume that in addition to ordinary telephone traffic the system also carries data traffic of lower priority. Using a deterministic fluid approximation as in Example 1.10, there exist constants a, b such that data is received at rate a and processed at rate b per idle channel. The content of unprocessed data traffic is therefore a reflected fluid corresponding to $r(n) = a - (K-n)b$. □

Problems

1.4 Write up the intensity matrix for $\{J_t\}$ and the $r(i)$ in some of the models suggested in the last paragraph of Example 1.10.

Notes Two seminal early papers on Markovian fluids are Anick *et al.* (1982) and Gaver and Lehocky (1982). The literature is by now extensive.

1c The $MArP/PH/1$ Workload

Consider a queue where customers arrive according to a Markovian point process, say with state space E and intensity matrix $\boldsymbol{C}+\boldsymbol{D}$ for the background Markov process $\{J_t\}$ as above, and where a customer arriving at a transition $i \to j$ has a phase–type service time distribution B_{ij} with representation say $(E^{(ij)}, \boldsymbol{T}^{(ij)}, \alpha^{(ij)})$. The netput S_t up to time t is then the sum of the service times of the customers who have arrived arrived minus t, and the reflected version $\{V_t\}$ of $\{S_t\}$ is the $MArP/PH/1$ workload process. A main issue in applications is to evaluate the steady–state distribution of (J_t, V_t), say $\mathbb{P}_e(J_t = j, V_t > x)$ where $\mathbb{P}_e$ refers to the stationary situation, and we will explain here that this reduces to calculating the steady–state distribution in an associated Markovian fluid model $\big\{(I_t, F_t)\big\}$.

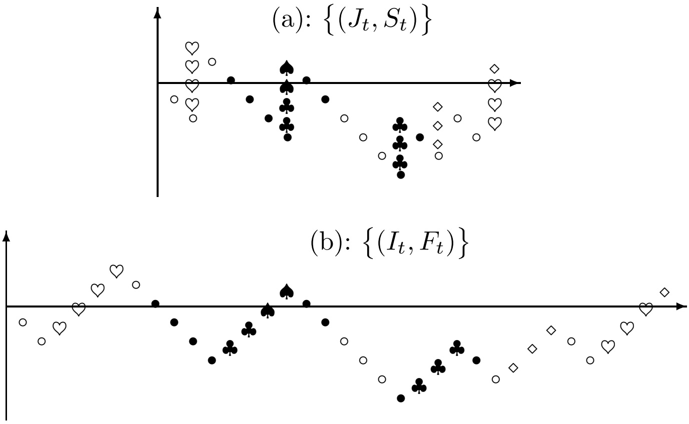

Figure 1.3

The idea is explained on Fig. 1.3, where there are two environmental states denoted $\circ, \bullet$. The phase space $E^{(\circ\circ)}$ for $B_{\circ\circ}$ has states $\diamond, \heartsuit$, and the one $E^{(\bullet\bullet)}$ for $B_{\bullet\bullet}$ states $\clubsuit, \spadesuit$, whereas $B_{\circ\bullet}$, $B_{\bullet\circ}$ are both degenerate at 0. An upwards jump in state i can be represented by an $E^{(ii)}$–valued Markov process as on Fig. 1.3(a). The fluid model $\{(I_t, F_t)\}$ on Fig. 1.3(b) is then obtained by changing the vertical jumps to segments with slope 1. Thus the state space is $F = \{\circ, \diamond, \heartsuit, \bullet, \clubsuit, \spadesuit\}$.

In the general formulation, the set E_I of Markov states for $\big\{I_t\big\}$ is the disjoint union of E and the $E^{(ij)}$,

$$E_I \;=\; E \cup \big\{(ij\alpha) : i,j \in E, \alpha \in E^{(ij)}\big\}, \quad r(i) = -1,\ i \in E,\ \ r(ij\alpha) = 1.$$

The intensity matrix for $\{I_t\}$ is (taking E with two elements for simplicity)

$$\left(\begin{array}{cc|cccc} c_{11} & c_{12} & d_{11}\boldsymbol{\alpha}^{(11)} & d_{12}\boldsymbol{\alpha}^{(12)} & \mathbf{0} & \mathbf{0} \\ c_{21} & c_{22} & \mathbf{0} & \mathbf{0} & d_{21}\boldsymbol{\alpha}^{(21)} & d_{22}\boldsymbol{\alpha}^{(22)} \\ \hline \boldsymbol{t}^{(11)} & \mathbf{0} & \boldsymbol{T}^{(11)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{t}^{(12)} & \mathbf{0} & \boldsymbol{T}^{(12)} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{t}^{(21)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{T}^{(21)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{t}^{(22)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{T}^{(22)} \end{array}\right)$$

where $\boldsymbol{t}^{(i)} = -\boldsymbol{T}^{(i)}\mathbf{1}$ is the exit rate vector. The following observation is immediate:

Proposition 1.12 *For $j \in E$, $x \geq 0$ one has*

$$\mathbb{P}_e(J_t = j, V_t > x) \;=\; \frac{\mathbb{P}_e(I_t = j, F_t > x)}{\mathbb{P}_e(I_t \in E)}\,.$$

2 Markov Additive Processes

2a. Definition and Structure
2b. Matrix M.G.F.'s
2c. Wald Martingales
2d. Mean, Variance and Related Limit Theorems
2e. Reflection and Time Reversion
2f. Wiener–Hopf Factorization

2a Definition and Structure

A Markov additive processes, abbreviated as MAP in this section,[2] is defined as a bivariate Markov process $\{X_t\} = \{(J_t, S_t)\}$ where $\{J_t\}$ is a Markov process with state space E (say) and the increments of $\{S_t\}$ are governed by $\{J_t\}$ in the sense that

$$\mathbb{E}\big[f(S_{t+s} - S_t)g(J_{t+s})\,\big|\,\mathscr{F}_t\big] \;=\; \mathbb{E}_{J_t,0}[f(S_s)g(J_s)]. \tag{2.1}$$

For shorthand, we write $\mathbb{P}_i, \mathbb{E}_i$ instead of $\mathbb{P}_{i,0}, \mathbb{E}_{i,0}$ in the following.

The structure of MAP's is completely understood when E is finite (the proofs that a MAP must look as postulated is elementary in the discrete–time case and somewhat similar to the derivation of the form of a Lévy process in IX.1 in the continuous–time case):

In discrete time, a MAP is specified by the measure–valued matrix (kernel) $\boldsymbol{F}(\mathrm{d}x)$ whose ijth element is the defective probability distribution $F_{ij}(\mathrm{d}x) = \mathbb{P}_{i,0}(J_1 = j, Y_1 \in \mathrm{d}x)$ where $Y_n = S_n - S_{n-1}$.

[2] But see footnote to Section 1.

An alternative description is in terms of the transition matrix $\boldsymbol{P} = (p_{ij})_{i,j\in E}$ (here $p_{ij} = \mathbb{P}_i(J_1 = j)$) and the probability measures

$$H_{ij}(\mathrm{d}x) \;=\; \mathbb{P}\bigl(Y_1 \in \mathrm{d}x \,\big|\, J_0 = i, J_1 = j\bigr) \;=\; \frac{F_{ij}(\mathrm{d}x)}{p_{ij}}.$$

In simulation language, this means that the MAP can be simulated by first simulating the Markov chain $\{J_n\}$ and next the $Y_1, Y_2, \ldots$ by generating Y_n according to H_{ij} when $J_{n-1} = i$, $J_n = j$.

If all F_{ij} are concentrated on $(0,\infty)$, a MAP is the same as a semi–Markov or Markov renewal process (cf. VII.4), with the Y_n being interpreted as interarrival times.

It is often convenient to assume that F_{ij} is independent of j. In fact, this can always be achieved by changing the driving Markov chain to $\{(J_n, J_{n+1})\}$. The particular case where F_{ij} is degenerate, say at $f(i)$, is often encountered, and then $S_n = f(J_0) + \cdots + f(J_{n-1})$.

In continuous time (assuming D–paths), $\{J_t\}$ is specified by its intensity matrix $\boldsymbol{\Lambda} = (\lambda_{ij})_{i,j\in E}$. On an interval $[t, t+s)$ where $J_t \equiv i$, $\{S_t\}$ evolves like a Lévy process $\bigl\{S_t^{(i)}\bigr\}$ with characteristic triplet $(\nu_i, \mu_i, \sigma_i^2)$, i.e. Lévy exponent

$$\kappa^{(i)}(\alpha) \;=\; \alpha\mu_i + \alpha^2\sigma_i^2/2 + \int_{-\infty}^{\infty} \bigl[\mathrm{e}^{\alpha x} - 1 - \alpha x I(|x| \le 1)\bigr]\, \nu_i(\mathrm{d}x)\,;$$

cf. IX.(1.4). In addition, a jump of $\{J_t\}$ from i to $j \ne i$ has probability q_{ij} of triggering to a jump of $\{S_t\}$ at the same time, the distribution of which has some distribution B_{ij}.

It follows immediately that the models of Sections 1a–1c are MAP's. It also follows in particular that the *MArP* is the most general point process such that the counting process $\{N_t\}$ is a MAP.

If E is infinite a MAP may be much more complicated. As an example, let $\{J_t\}$ be standard Brownian motion on the line. Then a Markov additive process can be defined by letting

$$S_t \;=\; \lim_{\epsilon\downarrow 0} \frac{1}{2\epsilon t} \int_0^t I\bigl(|J_s| \le \epsilon\bigr)\,\mathrm{d}s$$

be the local time at 0 up to time t.

In the following, we assume throughout E to be finite.

Problems

2.1 Show that a MAP in discrete time must look as postulated.

Notes Some basic sources for the foundations and structure of MAP's are Neveu (1961) and Çinlar (1972). The theory exposed later in this section was to a large extent developed in papers in the 1960s by Nagaev, Miller and Keilson and Wishart (only discrete time is treated there, but for most topics the continuous time theory is entirely similar); Asmussen (1989) contains a list of references to this work as well as to later studies of Wiener–Hopf factorization in discrete time. Work on special cases of Wiener–Hopf factorization in continuous time was done by Barlow, Rogers and Williams; see Rogers (1994) for references.

Some selected references on *MAP*'s on an infinite state space are Ney and Nummelin (1987), Fuh and Lai (1998), de Acosta and Ney (1998) and (for Brownian motion) Revuz and Yor (1999). Discrete time *MAP*'s are essentially the same as hidden Markov models in statistics, see Elliot *et al.* (1995) and MacDonald and Zucchini (1997).

2b Matrix M.G.F.'s

As a generalization of the m.g.f., consider the matrix $\widehat{\boldsymbol{F}}_t[\alpha]$ with ijth element $\mathbb{E}_i\bigl[\mathrm{e}^{\alpha S_t};\, J_t = j\bigr]$.

Proposition 2.1 *For a MAP in discrete time,* $\widehat{\boldsymbol{F}}_n[\alpha] = \widehat{\boldsymbol{F}}[\alpha]^n$ *where*

$$\begin{aligned}\widehat{\boldsymbol{F}}[\alpha] &= \widehat{\boldsymbol{F}}_1[\alpha] = \bigl(\mathbb{E}_i[\mathrm{e}^{\alpha S_1}; J_1 = j]\bigr)_{i,j\in E} = \bigl(\widehat{F}_{ij}[\alpha]\bigr)_{i,j\in E}\\ &= \bigl(p_{ij}\widehat{H}_{ij}[\alpha]\bigr)_{i,j\in E}.\end{aligned}$$

Proof. Conditioning upon (J_n, S_n) yields

$$\mathbb{E}_i[\mathrm{e}^{\alpha S_{n+1}};\, J_{n+1} = j] = \sum_{k\in E}\mathbb{E}_i[\mathrm{e}^{\alpha S_n};\, J_n = k]\,\mathbb{E}_k[\mathrm{e}^{\alpha Y_1};\, J_1 = j],$$

which in matrix formulation is the same as $\widehat{\boldsymbol{F}}_{n+1}[\alpha] = \widehat{\boldsymbol{F}}_n[\alpha]\widehat{\boldsymbol{F}}[\alpha]$. □

Proposition 2.2 *Consider a continuous–time MAP with parameters* $\boldsymbol{\Lambda}$, μ_i, σ_i^2, $\nu_i(\mathrm{d}x)$ *for* $i \in E$ *and* q_{ij}, B_{ij} *for* $i, j \in E$. *Then the matrix* $\widehat{\boldsymbol{F}}_t[\alpha]$ *with* ij*th element* $\mathbb{E}_i[\mathrm{e}^{\alpha S_t};\, J_t = j]$ *is given by* $\mathrm{e}^{t\boldsymbol{K}[\alpha]}$, *where*

$$\boldsymbol{K}[\alpha] = \boldsymbol{\Lambda} + \bigl(\kappa^{(i)}(\alpha)\bigr)_{\mathrm{diag}} + \bigl(\lambda_{ij}q_{ij}(\widehat{B}_{ij}[\alpha] - 1)\bigr).$$

Proof. Up to o(h) terms,

$$\begin{aligned}&\mathbb{E}_i[\mathrm{e}^{\alpha S_{t+h}};\, J_{t+h} = j]\\ &= (1 + \lambda_{jj}h)\mathbb{E}_i[\mathrm{e}^{\alpha S_t};\, J_t = j]\mathbb{E}_j\mathrm{e}^{\alpha S_h^{(j)}}\\ &\quad + \sum_{k\neq j}\lambda_{kj}h\,\mathbb{E}_i[\mathrm{e}^{\alpha S_t}; J_t = k]\bigl\{1 - q_{kj} + q_{kj}\widehat{B}_{kj}[\alpha]\bigr\}\\ &= \mathbb{E}_i[\mathrm{e}^{\alpha S_t}; J_t = j]\bigl(1 + h\kappa^{(j)}(\alpha)\bigr)\\ &\quad + h\sum_{k\in E}\mathbb{E}_i[\mathrm{e}^{\alpha S_t}; J_t = k]\bigl\{\lambda_{kj} + \lambda_{kj}q_{kj}(\widehat{B}_{kj}[\alpha] - 1)\bigr\}\end{aligned}$$

(recall that $q_{jj} = 0$). In matrix formulation, this means that

$$\widehat{\boldsymbol{F}}_{t+h}[\alpha] = \widehat{\boldsymbol{F}}_t[\alpha]\Bigl(\boldsymbol{I} + h\bigl(\kappa^{(i)}(\alpha)\bigr)_{\mathrm{diag}} + h\boldsymbol{\Lambda} + h\bigl(\lambda_{ij}q_{ij}(\widehat{B}_{ij}[\alpha] - 1)\bigr)\Bigr)$$

(up to $o(h)$ terms) so that $\widehat{\boldsymbol{F}}_t'[\alpha] = \widehat{\boldsymbol{F}}_t[\alpha]\boldsymbol{K}[\alpha]$, which in conjunction with $\widehat{\boldsymbol{F}}_0[\alpha] = \boldsymbol{I}$ implies $\widehat{\boldsymbol{F}}_t[\alpha] = \mathrm{e}^{t\boldsymbol{K}[\alpha]}$ according to the standard solution formula for systems of linear differential equations. □

2c Wald Martingales

In the following, assume that the Markov chain/process $\{J_t\}$ is ergodic. By Perron–Frobenius theory (I.6 and II.4d), we infer that in the discrete–time case the matrix $\widehat{\boldsymbol{F}}[\alpha]$ has a positive real eigenvalue with maximal absolute value, which we write as $\mathrm{e}^{\kappa(\alpha)}$, and that in the continuous–time case $\boldsymbol{K}[\alpha]$ has a real eigenvalue $\kappa(\alpha)$ with maximal real part. The corresponding left and right eigenvectors $\boldsymbol{\nu}^{(\alpha)} = (\nu_i)_{i\in E}$, $\boldsymbol{h}^{(\alpha)} = (h_i)_{i\in E}$ may be chosen with strictly positive components. Since $\boldsymbol{\nu}^{(\alpha)}$, $\boldsymbol{h}^{(\alpha)}$ are only given up to constants, we are free to impose two normalizations, and we shall take

$$\boldsymbol{\nu}^{(\alpha)}\boldsymbol{h}^{(\alpha)} = 1, \quad \boldsymbol{\pi}\boldsymbol{h}^{(\alpha)} = 1, \tag{2.2}$$

where $\boldsymbol{\pi} = \boldsymbol{\nu}^{(0)}$ is the stationary distribution. Then $\boldsymbol{h}^{(0)} = \boldsymbol{1}$.

The function $\kappa(\alpha)$ plays in many respects the same role as the cumulant g.f. of a random walk, as will be seen from the following results. In particular, its derivatives are "asymptotic cumulants", cf. Corollary 2.7, and appropriate generalizations of the Wald martingale (and the associated change of measure; cf. XIII.8) can be defined in terms of $\kappa(\alpha)$ (and $\boldsymbol{h}^{(\alpha)}$); cf. Proposition 2.4.

Corollary 2.3 $\mathbb{E}_i[\mathrm{e}^{\alpha S_t}; J_t = j] \sim h_i^{(\alpha)}\nu_j^{(\alpha)}\mathrm{e}^{t\kappa(\alpha)}$.

Proof. By Perron–Frobenius theory, $\widehat{\boldsymbol{F}}_t[\alpha] \sim \boldsymbol{h}(\alpha)\boldsymbol{\nu}(\alpha)\mathrm{e}^{t\kappa(\alpha)}$. □

Proposition 2.4 $\mathbb{E}_i\mathrm{e}^{\alpha S_t}h_{J_t}^{(\alpha)} = h_i^{(\alpha)}\mathrm{e}^{t\kappa(\alpha)}$. *Furthermore,* $\bigl\{\mathrm{e}^{\alpha S_t - t\kappa(\alpha)}h_{J_t}^{(\alpha)}\bigr\}$ *is a martingale.*

Proof. For the first assertion, just note that

$$\mathbb{E}_i\mathrm{e}^{\alpha S_t}h_{J_t}^{(\alpha)} = \boldsymbol{1}_i'\widehat{\boldsymbol{F}}_t[\alpha]\boldsymbol{h}^{(\alpha)} = \boldsymbol{1}_i'\mathrm{e}^{t\boldsymbol{K}[\alpha]}\boldsymbol{h}^{(\alpha)} = \boldsymbol{1}_i'\mathrm{e}^{t\kappa(\alpha)}\boldsymbol{h}^{(\alpha)} = \mathrm{e}^{t\kappa(\alpha)}h_i^{(\alpha)}.$$

It then follows that

$$\begin{aligned}
&\mathbb{E}\Bigl[\mathrm{e}^{\alpha S_{t+v}-(t+v)\kappa(\alpha)}h_{J_{t+v}}^{(\alpha)}\Bigm|\mathscr{F}_t\Bigr]\\
&= \mathrm{e}^{\alpha S_t - t\kappa(\alpha)}\mathbb{E}\Bigl[\mathrm{e}^{\alpha(S_{t+v}-S_t)-v\kappa(\alpha)}h_{J_{t+v}}^{(\alpha)}\Bigm|\mathscr{F}_t\Bigr]\\
&= \mathrm{e}^{\alpha S_t - t\kappa(\alpha)}\mathbb{E}_{J_t}\Bigl[\mathrm{e}^{\alpha S_v - v\kappa(\alpha)}h_{J_v}^{(\alpha)}\Bigr] = \mathrm{e}^{\alpha S_t - t\kappa(\alpha)}h_{J_t}^{(\alpha)}.
\end{aligned}$$

□

2d Mean, Variance and Related Limit Theorems

Let $\boldsymbol{k}^{(\alpha)}$ denote the derivative of $\boldsymbol{h}^{(\alpha)}$ w.r.t. α, and write $\boldsymbol{k} = \boldsymbol{k}^{(0)}$.

Corollary 2.5 $\mathbb{E}_i S_t \;=\; t\kappa'(0) + k_i - \mathbb{E}_i k_{J_t} \;=\; t\kappa'(0) + k_i - \mathbf{1}_i' \mathrm{e}^{\mathbf{\Lambda} t}\boldsymbol{k}.$

Proof. By differentiation in Proposition 2.4,

$$\mathbb{E}_i\left[S_t \mathrm{e}^{\alpha S_t} h_{J_t}^{(\alpha)} + \mathrm{e}^{\alpha S_t} k_{J_t}^{(\alpha)}\right] \;=\; \mathrm{e}^{t\kappa(\alpha)}\left(k_i^{(\alpha)} + t\kappa'(\alpha) h_i^{(\alpha)}\right). \tag{2.3}$$

Let $\alpha = 0$ and recall that $\boldsymbol{h}^{(0)} = \mathbf{1}$ so that $h_i^{(0)} = h_{J_t}^{(0)} = 1$. □

The argument is slightly heuristic (e.g. the existence of exponential moments is assumed) but can be made rigorous by passing to characteristic functions. In the same way, one obtains a generalization of Wald's identity $\mathbb{E}S_\tau = \mathbb{E}\tau \cdot \mathbb{E}S_1$ for a random walk:

Corollary 2.6 *For any stopping time τ with $\mathbb{E}_i\tau < \infty$, one has $\mathbb{E}_i S_\tau = \kappa'(0)\mathbb{E}\tau + k_i - \mathbb{E}_i k_{J_\tau}$.*

Corollary 2.7 *No matter the initial distribution $\boldsymbol{\mu}$ of J_0,*

$$\lim_{t\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}} S_t}{t} \;=\; \kappa'(0), \qquad \lim_{t\to\infty} \frac{\mathbb{V}ar_{\boldsymbol{\mu}} S_t}{t} \;=\; \kappa''(0).$$

Proof. The first assertion is immediate by dividing by t in Corollary 2.5. For the second, we differentiate (2.3) to get

$$\begin{aligned}
&\mathbb{E}_i\left[S_t^2 \mathrm{e}^{\alpha S_t} h_{J_t}^{(\alpha)} + 2S_t \mathrm{e}^{\alpha S_t} k_{J_t}^{(\alpha)} + \mathrm{e}^{\alpha S_t} k_{J_t}^{(\alpha)'}\right] \\
&\quad=\; \mathrm{e}^{t\kappa(\alpha)}\left(k_i^{(\alpha)'} + t\kappa'(\alpha)k_i^{(\alpha)} + t\bigl\{\kappa''(\alpha)h_i^{(\alpha)} + t\kappa'(\alpha)^2 h_i^{(\alpha)} + \kappa'(\alpha)k_i^{(\alpha)}\bigr\}\right).
\end{aligned}$$

Multiplying by μ_i, summing and letting $\alpha = 0$ yields

$$\mathbb{E}_{\boldsymbol{\mu}}[S_t^2 + 2S_t k_{J_t}] + \mathrm{O}(1) \;=\; t^2\kappa'(0)^2 + 2t\kappa'(0)\boldsymbol{\mu}\boldsymbol{k} + t\kappa''(0) + \mathrm{O}(1).$$

Squaring in Corollary 2.5 yields

$$[\mathbb{E}_{\boldsymbol{\mu}} S_t]^2 \;=\; t^2\kappa'(0)^2 + 2t\kappa'(0)\boldsymbol{\mu}\boldsymbol{k} - 2t\kappa'(0)\mathbb{E}_{\boldsymbol{\mu}} k_{J_t} + \mathrm{O}(1).$$

Since it is easily seen by an asymptotic independence argument that $\mathbb{E}_{\boldsymbol{\mu}}[S_t k_{J_t}] = t\kappa'(0)\mathbb{E}_{\boldsymbol{\mu}} k_{J_t} + \mathrm{O}(1)$, subtraction yields $\mathbb{V}ar_{\boldsymbol{\mu}} S_t = t\kappa''(0) + \mathrm{O}(1)$. □

Corollary 2.8 *$S_t/t \to \kappa'(0)$ $\mathbb{P}_{ix}$–a.s. for all i, x. If in addition $\kappa''(0)$ is welldefined and satisfies $0 < \kappa''(0) < \infty$, then $\bigl(S_t - t\kappa'(0)\bigr)/t^{1/2}$ has a limiting normal $(0, \kappa''(0))$ distribution.*

Proof. By general results on cumulative processes from VI.3. □

Corollary 2.9 (a) *In discrete time,* $\kappa'(0) = \sum_{i,j\in E} \pi_i p_{ij} \int_{\mathbb{R}} x\, H_{ij}(\mathrm{d}x)$.
(b) *In continuous time,*

$$\kappa'(0) \;=\; \sum_{i\in E} \pi_i\Bigl\{\mu_i + \int_{|x|>1} x\,\nu_i(\mathrm{d}x)\Bigr\} + \sum_{i\neq j} \pi_i\lambda_{ij}q_{ij}\int_{\mathbb{R}} x\, B_{ij}(\mathrm{d}x).$$

Proof. (a) For N large, there are approximately $N\pi_i p_{ij}$ pairs $(n-1,n)$ with $X_{n-1}=i, X_n=j$, and the sum of the corresponding Y_n is therefore approximately $N\pi_i p_{ij}\int_{\mathbb{R}} x\,H_{ij}(\mathrm{d}x)$ so that S_N/N is approximately the stated expression for $\kappa'(0)$. But on the other hand, we know that $S_N/N \overset{\text{a.s.}}{\to} \kappa'(0)$. The proof of (b) is similar. □

We call the *MAP degenerate* if $\sup|S_t|<\infty$.

Proposition 2.10 *We have* (a′) $S_t\to-\infty$, (b′) $\underline{\lim}\,S_t=-\infty$, $\overline{\lim}\,S_t=\infty$, (c′) $S_t\to\infty$ $\mathbb{P}_{ix}$*-a.s. for all* i,x *according as* (a) $\kappa'(0)<0$, (b) $\kappa'(0)=0$ *and the MAP is nondegenerate,* (c) $\kappa'(0)>0$. *Further, letting* $\tau(x)=\inf\{t: S_t>x\}$, *it holds for* $i\in E$ *in case* (c) *that* $\mathbb{E}_i\tau(x)\sim x/\kappa'(0)$, $x\to\infty$, *and in case* (b) *that* $\mathbb{E}_i\tau(x)=\infty$ *for all large* x.

Proof. By Corollary 2.8, $S_t/t \overset{\text{a.s.}}{\to} \kappa'(0)$ which immediately shows that (a) → (a′), (c) → (c′). In case (b), let $\tau(i,k)$ be the time of the kth entrance of $\{J_t\}$ to $i\in E$. Then $\{S_{\tau(i,k)}\}_{k=0,1,\dots}$ is a discrete random walk with (by regenerative process theory) mean 0. Nondegeneracy implies that the increment distribution is not concentrated at 0, and hence $\underline{\lim}\,S_{\tau(i,k)}=-\infty$, $\overline{\lim}=\infty$ by VIII.2.4, implying (b′). The last statement is an easy consequence of Corollary 2.6. □

2e Reflection and Time Reversion

The time–reversed *MAP* $\{(J_t^*,S_t^*)\}$ and the reflected *MAP* $\{(J_t,V_t)\}$ are defined as in IX.2.

The definition of the time–reversed MAP means that $\{J_t^*\}$ is the usual reversed Markov process, with transition matrix $\boldsymbol{P}^*=\boldsymbol{\Delta}_{\boldsymbol{\pi}}^{-1}\boldsymbol{P}^{\mathsf{T}}\boldsymbol{\Delta}_{\boldsymbol{\pi}}$ in discrete time and intensity matrix $\boldsymbol{\Lambda}^*=\boldsymbol{\Delta}_{\boldsymbol{\pi}}^{-1}\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Delta}_{\boldsymbol{\pi}}$ in continuous time, where $\boldsymbol{\Delta}_{\boldsymbol{\pi}}$ is the diagonal matrix with $\boldsymbol{\pi}$ on the diagonal. For the additive part, we have in discrete time to replace H_{ij} by $H_{ij}^*=H_{ji}$ (or, equivalently, F_{ij} by $F_{ij}^*=\pi_jF_{ji}/\pi_i$), whereas in continuous time the Lévy processes $\{S_t^{(i)}\}$ remain unchanged and we only have to replace B_{ij} by $B_{ij}^*=B_{ji}$ and q_{ij} by $q_{ij}^*=q_{ji}$. From IX.2, we immediately get:

Proposition 2.11 *The reflected MAP* $\{(J_t,V_t)\}$ *is a Markov process. If* $\kappa'(0)>0$, *then* $V_t\to\infty$ *a.s., whereas* $V_t\to\infty$ *in distribution if* $\kappa'(0)=0$. *If* $\kappa'(0)<0$, *then* (J_t,V_t) *has a total variation limit which coincides with the distribution of* $\bigl(J_0^*,\sup_{t\ge0}S_t^*\bigr)$ *where* J_0^* *is chosen with distribution* $\boldsymbol{\pi}$.

2f Wiener–Hopf Factorization

Recall the Wiener–Hopf factorization identity in VIII.3 for a simple random walk: $\delta_0-F=(\delta_0-G_-)*(\delta_0-G_+)$ where F is the increment distribution, and G_+, G_- the ladder height distributions (δ_0 is the distribution degenerate at 0).

For MAP's, we only consider the discrete–time case and define

$$\begin{aligned} \tau_+ &= \inf\{n>0:\, S_n>0\}, \quad \tau_- \;=\; \inf\{n>0:\, S_n\le 0\}, \\ G_+(i,j;A) &= \mathbb{P}_i\bigl(J_{\tau_+}=j, S_{\tau_+}\in A, \tau_+<\infty\bigr), \\ G_-(i,j;A) &= \mathbb{P}_i\bigl(J_{\tau_-}=j, S_{\tau_-}\in A, \tau_-<\infty\bigr). \end{aligned}$$

Further, $\boldsymbol{G}_+$ denotes the (measure–valued) matrix with ijth element $G_+(i,j;\cdot)$ (and similarly for $\boldsymbol{G}_-$), and $\boldsymbol{G}^*_+$ etc. refer to the time–reversed MAP $\{(J^*_t, S^*_t)\}$. We write ${}^{\#}\boldsymbol{G}_- = \boldsymbol{\Delta}_{\boldsymbol{\pi}}^{-1}\boldsymbol{G}^{*\mathsf{T}}_-\boldsymbol{\Delta}_{\boldsymbol{\pi}}$ and similarly for ${}^{\#}\boldsymbol{G}_+$. In the measure–valued case, $\boldsymbol{A}*\boldsymbol{B}$ denotes the matrix with ijth element $\sum_{k\in E}A(i,k;\cdot)*B(k,j;\cdot)$, and $\boldsymbol{I}=(\delta_0)_{\mathrm{diag}}$.

Theorem 2.12 $\boldsymbol{I}-\boldsymbol{F} \;=\; (\boldsymbol{I}-{}^{\#}\boldsymbol{G}_-)*(\boldsymbol{I}-\boldsymbol{G}_+) \;=\; (\boldsymbol{I}-{}^{\#}\boldsymbol{G}_+)*(\boldsymbol{I}-\boldsymbol{G}_-)$.

For the proof, we define

$$R_+(i,j;A) \;=\; \mathbb{E}_i\sum_{n=0}^{\tau_+-1} I(J_n=j, S_n\in A), \quad {}^{\#}\boldsymbol{U}_- \;=\; \sum_{n=0}^{\infty}({}^{\#}\boldsymbol{G}_-)^{*n}.$$

Proposition 2.13 $\boldsymbol{R}_+ \;=\; {}^{\#}\boldsymbol{U}_-$.

Proof. Let n be fixed and write $i=i_0$, $j=i_n$,

$$\begin{aligned} B &= \bigl\{J_0=i_0, J_1=i_1,\ldots,J_n=i_n\bigr\}, \\ B^* &= \bigl\{J^*_0=i_n, J^*_1=i_{n-1},\ldots,J^*_n=i_0\bigr\}. \end{aligned}$$

For $A\subseteq(-\infty,0]$, we have

$$\begin{aligned} &\mathbb{P}_i\bigl(\tau_+>n, J_n=j, S_n\in A\bigr) \\ &= \sum_{i_1,\ldots,i_{n-1}} p_{i_0i_1}\cdots p_{i_{n-1}i_n}\mathbb{P}\bigl(S_k\le 0, k\le n, S_n\in A\,\big|\,B\bigr) \\ &= \frac{\pi_{i_n}}{\pi_{i_0}}\sum_{i_1,\ldots,i_{n-1}} p^*_{i_ni_{n-1}}\cdots p^*_{i_1i_0}\mathbb{P}\bigl(S^*_k\ge S^*_n, k\le n, S^*_n\in A\,\big|\,B^*\bigr) \\ &= \frac{\pi_{i_n}}{\pi_{i_0}}\mathbb{P}_j\bigl(J^*_n=i, S^*_n\in A;\, F^*_n\bigr) \;=\; \frac{\pi_j}{\pi_i}\mathbb{P}_j\bigl(J^*_n=i, S^*_n\in A;\, F^*_n\bigr), \end{aligned}$$

where F^*_n is the event that n is a descending ladder epoch for $\{S^*_n\}$ (in the second step, we used $\{S_k\}_0^n \overset{\mathscr{D}}{=} \bigl\{S^*_n-S^*_{n-k}\bigr\}_0^n$). Summing over n, we get

$$R_+(i,j;A) \;=\; \frac{\pi_j}{\pi_i}\sum_{m=0}^{\infty}(\boldsymbol{G}^*_-)^{*m}(j,i;A) \;=\; \sum_{m=0}^{\infty}({}^{\#}\boldsymbol{G}_-)^{*m}(i,j;A),$$

where the last equality easily follows by induction. □

Proof of Theorem 2.12. Adapting the general stopping time identity I.3.3 as in the proof of VIII.3.1 easily gives $\boldsymbol{R}_++\boldsymbol{G}_+ = \boldsymbol{I}+\boldsymbol{R}_+*\boldsymbol{F}$ which by Proposition 2.13 we may rewrite as ${}^{\#}\boldsymbol{U}_-+\boldsymbol{G}_+ = \boldsymbol{I}+{}^{\#}\boldsymbol{U}_-*\boldsymbol{F}$. Convolving with ${}^{\#}\boldsymbol{G}_-$ to the left, we get ${}^{\#}\boldsymbol{U}_--\boldsymbol{I}+{}^{\#}\boldsymbol{G}_-*\boldsymbol{G}_+ = {}^{\#}\boldsymbol{G}_-+{}^{\#}\boldsymbol{U}_-*\boldsymbol{F}-\boldsymbol{F}$.

Subtracting the two last identities yields $\boldsymbol{G}_+ + \boldsymbol{I} - {}^{\#}\boldsymbol{G}_- * \boldsymbol{G}_+ = \boldsymbol{I} - {}^{\#}\boldsymbol{G}_- + \boldsymbol{F}$ which is the same as the first of the stated identities. The proof of the second is similar. □

For a measure–valued matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|$ denote the matrix whose ijth element is the total mass $A(i,j;\mathbb{R})$ of $A(i,j;\cdot)$.

Proposition 2.14 (a) *If* $\kappa'(0) < 0$*, then* $\|\boldsymbol{G}_+\|$ *is substochastic* $(\mathrm{spr}\big(\|\boldsymbol{G}_+\|\big) < 1)$*, whereas* $\mathrm{spr}\big(\|{}^{\#}\boldsymbol{G}_-\|\big) = 1$ *and* $\boldsymbol{\pi}\|{}^{\#}\boldsymbol{G}_-\| = \boldsymbol{\pi}$*.*
(b) *If* $\kappa'(0) > 0$*, then* $\mathrm{spr}\big(\|{}^{\#}\boldsymbol{G}_-\|\big) < 1$*, whereas* $\|\boldsymbol{G}_+\|$ *is stochastic with* $\boldsymbol{\pi}_+ = \boldsymbol{\pi}(\boldsymbol{I} - {}^{\#}\boldsymbol{G}_-)$ *as left eigenvector.*
(c) *If* $\kappa'(0) = 0$*, then* $\mathrm{spr}\big(\|\boldsymbol{G}_+\|\big) = \mathrm{spr}\big(\|{}^{\#}\boldsymbol{G}_-\|\big) = 1$*.*

Proof. In (a), we have $S_n \overset{\text{a.s.}}{\to} -\infty$ and $S_n^* \overset{\text{a.s.}}{\to} -\infty$, which immediately yields that $\|\boldsymbol{G}_+\|$ is substochastic and $\|\boldsymbol{G}_-^*\|$ stochastic so that $\mathrm{spr}(\|\boldsymbol{G}_-^*\|) = 1$. The way ${}^{\#}\boldsymbol{G}_-$ is constructed from $\boldsymbol{G}_-^*$ then ensures that also $\mathrm{spr}(\|{}^{\#}\boldsymbol{G}_-\|) = 1$, and $\boldsymbol{\pi}\|{}^{\#}\boldsymbol{G}_-\| = \boldsymbol{\pi}$ follows from

$$\sum_{i\in E} \pi_i {}^{\#}G_-(i,j;\mathbb{R}) \;=\; \pi_j \sum_{i\in E} G_-^*(j,i;\mathbb{R}) \;=\; \pi_j .$$

Also (c) and the first part of (b) is similar to (a). For the last claim in (b), note that $\mathrm{spr}(\|{}^{\#}\boldsymbol{G}_-\|) < 1$ implies that $\boldsymbol{\pi}_+ \neq \boldsymbol{0}$. Also Theorem 2.12 yields $\boldsymbol{I} - \|\boldsymbol{F}\| = (\boldsymbol{I} - \|{}^{\#}\boldsymbol{G}_-\|)(\boldsymbol{I} - \|\boldsymbol{G}_+\|)$, and multiplying by $\boldsymbol{\pi} = \boldsymbol{\pi}\|\boldsymbol{F}\|$ to the left, we get $\boldsymbol{0} = \boldsymbol{\pi}_+(\boldsymbol{I} - \|\boldsymbol{G}_+\|)$. □

3 The Matrix Paradigms $GI/M/1$ and $M/G/1$

3a. Recurrence and Positive Recurrence
3b. $GI/M/1$ Type Queueing Models
3c. $M/G/1$ Type Queueing Models. QBD's
3d. The Sengupta Process

3a Recurrence and Positive Recurrence

Assume now that the additive component S_t of the discrete–time MAP $\{(J_n, S_n)\}$ is lattice, i.e. the MAP has state space $E \times \mathbb{Z}$. We can write $f_{ij}(k) = F_{ij}(\{k\})$ (the probability of adding $k \in \mathbb{Z}$ to the level and changing the phase from i to j), $\boldsymbol{F}(k) = \big(f_{ij}(k)\big)$, and can then write the $(E \times \mathbb{Z}) \times$

$(E \times \mathbb{Z})$ transition matrix for the MAP as

$$\begin{pmatrix} \ddots & & & \vdots & & & \\ & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{F}(2) & \boldsymbol{F}(3) & \boldsymbol{F}(4) & \\ & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{F}(2) & \boldsymbol{F}(3) & \\ \cdots & \boldsymbol{F}(-2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{F}(2) & \cdots \\ & \boldsymbol{F}(-3) & \boldsymbol{F}(-2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \\ & \boldsymbol{F}(-4) & \boldsymbol{F}(-3) & \boldsymbol{F}(-2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \\ & & & \vdots & & & \ddots \end{pmatrix} \tag{3.1}$$

by partitioning into $E \times E$ blocks corresponding to levels.

We shall consider some modifications of the MAP at 0, which are more complicated than the simple reflection considered in Section 2e and highly adaptable from the point of view of specific applications. The processes $\{(I_n, L_n)\}$ in question are Markov chains on $E_0 \times \{0\} \cup E \times \{1, 2, \ldots\}$; the I–component is called the *phase* and the L–component the *level*. Thus E_0 is a set of boundary states in which I_n takes its values when $L_n = 0$ (otherwise $I_n \in E$). The state space is shown in Fig. 3.1 for the case where E_0 has seven elements and E four.

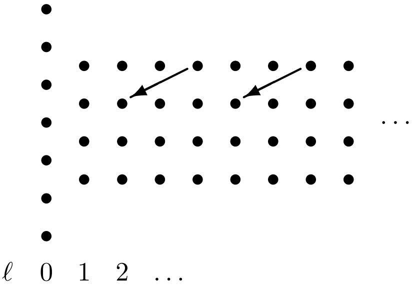

Figure 3.1

Away from level 0, $\{(I_n, L_n)\}$ moves as $\{(J_n, S_n)\}$ (e.g., the two transitions marked on Fig. 3.1 have the same probability). However, when the MAP at level $\ell > 0$ attempts to go to a level < 0, $\{(I_n, L_n)\}$ is reset to level 0 and a phase in E_0 with probabilities depending on ℓ; the jump out of level 0 may have any distribution. It follows that the transition matrix $\boldsymbol{Q}$ can be written in the block–partitioned form as

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{A}(1) & \boldsymbol{A}(2) & \boldsymbol{A}(3) & \ldots \\ \boldsymbol{B}(1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{F}(2) & \ldots \\ \boldsymbol{B}(2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \ldots \\ \boldsymbol{B}(3) & \boldsymbol{F}(-2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \ldots \\ \vdots & & & & \ddots \end{pmatrix} \tag{3.2}$$

where the dimensions are $\boldsymbol{C} : E_0 \times E_0$, $\boldsymbol{A}(k) : E_0 \times E$, $\boldsymbol{B}(\ell) : E \times E_0$, $\boldsymbol{F}(\ell) : E \times E$. Particular important forms are obtained by letting the *MAP* be right– or left–continuous (skip–free) for levels, i.e. of one of the forms[3]

$$\boldsymbol{Q} \;=\; \begin{pmatrix} \boldsymbol{B}(0) & \boldsymbol{F}(1) & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{B}(1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{0} & \dots \\ \boldsymbol{B}(2) & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \dots \\ \boldsymbol{B}(3) & \boldsymbol{F}(-2) & \boldsymbol{F}(1) & \boldsymbol{F}(0) & \dots \\ \vdots & & & & \ddots \end{pmatrix}, \tag{3.3}$$

$$\boldsymbol{Q} \;=\; \begin{pmatrix} \boldsymbol{A}(0) & \boldsymbol{A}(1) & \boldsymbol{A}(2) & \boldsymbol{A}(3) & \dots \\ \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{F}(2) & \dots \\ \boldsymbol{0} & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \dots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \dots \\ \vdots & & & & \ddots \end{pmatrix}. \tag{3.4}$$

We say that matrices of the form (3.3) are of *the $GI/M/1$ type* (note that here $E_0 = E$) and those of the form (3.4) of *the $M/G/1$ type*, the obvious motivation being the imbedded Markov chains in the queues $GI/M/1$ and $M/G/1$ (here E and E_0 are one–point sets so that blocks reduce to numbers); see III.6 and X.5.

The following ergodicity criterion covers most examples:

Proposition 3.1 *Assume in* (3.2) *that both $\boldsymbol{Q}$ and $\boldsymbol{P} = \sum_{-\infty}^{\infty} \boldsymbol{F}(k)$ are irreducible and stochastic, and let $\boldsymbol{\nu}$ be the stationary distribution of $\boldsymbol{P}$ and $\boldsymbol{\mu} = \sum_{-\infty}^{\infty} k\boldsymbol{F}(k)\boldsymbol{1}$. Then $\{(I_n, L_n)\}$ is recurrent if and only if $\boldsymbol{\nu\mu} \le 0$, and positive recurrent if and only if* (a) $\sum_{-\infty}^{\infty} k\boldsymbol{A}(k)\boldsymbol{1} < \infty$ *and* (b) $\boldsymbol{\nu\mu} < 0$.

Proof. The two Markov chains $\{(J_n, S_n)\}$, $\{(I_n, L_n)\}$ evolve in the same way in levels $\ell \ge 1$. More precisely, if we let $\sigma = \inf\{n \ge 1 : S_n \le 0\}$, $\tau = \inf\{n \ge 1 : L_n = 0\}$ and start in such a way that $J_0 = I_0$, $S_0 = L_0 \ge 1$, then $\tau = \sigma$ and $I_n = J_n$, $L_n = S_n$ for $n < \tau$. Hence recurrence of $\{(I_n, L_n)\}$ is equivalent to $\mathbb{P}_{j\ell}(\sigma < \infty)$ for all $j \in E$, $\ell \ge 1$, i.e. by Proposition 2.10 to $\boldsymbol{\nu\mu} \le 0$ (note that $\kappa'(0) = \boldsymbol{\nu\mu}$ by Corollary 2.9), and (by I.3.10) positive recurrence is equivalent to $\mathbb{E}_{i0}\tau < \infty$ for all $i \in E_0$. Now clearly,

$$\mathbb{E}_{i0}\tau \;=\; \sum_{j\in E} c_{ij} \;+\; \sum_{\ell=1}^{\infty}\sum_{j\in E} a_{ij}(\ell)\left\{1 + \mathbb{E}_{j\ell}\sigma\right\}. \tag{3.5}$$

If (a) fails, Proposition 2.10 shows that (irrespective of $\boldsymbol{\nu\mu} < 0$ or $\boldsymbol{\nu\mu} = 0$) this expression cannot be finite. If conversely (a) and (b) hold, then Proposition 2.10 shows immediately that (3.5) is finite. Finally, if (a) holds but not (b), we can choose ℓ_0 such that $\mathbb{E}_{j\ell}\sigma = \infty$ for all $\ell \ge \ell_0$ and all

[3] See the Notes at the end of the section for notational issues.

$j \in E$, and (by irreducibility) $i \in E$ such that $\mathbb{P}_{i0}(L_n \geq \ell_0, \tau > n) > 0$ for some n. Then $\mathbb{E}_{i0}\tau = \infty$, and positive recurrence fails. □

3b $GI/M/1$ *Type Queueing Models*

We consider a Markov chain $\{(I_n, L_n)\}$ with transition matrix $\boldsymbol{Q}$ of the form (3.3). The crucial feature is the skip–free–to–the–right–for–levels property, that $\{L_n\}$ can increase at most one in one time step.

Example 3.2 Consider the $GI/PH/1$ queue where the service time distribution has representation $(E, \boldsymbol{T}, \boldsymbol{\alpha})$, and let the level L_n be the number of customers just *before* the nth arrival and the phase I_n the phase in which the server is working just *after* the nth arrival. Let $p^t_{ij}(k)$ be the probability that the server, when facing an infinitely long queue and starting service at time 0 in phase i, will by time t have completed service of k customers and work in phase j. With A the interarrival distribution, it is seen that $\boldsymbol{Q}$ is of the form (3.3) corresponding to

$$f_{ij}(\ell) \;=\; \int_0^\infty p^t_{ij}(1-\ell)\, A(\mathrm{d}t), \quad b_{ij}(\ell) \;=\; \alpha_j \sum_{k=-\infty}^{\ell} \sum_{r \in E} f_{ir}(k).$$

Indeed, the expression for $f_{ij}(\ell)$ is clear, and so is the one for $b_{ij}(\ell)$ once one observes that the double sum is the probability that the server starting in phase i will serve at least ℓ customers within a service period (α_j is the probability that he will start the service of the next customer in phase j). □

Corollary 3.3 *Suppose that both* $\boldsymbol{Q}$ *in* (3.3) *and* $\boldsymbol{P} = \sum_{-\infty}^{1} \boldsymbol{F}(k)$ *are irreducible and stochastic, and let* $\boldsymbol{\nu}$ *be the stationary distribution of* $\boldsymbol{P}$ *and* $\boldsymbol{\mu} = \sum_{-\infty}^{1} k\boldsymbol{A}(k)\boldsymbol{1}$. *Then recurrence of* $\boldsymbol{Q}$ *is equivalent to* $\boldsymbol{\nu\mu} \leq 0$ *and positive recurrence to* $\boldsymbol{\nu\mu} < 0$.

Proof. Just note that $\boldsymbol{A}(k) = 0$ for $k > 1$ so that condition (a) in Proposition 3.1 is vacuous. □

In the ergodic case, the key to the analysis is to fix a level ℓ and think of $\{(I_n, L_n)\}$ as a semi–regenerative process with the semi–regeneration points as the returns to level ℓ (which may be of one the types $i \in E$). To this end, in the notation of VII.5, we let $C = \inf\{n \geq 1 : L_n = \ell\}$. The role of the J_n–chain in VII.5 is then taken by $\{I^{(\ell)}_n\}$ where $I^{(\ell)}_k$ is the phase at the kth return of $\{L_n\}$ to level ℓ. Define $\boldsymbol{R}(k)$ as the matrix with elements

$$r_{ij}(k) \;=\; \mathbb{E}_{i\ell} \sum_{n=0}^{C-1} I(I_n = j, L_n = \ell + k) \tag{3.6}$$

and write $\boldsymbol{R} = \boldsymbol{R}(1) = (r_{ij})$. For measures $\boldsymbol{\pi}$ on $E \times \mathbb{N}$, we use notation like $\boldsymbol{\pi} = (\boldsymbol{\pi}_\ell)_{\ell \in \mathbb{N}} = (\boldsymbol{\pi}_0\, \boldsymbol{\pi}_1\, \cdots)$ where $\boldsymbol{\pi}_\ell = (\pi_{i\ell})_{i \in E}$ denotes the restriction to level ℓ.

Lemma 3.4 (i) *The matrices* $\boldsymbol{R}(k)$ *do not depend on the choice of* $\ell = 0, 1, 2, \ldots$*;* (ii) $\boldsymbol{R}(k) = \boldsymbol{R}^k$*;* (iii) *in the recurrent case,* $\boldsymbol{\pi} = (\boldsymbol{\pi}_0\ \boldsymbol{\pi}_1\ \cdots)$ *is a stationary measure for* $\{(I_n, L_n)\}$ *if and only if* $\boldsymbol{\pi}_0$ *is stationary for* $\{I_n^{(0)}\}$ *and* $\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\boldsymbol{R}^k$, $k = 1, 2, \ldots$.

Proof. (i) is a consequence of the structure of $\boldsymbol{Q}$, which shows that only upward excursions from ℓ contribute to $r_{ij}(k)$, and also that such excursions are homogeneous in the level. Consider an excursion from ℓ and upward. Each visit to level $\ell + k - 1$ in phase i generates an average of r_{ij} visits to level $\ell + k$ in phase j before any of the levels $\ell + k - 1, \ell + k - 2, \ldots$ is entered. Therefore $\boldsymbol{R}(k) = \boldsymbol{R}(k-1)\boldsymbol{R}$, showing (ii). For (iii) let $\ell = 0$ and combine the definition of $\boldsymbol{R}(k) = \boldsymbol{R}$ with Corollary VII.5.3 to see that $\boldsymbol{\pi}_0$ is necessarily stationary for $\{I_n^{(0)}\}$ and connected to $\boldsymbol{\pi}_k$ by

$$\pi_{jk} \;=\; \sum_{i\in E}\pi_{i0}\,\mathbb{E}_{i0}\sum_{n=0}^{C-1} I(I_n = j, L_n = k) \;=\; \sum_{i\in E}\pi_{i0}r_{ij}(k)\,. \qquad \Box$$

For any $E \times E$ matrix $\boldsymbol{X}$ such that the series are welldefined and convergent, define (left and right)

$$\widehat{\boldsymbol{F}}_{\mathrm{L}}[\boldsymbol{X}] \;=\; \sum_{k=-\infty}^{\infty}\boldsymbol{X}^k\boldsymbol{F}(k), \quad \widehat{\boldsymbol{F}}_{\mathrm{R}}[\boldsymbol{X}] \;=\; \sum_{k=-\infty}^{\infty}\boldsymbol{F}(k)\boldsymbol{X}^k. \qquad (3.7)$$

Lemma 3.5 *The Markov chain* $\{I_n^{(0)}\}$ *has transition matrix* $\widehat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}]$.

Proof. Consider an excursion away from level 0, say of length $\omega = \inf\{n \ge 1 : L_n = 0\}$. Then $\mathbb{P}_i(I_1^{(0)} = j) = \mathbb{P}_{i0}(J_\omega = j)$. The contribution from $\{\omega = 1\} = \{L_{\omega-1} = 0\}$ to this matrix is clearly the term $\boldsymbol{B}(0)$ in $\widehat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}] = \sum_0^\infty \boldsymbol{R}^k\boldsymbol{B}(k)$. Further, collecting phases at visits to level $k \ge 1$ during the excursion gives an average governed by $\boldsymbol{R}^k$ so that the the contribution from $\{L_{\omega-1} = k\}$ is $\boldsymbol{R}^k\boldsymbol{B}(k)$. Collecting terms, the result follows. □

Theorem 3.6 *Consider an irreducible positive recurrent transition matrix* $\boldsymbol{Q}$ *of the GI/M/1 type* (3.3)*. Then* $\{I_n^{(0)}\}$ *is positive recurrent as well, i.e.* $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\hat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}]$ *has a solution* $\boldsymbol{\pi}_0$ *with all* $\pi_{i0} > 0$ *which is unique up to a constant. If we normalize* $\boldsymbol{\pi}_0$ *by* $1 = \boldsymbol{\pi}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1}$*, then the stationary distribution for* $\boldsymbol{Q}$ *is* $(\boldsymbol{\pi}_0\ \boldsymbol{\pi}_1\ \cdots)$ *with* $\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_0\boldsymbol{R}^\ell$, $\ell = 1, 2, \ldots$.

Proof. That $\{I_n^{(0)}\}$ is positive recurrent follows from I.3.9. The remaining statements are immediate from Lemmas 3.4 and 3.5 once one notes that the summability of $\boldsymbol{\pi}_\ell = \boldsymbol{\pi}_0\boldsymbol{R}^\ell$ and all $\pi_{i0} > 0$ implies $\mathrm{spr}(\boldsymbol{R}) < 1$ and $\sum_0^\infty \boldsymbol{R}^k = (\boldsymbol{I} - \boldsymbol{R})^{-1}$. □

It remains to evaluate $\boldsymbol{R}$. We have:

Theorem 3.7 *$\boldsymbol{R}$ is solution to $\boldsymbol{R} = \Psi(\boldsymbol{R})$ where $\Psi(\boldsymbol{R}) = \boldsymbol{F}(1) + \boldsymbol{R}\boldsymbol{F}(0) + \boldsymbol{R}^2\boldsymbol{F}(-1) + \cdots$ and the minimal nonnegative solution to this equation. Furthermore, $\boldsymbol{R}$ can be evaluated by successive iterations, say as limit of the nondecreasing sequence $\boldsymbol{R}^{(n)}$ given by $\boldsymbol{R}^{(0)} = \boldsymbol{0}$, $\boldsymbol{R}^{(n+1)} = \Psi\big(\boldsymbol{R}^{(n)}\big)$.*

Proof. Consider again an excursion away from level 0 so that $\boldsymbol{R}$ gives the average phase distribution at visits to level 1 during the excursion. Collecting phases at visits to level 1 where $L_n = k$ in the preceding step gives the contribution $\boldsymbol{R}^k\boldsymbol{F}(1-k)$ to $\boldsymbol{R}$ for $k \geq 1$ and $\boldsymbol{F}(1)$ for $k = 0$. Collecting terms yields $\boldsymbol{R} = \Psi(\boldsymbol{R})$. The rest of the proof is deferred to Section 4b where we also present other ways to evaluate $\boldsymbol{R}$ than the iterative solution. □

We state without proof the following alternative criterion for positive recurrence, which is valid also without assuming $\boldsymbol{P}$ irreducible as in Corollary 3.3 and which can be applied once $\boldsymbol{R}$ has been evaluated:

Corollary 3.8 *Suppose that $\boldsymbol{Q}$ is irreducible. Then positive recurrence is equivalent to* $\operatorname{spr}(\boldsymbol{R}) < 1$.

An analogous theory can be developed in continuous time:

Corollary 3.9 *Consider an irreducible Markov jump process $\{(J_t, L_t)\}_{t\geq 0}$ with intensity matrix $\boldsymbol{Q}$ of the form* (3.3), *and let $\Psi(\boldsymbol{R}) = \boldsymbol{F}(1) + \boldsymbol{R}\boldsymbol{F}(0) + \boldsymbol{R}^2\boldsymbol{F}(-1) + \cdots$. Then the equation $\boldsymbol{0} = \Psi(\boldsymbol{R})$ has a minimal nonnegative solution, and the process is ergodic if and only if* $\operatorname{spr}(\boldsymbol{R}) < 1$. *In that case, the stationary distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0\, \boldsymbol{\pi}_1 \cdots)$ is given by $\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\boldsymbol{R}^k$ where $\boldsymbol{\pi}_0$ solves $\boldsymbol{\pi}_0\widehat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}] = \boldsymbol{0}$ and is normalized by $\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{1} = 1$. If furthermore $\boldsymbol{\Lambda} = \sum_{-\infty}^{1}\boldsymbol{F}(k)$ is irreducible, then ergodicity of $\{(J_t, L_t)\}$ is equivalent to $\boldsymbol{\nu}\boldsymbol{\mu} < 0$ where $\boldsymbol{\nu}$ is the stationary distribution of $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu} = \sum_{-\infty}^{1} k\boldsymbol{F}(k)\boldsymbol{1}$.*

Proof. We use uniformization; cf. Problem II.4.1. Choose $a > 0$ with $a > b_{ii}(0)$, $a > |f_{ii}(0)|$ for all i. Then $\boldsymbol{Q}^* = \boldsymbol{I} + \boldsymbol{Q}/a$ is a transition matrix, $\boldsymbol{Q}^*$ is positive recurrent if and only if $\boldsymbol{Q}$ is ergodic and in that case the stationary distributions are the same; cf. again Problem II.4.1. But $\boldsymbol{Q}^*$ is of the $GI/M/1$ type corresponding to

$$\boldsymbol{F}^*(k) = \begin{cases} \boldsymbol{F}(k)/a & k \neq 0 \\ \boldsymbol{I} + \boldsymbol{F}(k)/a & k = 0 \end{cases}, \quad \boldsymbol{B}^*(k) = \begin{cases} \boldsymbol{B}(k)/a & k \neq 0 \\ \boldsymbol{I} + \boldsymbol{B}(k)/a & k = 0 \end{cases}.$$

Hence if $\boldsymbol{R}$, $\boldsymbol{\pi}_0$ are chosen as for this discrete time model,

$$\begin{aligned} \widehat{\boldsymbol{B}}^*_{\mathrm{L}}[\boldsymbol{R}] &= \boldsymbol{I} + \widehat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}]/a, \quad \boldsymbol{\pi}_0 \;=\; \boldsymbol{\pi}_0\widehat{\boldsymbol{B}}^*_{\mathrm{L}}[\boldsymbol{R}] \;=\; \boldsymbol{\pi}_0 + \boldsymbol{\pi}_0\widehat{\boldsymbol{B}}_{\mathrm{L}}[\boldsymbol{R}]/a, \\ \boldsymbol{R} &= \boldsymbol{F}^*(1) + \boldsymbol{R}\boldsymbol{F}^*(0) + \boldsymbol{R}^2\boldsymbol{F}^*(-1) + \cdots \\ &= \boldsymbol{R} + \big(\boldsymbol{F}(1) + \boldsymbol{R}\boldsymbol{F}(0) + \boldsymbol{R}^2\boldsymbol{F}(-1) + \cdots\big)/a. \end{aligned}$$

The rest of the proof is a straightforward translation of results derived above. □

Example 3.10 Let L_t be the queue length at time t in a single–server queue where the arrival and service rates at time t are β_i, δ_i when $J_t = i$; here $\{J_t\}_{t\geq 0}$ is a finite ergodic Markov process with intensity matrix $\boldsymbol{\Lambda}$ and stationary distribution $\boldsymbol{\nu}$. For obvious reasons, this system is denoted as *the* $M/M/1$ *queue in a random environment* or *the Markov–modulated* $M/M/1$ *queue* ($MM/MM/1$). Obviously, $\{(J_t, L_t)\}_{t\geq 0}$ is a Markov process with intensity matrix $\boldsymbol{Q}$ of the form (3.3), which, writing $\boldsymbol{\Delta_\beta} = (\beta_i)_{\mathrm{diag}}$, etc. corresponds to

$$\begin{aligned} \boldsymbol{F}(-1) &= \boldsymbol{\Delta_\delta},\ \ \boldsymbol{F}(1) \ = \ \boldsymbol{\Delta_\beta},\ \ \boldsymbol{F}(0) \ = \ \boldsymbol{\Lambda} - \boldsymbol{\Delta_{\beta+\delta}}, \\ \boldsymbol{B}(0) &= \boldsymbol{\Lambda} - \boldsymbol{\Delta_\beta},\ \ \boldsymbol{B}(1) \ = \ \boldsymbol{\Delta_\delta} \end{aligned}$$

(all other $\boldsymbol{F}(k)$, $\boldsymbol{B}(k)$ are zero). Hence ergodicity is equivalent to $\boldsymbol{\nu\beta} < \boldsymbol{\nu\delta}$ which is intuitive since $\boldsymbol{\nu\beta}$ is the average arrival rate and $\boldsymbol{\nu\delta}$ the average service rate. For further aspects of this model, see the discussion of QBD's below. □

3c $M/G/1$ *Type Queueing Models. QBD's*

Example 3.11 Consider the $MArP/G/1$ queue, i.e. the arrivals occur according to a Markovian point process (specified by $\boldsymbol{C}$, $\boldsymbol{D}$) and service times have some general distribution B. We define I_n as the phase of the arrival process at the nth departure and L_n as the number of customers just after.

For $k \geq -1$, define $\boldsymbol{F}(k)$, $k \geq 0$, as the (defective) transition matrix for the phase changes in a service period having $k+1$ arrivals. It is then elementary to check that $\{(I_n, L_n)\}$ has a transition matrix where rows $i = 1, 2, \ldots$ have the (spatial homogeneous) form in (3.4). Clearly,

$$\boldsymbol{F}(-1) \ = \ \int_0^\infty \mathrm{e}^{\boldsymbol{C}t}\, B(\mathrm{d}t), \quad \boldsymbol{F}(0) \ = \ \int_0^\infty B(\mathrm{d}t) \int_0^t \mathrm{e}^{\boldsymbol{C}s} \boldsymbol{D} \mathrm{e}^{\boldsymbol{C}(t-s)}\,\mathrm{d}s;$$

however, the expressions quickly become complicated and it is discussed thoroughly in, e.g., Neuts (1989) how to compute the $\boldsymbol{F}(k)$. For row 0, let as in the proof of Proposition 1.4 $\boldsymbol{P} = -\boldsymbol{C}^{-1}\boldsymbol{D}$ be the transition matrix for the sequence of phases just after arrivals. Then $\boldsymbol{P}$ also governs the transitions of phases from a departure instant until the next arrival, and using this, it is seen that row 0 corresponds to $\boldsymbol{A}(k) = \boldsymbol{PF}(k-1)$, $k = 0, 1, \ldots$. □

In the general theory of $M/G/1$ type models, a certain matrix $\boldsymbol{G}$ plays an equally fundamental role as the $\boldsymbol{R}$–matrix for $GI/M/1$ type models. It is defined in terms of the underlying MAP by $g_{ij} = \mathbb{P}_i(S_{\tau_-^{\mathrm{s}}} = j)$ where $\tau_-^{\mathrm{s}} = \inf\{n > 0 : S_n < 0\}$ is the strict descending ladder epoch. That is, $\boldsymbol{G} = \boldsymbol{G}_-^{\mathrm{s}}$ in the terminology of Chapter VIII and Section 2f. In Section 4, we prove the following result (and also present other algorithms than the iterative one):

Theorem 3.12 *$\boldsymbol{G}$ is solution to $\boldsymbol{G} = \varphi(\boldsymbol{G})$ where $\varphi(\boldsymbol{G}) = \boldsymbol{F}(-1) + \boldsymbol{F}(0)\boldsymbol{G} + \boldsymbol{F}(1)\boldsymbol{G}^2 + \cdots$, and the minimal solution to this equation. Furthermore, $\boldsymbol{G}$ can be evaluated by successive iterations, say as limit of the nondecreasing sequence $\boldsymbol{G}^{(n)}$ given by $\boldsymbol{G}^{(0)} = \boldsymbol{0}$, $\boldsymbol{G}^{(n+1)} = \varphi\big(\boldsymbol{R}^{(n)}\big)$.*

Remark 3.13 For the case $\kappa'(0) > 0$ where $\boldsymbol{G}$ is invertible, the equation $\boldsymbol{G} = \varphi(\boldsymbol{G})$ means $\widehat{\boldsymbol{F}}_{\mathrm{R}}[\boldsymbol{G}] = \boldsymbol{I}$ in the notation of (3.7). Similarly, when $\kappa'(0) < 0$ the equation $\boldsymbol{R} = \Psi(\boldsymbol{R})$ means $\widehat{\boldsymbol{F}}_{\mathrm{L}}[\boldsymbol{R}^{-1}] = \boldsymbol{I}$. The close analogy with the equations $\widehat{F}[z] = 1$ and $\widehat{F}[z^{-1}] = 1$ for simple random walks that played a fundamental role in VIII.5 supports intuitively that $\boldsymbol{G} = \varphi(\boldsymbol{G})$ and $\boldsymbol{R} = \Psi(\boldsymbol{R})$ are equally important for discrete Markov additive models. □

We will not discuss the general theory of $M/G/1$ type models any further (we refer again to Neuts, 1989), but will look at the particular case of a *quasi birth–death process* (*QBD*), defined by the requirement that (3.4) takes the form

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{F}(1) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{F}(-1) & \boldsymbol{F}(0) & \boldsymbol{F}(1) & \cdots \\ \vdots & & & & & \ddots \end{pmatrix}. \tag{3.8}$$

Since the model (3.8) is also of $GI/M/1$ type, we can immediately apply the results of Section 3b. However, the special features allow for various simplifications and alternative point of views. These involve a matrix $\boldsymbol{U}$, defined in terms of the underlying MAP by $u_{ij} = \mathbb{P}_i(J_{\theta(0)} = j, \theta(0) < \tau_-^{\mathrm{s}})$ where $\theta(\ell) = \inf\{n > 0 : S_n = \ell\}$. If any of the matrices $\boldsymbol{R}, \boldsymbol{G}, \boldsymbol{U}$ is known, the remaining two are easily computed as follows:

Proposition 3.14 *For a QBD,*

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{F}(1)(\boldsymbol{I} - \boldsymbol{U})^{-1}, & (3.9) \\ \boldsymbol{G} &= (\boldsymbol{I} - \boldsymbol{U})^{-1}\boldsymbol{F}(-1), & (3.10) \\ \boldsymbol{U} &= \boldsymbol{F}(0) + \boldsymbol{F}(1)\boldsymbol{G} \;=\; \boldsymbol{F}(0) + \boldsymbol{R}\boldsymbol{F}(-1) & (3.11) \\ &= \boldsymbol{F}(0) + \boldsymbol{F}(1)(\boldsymbol{I} - \boldsymbol{U})^{-1}\boldsymbol{F}(-1). & (3.12) \end{aligned}$$

Further, in the ergodic case $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0\boldsymbol{R}^n$ where $\boldsymbol{\pi}_0$ satisfies $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\big(\boldsymbol{C} + \boldsymbol{R}\boldsymbol{F}(-1)\big) = \boldsymbol{\pi}_0\big(\boldsymbol{C} + \boldsymbol{F}(1)\boldsymbol{G}\big)$ and $\boldsymbol{\pi}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1} = 1$. If $\boldsymbol{C} = \boldsymbol{F}(-1) + \boldsymbol{F}(0)$, then $\boldsymbol{\pi}_0 = \boldsymbol{\nu}(\boldsymbol{I} - \boldsymbol{R})$ where $\boldsymbol{\nu}$ is the stationary distribution of $\boldsymbol{P} = \boldsymbol{F}(-1) + \boldsymbol{F}(0) + \boldsymbol{F}(1)$.

Proof. Assume $S_0 = 0$. Then $\boldsymbol{U}$ is the (defective) transition matrix of $\{J_n\}$ observed at visits of $\{S_n\}$ to 0 before τ_-^{s}, and hence the ijth element of $\sum_0^\infty \boldsymbol{U}^n = (\boldsymbol{I} - \boldsymbol{U})^{-1}$ is the expected number of visits before τ_-^{s} to j in level 0 starting from $J_0 = i$. From this (3.10) follows by conditioning on $J_{\tau_-^{\rm s}-1} = j$, and the proof of (3.9) is similar. That $\boldsymbol{U} = \boldsymbol{F}(0)+\boldsymbol{F}(1)\boldsymbol{G}$ follows by conditioning upon the first transition of $\{(J_n, S_n)\}$, and inserting (3.9), (3.10) then yields first (3.12) and next the second identity in (3.11).

That $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\bigl(\boldsymbol{C} + \boldsymbol{R}\boldsymbol{F}(-1)\bigr)$ is the same as $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\widehat{\boldsymbol{B}}_{\rm L}[\boldsymbol{R}]$, cf. Theorem 3.6, and the second expression then follows from (3.11). Assume $\boldsymbol{C} = \boldsymbol{F}(-1) + \boldsymbol{F}(0)$ and let $\boldsymbol{\pi}_0 = \boldsymbol{\nu}(\boldsymbol{I} - \boldsymbol{R})$. Clearly, $\boldsymbol{\pi}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1} = 1$ and further

$$\begin{aligned}
\boldsymbol{\pi}_0\bigl(\boldsymbol{C} + \boldsymbol{R}\boldsymbol{F}(-1)\bigr) \;&=\; \boldsymbol{\nu}(\boldsymbol{I} - \boldsymbol{R})\bigl(\boldsymbol{F}(-1) + \boldsymbol{F}(0) + \boldsymbol{R}\boldsymbol{F}(-1)\bigr)\\
=\;& \boldsymbol{\nu}\bigl(\boldsymbol{F}(-1) + \boldsymbol{F}(0) + \boldsymbol{R}\boldsymbol{F}(-1) - \boldsymbol{R}\boldsymbol{F}(-1) - \boldsymbol{R}\boldsymbol{F}(0) - \boldsymbol{R}^2\boldsymbol{F}(-1)\bigr)\\
=\;& \boldsymbol{\nu}\bigl(\boldsymbol{F}(-1) + \boldsymbol{F}(0) + \boldsymbol{F}(1) - \boldsymbol{R}\bigr) \;=\; \boldsymbol{\nu}(\boldsymbol{I} - \boldsymbol{R}) \;=\; \boldsymbol{\pi}_0,
\end{aligned}$$

where we used $\boldsymbol{R} = \Psi(\boldsymbol{R}) = \boldsymbol{F}(1) + \boldsymbol{R}\boldsymbol{F}(0) + \boldsymbol{R}^2\boldsymbol{F}(-1)$ in the third step and the definition of $\boldsymbol{\nu}$ in the next. Thus $\boldsymbol{\pi}_0$ satisfies the requirements of Theorem 3.6. □

Remark 3.15 As in Corollary 3.9, it is straighforward to translate to QBD's in continuous time. One key example is the Markov–modulated $M/M/1$ queue in Example 3.10, another the $MArP/PH/1$ queue. Let the service time distribution have phase representation (F, α, T) and the parameters of the $MArP$ be $\boldsymbol{C}^A, \boldsymbol{D}^A$, and define $\{(I_t, L_t)\}$ by L_t being the number in system and I_t the current phase of service if $L_t > 0$, the phase in which the next customer will start service if $L_t = 0$. Then the intensity matrix of $\{(I_t, L_t)\}$ has the form (3.8) corresponding to $\boldsymbol{F}(-1) = \boldsymbol{I} \otimes \boldsymbol{t\alpha}$[4] (here $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{1}$ is the exit rate vector), $\boldsymbol{F}(0) = \boldsymbol{C}^A \oplus \boldsymbol{T}$, $\boldsymbol{F}(1) = \boldsymbol{D}^A \otimes \boldsymbol{I}$, $\boldsymbol{C} = \boldsymbol{C}^A \otimes \boldsymbol{I}$. □

3d The Sengupta Process

The model may be seen as a continuous–time and –state version of the $GI/M/1$ paradigm, the most apparent analogue being that paths are upward skip–free for levels.

The underlying MAP $\{(J_t, S_t)\}$ has a Markov component $\{J_t\}$, with a finite state space E and intensity matrix $\boldsymbol{\Lambda}$, and a level component $\{S_t\}$ having only downward jumps and increasing linearly in between jumps. As for the $MArP$, we decompose $\boldsymbol{\Lambda}$ as $\boldsymbol{C} + \boldsymbol{D}$ where the ijth element d_{ij} gives the intensity of a (downward) jump in phase i accompanied of a phase change to j and the elements of $\boldsymbol{C} = \boldsymbol{\Lambda} - \boldsymbol{D}$ describe phase changes without

[4] Kronecker (or tensor–) product $(\boldsymbol{A} \otimes \boldsymbol{B})_{ij,k\ell} = a_{ik}b_{j\ell}$ and Kronecker sum $\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B}$. Note that if $\{X_t^{(1)}\}$, $\{X_t^{(2)}\}$ are independent Markov processes with intensity matrices $\boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}$, then $\{(X_t^{(1)}, X_t^{(2)})\}$ has intensity matrix $\boldsymbol{\Lambda}^{(1)} \oplus \boldsymbol{\Lambda}^{(2)}$.

jump. The distribution of the size of the downward jump corresponding to a phase change $i \to j$ is denoted A_{ij} and is thus concentrated on $(0, \infty)$.

The Sengupta process $\{(I_t, L_t)\}$ evolves as $\{(J_t, S_t)\}$ on $(0, \infty)$. However. if a jump occuring from level x would have taken $\{L_t\}$ below 0, the level is reset to 0 and the phase may also be changed. Thus, the intensity of a transition $(i, x) \to (j, x - y)$ is $d_{ij}A(\mathrm{d}y)$ for $0 < y < x$ and say $q_{ij}(x)$ for $y = x$ where the $q_{ij}(x)$ satisfy $\sum_j q_{ij}(x) = \sum_j d_{ij}\overline{A}(x)$ but are otherwise arbitrary. Note that if the jump takes the level to 0, the level starts to increase right away so that there are no intervals on which the level stays at 0.

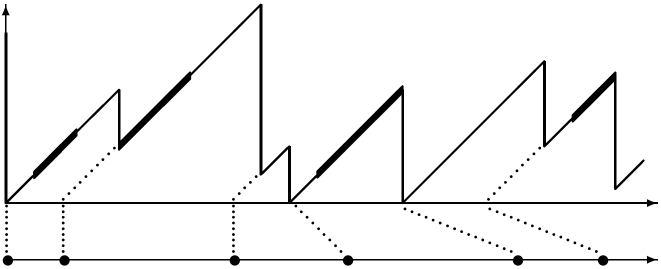

Figure 3.2

An example of a sample path with two phases (1=thin, 2=thick) is in Fig. 3.2 (upper part; the part with the broken lines refers to Example 3.17).

It is clear that if a stationary distribution of $\{(I_t, L_t)\}$ exists, the level part must be absolutely continuous. Accordingly, we let $\pi_i(x)$ denote the steady–state density of L_t restricted to $\{J_t = i\}$ and write $\boldsymbol{\pi}(x) = \big(\pi_i(x)\big)_{i \in E}$. Write further $\boldsymbol{Q}(x) = \big(q_{ij}(x)\big)_{i,j \in E}$ and $\boldsymbol{\Pi}(\mathrm{d}x) = \big(d_{ij}A_{ij}(\mathrm{d}x)\big)_{i,j \in E}$

Theorem 3.16 *Assume that $\{J_t\}$ is irreducible. Then $\{(I_t, L_t)\}$ is Harris ergodic if and only if $\kappa'(0) < 0$. Then $\boldsymbol{\pi}(x) = \boldsymbol{\pi}(0)\mathrm{e}^{\boldsymbol{T}x}$ where the $E \times E$ matrix $\boldsymbol{T}$ solves*

$$\boldsymbol{T} \;=\; \boldsymbol{C} + \int_0^\infty \mathrm{e}^{\boldsymbol{T}x}\boldsymbol{\Pi}(\mathrm{d}x) \tag{3.13}$$

and is the minimal solution subject to $\boldsymbol{T} \geq \boldsymbol{C}$. Further, $\boldsymbol{\pi}(0) = \boldsymbol{\nu}/(-\boldsymbol{\nu}\boldsymbol{T}^{-1}\boldsymbol{1})$ where $\boldsymbol{\nu}$ is the stationary vector for the transition matrix $\int_0^\infty \mathrm{e}^{\boldsymbol{T}x}\boldsymbol{Q}(x)\,\mathrm{d}x$.

Proof. The proof of the statement concerning Harris ergodicity is easy and omitted. Define the occupation measure $\boldsymbol{R}$ by

$$R_{ij}(A) \;=\; \mathbb{E}_{i,0}\int_0^{\tau_-} I(S_t \in A, J_t = j)\,\mathrm{d}t$$

where $\tau_- = \inf\{t > 0 : S_t < 0\}$. The density $r_{ij}(x)$ can also be interpreted as the expected number of upcrossings of level x in phase j before τ_-. Following each such upcrossing, the expected number of upcrossings of level $x+y$ in phase k before level x is downcrossed is $r_{jk}(y)$, so the total expected number before τ_- is $\sum_k r_{ij}(x)r_{jk}(y)$. Equating this to $r_{ij}(x+y)$ shows that $\boldsymbol{r}(x) = (r_{ij}(x))_{i,j\in E}$ satisfies $\boldsymbol{r}(x+y) = \boldsymbol{r}(x)\boldsymbol{r}(y)$ and therefore must be of the form $\mathrm{e}^{\boldsymbol{T}x}$. Taking the t with $L_t = 0$ and $I_t = i$ for some i as semi–regeneration points and appealing to VII.5.2 then gives that the steady–state density must have the form $\boldsymbol{\pi}(0)\mathrm{e}^{\boldsymbol{T}x}$ for some $\boldsymbol{\pi}(0)$.

To derive (3.13), we divide possible upcrossings of level h in phase j before τ_- into three types, the one occuring at time h without a jump of the level, the ones following a jump from some level $y > h$ to $(0, h]$ and the rest. Starting from $I_0 = i$, the expected number is $\delta_{ij} + c_{ij}h + \mathrm{o}(h)$ for the first type, $\mathrm{o}(h)$ for the third type and (interpreting $\boldsymbol{R}(\mathrm{d}x) = \mathrm{e}^{\boldsymbol{T}x}\mathrm{d}x$ as an occupation measure)

$$\begin{aligned}
&\sum_{k\in E}\int_h^\infty (\mathrm{e}^{\boldsymbol{T}y})_{ik}\,\mathrm{d}y\,\cdot\, d_{kj}\int_{y-h}^{y} A_{kj}(\mathrm{d}x) \;+\; \mathrm{o}(h)\\
&=\;\sum_{k\in E}\int_0^\infty\int_x^{x+h}(\mathrm{e}^{\boldsymbol{T}y})_{ik}\,\mathrm{d}y\,\Pi_{kj}(\mathrm{d}x) \;+\; \mathrm{o}(h)\\
&=\; h\sum_{k\in E}\int_0^\infty (\mathrm{e}^{\boldsymbol{T}x})_{ik}\,\Pi_{kj}(\mathrm{d}x) \;+\; \mathrm{o}(h)
\end{aligned}$$

for the second type. Thus

$$r_{ij}(h) \;=\; \delta_{ij} + c_{ij}h \,+\, h\sum_{k\in E}\int_0^\infty (\mathrm{e}^{\boldsymbol{T}x})_{ik}\,\Pi_{kj}(\mathrm{d}x) \,+\, \mathrm{o}(h).$$

On the other hand, $r_{ij}(h) = (\mathrm{e}^{\boldsymbol{T}h})_{ij} = \delta_{ij} + t_{ij}h + \mathrm{o}(h)$, and equating the two expressions gives (3.13).

Finally $\boldsymbol{\pi}(0)$ must have the form $c\boldsymbol{\nu}$ where $\boldsymbol{\nu}$ is the stationary distribution for $\{I_t\}$ observed at semi–regeneration points. That this Markov chain has transition matrix $\int_0^\infty \mathrm{e}^{\boldsymbol{T}x}\boldsymbol{Q}(x)\,\mathrm{d}x$ is easily seen, and $-\boldsymbol{\pi}(0)\boldsymbol{T}^{-1}\boldsymbol{1}=$ $\int_0^\infty \boldsymbol{\pi}(0)\mathrm{e}^{\boldsymbol{T}x}\boldsymbol{1}\,\mathrm{d}x = 1$ shows that $c = 1/(-\boldsymbol{\nu}\boldsymbol{T}^{-1}\boldsymbol{1})$. □

Example 3.17 Consider the $GI/PH/1$ queue with interarrival distribution A and representation $(E, \boldsymbol{\alpha}, \boldsymbol{S})$ of the service time distribution. We define $\{(I_t, L_t)\}$ by cutting idle periods out and letting I_t be the phase in which the server is currently operating, L_t the time that passed since the customer in service arrived. See Fig. 3.2 where a customer has arrived at an idle system at $t = 0$ and the bullets on the lower t–axis indicate arrival times. Thus, the jumps are the interarrival times and with $\boldsymbol{s} = -\boldsymbol{S}\boldsymbol{1}$ the exit rate vector of the service time distribution, we have $A_{ij} = A$, $\boldsymbol{C} = \boldsymbol{S}$, $d_{ij} = s_i\alpha_j$ and $q_{ij}(x) = s_i\alpha_j\overline{A}(x)$.

Defining W_n as the value of $\{L_t\}$ just after the nth jump, the graphical representation shows that W_n is simply the waiting time of customer n. We will show that that the steady–state distribution is given by

$$\mathbb{P}_e(W_n \le x) \;=\; 1 - \mu_B\boldsymbol{\beta}\mathrm{e}^{\boldsymbol{T}x}(\boldsymbol{T}-\boldsymbol{S})\boldsymbol{1}, \tag{3.14}$$

where $\boldsymbol{\beta} = -\boldsymbol{\alpha}\boldsymbol{S}^{-1}/\mu_B$ with $\mu_B = -\boldsymbol{\alpha}\boldsymbol{S}^{-1}\mathbf{1}$ the mean service time. To this end, note first that $L_t = 0$ means that a busy period has just started so that the server has chosen his phase according to $\boldsymbol{\alpha}$. Thus the Markov chain having stationary distribution $\boldsymbol{\nu}$ is just i.i.d. replicates from $\boldsymbol{\alpha}$ so that $\boldsymbol{\nu} = \boldsymbol{\alpha}$ and $\boldsymbol{\pi}(0) = -\boldsymbol{\alpha}/(\boldsymbol{\alpha}\boldsymbol{T}^{-1}\mathbf{1})$. Marginally, $\{I_t\}$ is the Markov process obtained from the sequence of service time by piecing the terminating Markov processes corresponding to individual service times together as in III.5, so that by results given there $\boldsymbol{\beta}$ is the stationary distribution of $\{I_t\}$. On the other hand, this stationary distribution is $\int_0^\infty \boldsymbol{\pi}(x)\,\mathrm{d}x = -\boldsymbol{\pi}(0)\boldsymbol{T}^{-1}$ so that $\boldsymbol{\pi}(0) = -\boldsymbol{\beta}\boldsymbol{T}$.

By Palm theory, $\mathbb{P}_e(W_n \le x)$ has the form $cu(x)$ where $u(x)$ is the steady–state rate of a jump of $\{L_t\}$ terminating in $[0,x]$. Here

$$\begin{aligned} u(x) &= \int_0^\infty A(\mathrm{d}y)\int_0^{x+y}\boldsymbol{\pi}(z)\boldsymbol{s}\,\mathrm{d}z \;=\; \int_0^\infty A(\mathrm{d}y)\int_0^{x+y}\boldsymbol{\pi}(0)\mathrm{e}^{\boldsymbol{T}z}\boldsymbol{s}\,\mathrm{d}z \\ &= -\int_0^\infty A(\mathrm{d}y)\int_0^{x+y}\boldsymbol{\beta}\boldsymbol{T}\mathrm{e}^{\boldsymbol{T}z}\boldsymbol{s}\,\mathrm{d}z \;=\; \int_0^\infty \boldsymbol{\beta}\big(\boldsymbol{I}-\mathrm{e}^{\boldsymbol{T}(x+y)}\big)\boldsymbol{s}A(\mathrm{d}y) \\ &= \boldsymbol{\beta}\Big(\boldsymbol{I}-\mathrm{e}^{\boldsymbol{T}x}\int_0^\infty \mathrm{e}^{\boldsymbol{T}y}A(\mathrm{d}y)\Big)\boldsymbol{s} \;=\; \boldsymbol{\beta}\boldsymbol{s}-\boldsymbol{\beta}\mathrm{e}^{\boldsymbol{T}x}(\boldsymbol{T}-\boldsymbol{S})\mathbf{1}, \end{aligned}$$

where the last equality followed by postmultiplying by $1 = \boldsymbol{\alpha}\mathbf{1}$ and using $\boldsymbol{\Pi}(\mathrm{d}x) = \boldsymbol{s}\boldsymbol{\alpha}A(\mathrm{d}x)$ and (3.13). Here $\boldsymbol{\beta}\boldsymbol{s} = \boldsymbol{\alpha}\boldsymbol{S}^{-1}\boldsymbol{S}\mathbf{1}/\mu_B = 1/\mu_B$. Further, $\boldsymbol{\pi}(x) \to \mathbf{0}$, $x \to \infty$, implies $\mathrm{e}^{\boldsymbol{T}x} \to \mathbf{0}$. Since $\mathbb{P}_e(W_n \le x) \to 1$, we get $c = \mu_B$ and (3.14) follows. □

Example 3.18 Consider the *FCFS GI/PH/s* queue with interarrival distribution A and representation $(F, \boldsymbol{\alpha}, \boldsymbol{S})$ of the service time distribution. By a *non–all–busy period* we understand a time interval where at least one server is idle. We define $\{(I_t, L_t)\}$ by cutting out non–all–busy periods and letting $E = F^s$, $J_t \in E$ the set of phases in which the s servers are currently operating, L_t the time which passed since the last customer to enter server service arrived.

The same graphical representation as in Fig. 3.2 applies, so that W_n is the waiting time of customer n and the jumps are the interarrival times. Thus, with $\boldsymbol{s} = -\boldsymbol{S}\mathbf{1}$ the exit rate vector of the service time distribution, we have $A_{ij} = A$, $\boldsymbol{C} = \boldsymbol{S}\oplus\cdots\oplus\boldsymbol{S}$ (Kronecker sum; see footnote to Remark 3.15) and $d_{i_1\ldots i_s,j_1\ldots j_s} = s_{i_k}\alpha_{j_k}$ if $i_\ell = j_\ell$ for $\ell \ne k$ ($d_{i_1\ldots i_s,j_1\ldots j_s} = 0$ if $i_k \ne j_k$ for at least two k). The form of

$$q_{i_1\ldots i_s,j_1\ldots j_s}(x) \;=\; \overline{A}(x)\sum_{k=1}^{s} s_{i_k}p_{i_1\ldots i_{k-1}i_{k+1}\ldots i_s,j_1\ldots j_s}(x),$$

where $p_{\ldots}(x)$ is the probability that a non–all–busy period, initiated at residual interarrival time x with server k being idle and with server $\ell \ne k$ working in phase i_ℓ, is followed by an all–busy period with server m working in phase j_m, $m = 1,\ldots,s$, is, however, more complicated than when $s = 1$.

The representation allows in a straightforward way to deal with heterogeneous phase–type servers. If, as above, the servers are homogeneous, the dimension of E may be greatly reduced by instead letting $E = \{n_1 \ldots n_p : n_1 + \cdots + n_p = s\}$ where $p = |F|$ is the number of phases for the service time distribution and $n_r \in \{0, \ldots, s\}$ gives the number of servers operating in phase r. □

Notes The classical sources for $GI/M/1$ and $M/G/1$ type models are Neuts (1981), resp. Neuts (1989); for QBD's, a more comprehensive and up–to–date treatment is in Latouche and Ramaswami (1999). There is also much material in the series of volumes edited by Alfa and Chakravarty (1997, 1998) and Latouche and Taylor (2000, 2002).

The literature makes little use of the Markov additive point of view we have taken here and which is also reflected in our notation, which is logical for MAPs but is not the one in common use (see, however, the footnote on p. 130 in Latouche and Ramaswami, 1999).

The model of Section 3d was introduced in Sengupta (1989) who also derived the $GI/PH/1$ waiting time distribution via Example 3.17 (for different approaches, see VIII.5 and references there). For the $GI/PH/s$ waiting time via Example 3.18, see Asmussen and O'Cinneide (1998) and Asmussen and Møller (2001).

4 Solution Methods

4a. Martingale Calculations
4b. Iterative Solutions
4c. The Spectral Solution

We will concentrate on two examples: the stationary distribution of a reflected Markovian fluid, and evaluation of the $\boldsymbol{R}$–matrix for a $GI/M/1$ type model and the $\boldsymbol{G}$–matrix for a $M/G/1$ type model; the Sengupta model can be treated by entirely similar methods as the ones we present, but we omit the details.

In view of Proposition 1.12, algorithms for the stationary distribution of the fluid give also the stationary distribution of the workload in the $MArP/PH/1$ queue. After some translation, IX.2.7 shows that what we must find are ways to evaluate the ruin probability $\psi_i(x) = \mathbb{P}_i(\tau(x) < \infty)$ where $\{(J_t, S_t)\}$ is an unrestricted Markovian fluid and $\tau(x) = \inf\{t \geq 0 : S_t = x\}$.

For the $GI/M/1$ type model, it follows by Proposition 2.13 that

$$\boldsymbol{R} \;=\; \Delta_{\boldsymbol{\nu}}^{-1} \boldsymbol{G}_+^{*\mathsf{T}} \Delta_{\boldsymbol{\nu}} \tag{4.1}$$

where $\boldsymbol{G}_+^*$ is the strict ascending ladder height kernel of the time–reversed MAP. For the $M/G/1$ type model, $\boldsymbol{G}$ is the strict descending ladder height kernel, which is the same as the strict ascending ladder height kernel of the sign–reversed model. Thus, if we are able to calculate the strict ascending ladder height kernel of an upward skip–free MAP, we also have the necessary machinery to compute the $\boldsymbol{R}$– and $\boldsymbol{G}$–matrices.

4a Martingale Calculations

We write the state space E of $\{J_t\}$ in the fluid model as the disjoint union of E_-, the states with $r(i) < 0$ and E_+, the states with $r(i) > 0$ (for convenience, we assume that no $r(i) = 0$), and use corresponding block notation. We let q_- be the number of states of E_- and q_+ the number of states in E_+ and denote by $\boldsymbol{\Lambda}$ the intensity matrix of $\{J_t\}$, by $\boldsymbol{\Delta_r}$ the diagonal matrix with the $r(i)$ on the diagonal.

Proposition 4.1 *Let s be an eigenvalue of $\boldsymbol{\Delta_r^{-1}\Lambda}$ and $\boldsymbol{b}$ a corresponding right eigenvector. Then $\left\{\mathrm{e}^{-sS_t}b_{J_t}\right\}$ is a martingale.*
[Note that s and hence the martingale may be complex. Note also that since S_t is a bounded r.v., integrability is automatic.]

Proof. In the notation of Section 2b, $\boldsymbol{K}[\alpha] = \boldsymbol{\Lambda}+\alpha\boldsymbol{\Delta_r}$. Therefore $\boldsymbol{\Delta_r^{-1}\Lambda b} = s\boldsymbol{b}$ implies that 0 is an eigenvalue of $\boldsymbol{K}[-s]$ and $\boldsymbol{b}$ a corresponding right eigenvector. The rest of the proof is just as for Proposition 2.4. □

Theorem 4.2 *Consider a Markovian fluid with $\kappa'(0) < 0$, assume that $\boldsymbol{\Delta}_r^{-1}\boldsymbol{\Lambda}$ has q_+ distinct eigenvalues $s_1, \ldots, s_{q_+}$ with $\Re(s_\nu) < 0$ and let $\boldsymbol{b}^{(\nu)} = \left(\boldsymbol{c}^{(\nu)\mathsf{T}}\ \boldsymbol{d}^{(\nu)\mathsf{T}}\right)^{\mathsf{T}}$ be the right eigenvector corresponding to s_ν, $\nu = 1, \ldots, q_+$. Then*

$$\psi_i(x) \;=\; \boldsymbol{1}_i' \left(\mathrm{e}^{s_1 x}\boldsymbol{b}^{(1)} \cdots\ \mathrm{e}^{s_{q_+} x}\boldsymbol{b}^{(q_+)}\right)\left(\boldsymbol{d}^{(1)} \cdots\ \boldsymbol{d}^{(q_+)}\right)^{-1}\boldsymbol{1}.$$

Proof. For $x, y > 0$, define

$$\begin{aligned}
\omega(x,y) &= \inf\bigl\{t > 0 : S_t = x \text{ or } S_t = -y\bigr\},\\
p_i(x,y;j) &= \mathbb{P}_i\bigl(S_{\omega(x,y)} = x, J_{\omega(x,y)} = j\bigr),\quad j \in E_+,\\
r_i(x,y;j) &= \mathbb{P}_i\bigl(S_{\omega(x,y)} = -y, J_{\omega(x,y)} = j\bigr),\quad j \in E_-,\\
p_i(x;j) &= \mathbb{P}_i\bigl(\tau(x) < \infty, J_{\tau(x)} = j\bigr),\quad j \in E_+\,.
\end{aligned}$$

[Note that if $S_{\omega(x,y)} = x$, then the sample path must be increasing at time $\omega(x,y)$ and hence $J_{\omega(x,y)} \in E_+$; similarly, $S_{\omega(x,y)} = -y$ implies $J_{\omega(x,y)} \in E_-$.] Optional stopping of $\left\{\mathrm{e}^{-s_\nu S_t}b_{J_t}^{(\nu)}\right\}$ at time $\omega(x,y)$ yields

$$b_i^{(\nu)} \;=\; e^{-s_\nu x}\sum_{j\in E_+} p_i(x,y;j)d_j^{(\nu)} + e^{s_\nu y}\sum_{j\in E_-} r_i(x,y;j)c_j^{(\nu)}.$$

Letting $y \to \infty$ and using $\Re(s_\nu) < 0$ yields

$$\mathrm{e}^{s_\nu x}b_i^{(\nu)} \;=\; \sum_{j\in E_+} p_i(x;j)d_j^{(\nu)},\quad j \in E_+.$$

Solving for the $p_i(x;j)$ and noting that $\psi_i(x) = \sum_{j\in E_+} p_i(x;j)$, the result follows. □

Remark 4.3 Write $\boldsymbol{\Sigma} = \boldsymbol{\Delta}_r^{-1}\boldsymbol{\Lambda}$. Then s is an eigenvalue of $\boldsymbol{\Delta}_r^{-1}\boldsymbol{\Lambda}$ if and only if s is a root of the characteristic polynomial $\det(\boldsymbol{\Sigma}-s\boldsymbol{I}) = 0$. However, by a general matrix identity familiar from the theory of the multivariate normal distribution, we can write $\det(\boldsymbol{\Sigma}-s\boldsymbol{I})$ as

$$\det(\boldsymbol{\Sigma}_{--} - s\boldsymbol{I}) \cdot \det\bigl(\boldsymbol{\Sigma}_{++} - s\boldsymbol{I} - \boldsymbol{\Sigma}_{+-}(\boldsymbol{\Sigma}_{--} - s\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_{-+}\bigr), \tag{4.2}$$

where $\boldsymbol{\Sigma}_{--}$ is the upper left block in the decomposition of $\boldsymbol{\Sigma}$ according to E_-, E_+, etc. so that s must be root in one of the two determinants on the r.h.s. (the first is a polynomial and the next a ratio between two polynomials). We will see in connection with the spectral method that it is in fact only the roots of the second determinant that give the needed roots with $\Re(s) < 0$. □

A variant of the argument produces the ladder height distributions needed for the $GI/M/1$ and $M/G/1$ type models. In a general Markov additive setting, upward skip–freeness means that $\boldsymbol{F}(2) = \boldsymbol{F}(3) = \cdots = \boldsymbol{0}$ in (3.1). We write q for the number of elements of E, and identify $\boldsymbol{G}_+$ with the matrix with ijth element $g_+(i,j) = G_+(i,j;\{1\})$.

Theorem 4.4 *Consider an upward skip–free discrete–time lattice MAP with $\kappa'(0) = \boldsymbol{\nu\mu} < 0$, let $\widehat{\boldsymbol{F}}[z] = \sum_{-\infty}^{1} z^k\boldsymbol{F}(k)$, assume that there exist q complex numbers $z_1,\ldots,z_q$ with $|\Re(z_\nu)| > 1$ such that 1 is an eigenvalue of $\widehat{\boldsymbol{F}}[z_\nu]$ and let $\boldsymbol{b}^{(\nu)}$ be the right eigenvector corresponding to z_ν, $\nu = 1,\ldots,q$. Then*

$$\boldsymbol{G}_+ \;=\; \Bigl(\boldsymbol{b}^{(1)}\cdots\;\boldsymbol{b}^{(q)}\Bigr)\Bigl(z_1\boldsymbol{b}^{(1)}\cdots\;z_q\boldsymbol{b}^{(q)}\Bigr)^{-1}.$$

If instead $\kappa'(0) \ge 0$, the same formula holds except that one should look for roots with $\bigl|\Re(z_\nu)\bigr| \ge 1$.

Proof. For $m > 0$, $j \in E$ define

$$\begin{aligned}\omega(m) &= \inf\bigl\{n > 0 : S_n = 1 \text{ or } S_n = -m\bigr\},\\ p_i(m;j) &= \mathbb{P}_i\bigl(S_{\omega(m)} = 1, J_{\omega(m)} = j\bigr),\\ r_i(m;j) &= \mathbb{P}_i\bigl(S_{\omega(m)} = -m, J_{\omega(m)} = j\bigr).\end{aligned}$$

As in Proposition 4.1, it is easily seen that $\bigl\{z_\nu^{S_n} b_{J_n}^{(\nu)}\bigr\}$ is a martingale (integrability follows from $S_n \le n$ and $|\Re(z_\nu)| > 1$). Optional stopping at time $\omega(m)$ yields

$$b_i^{(\nu)} \;=\; z_\nu \sum_{j\in E} p_i(m;j) b_j^{(\nu)} + z_\nu^{-m} \sum_{j\in E} r_i(m;j) b_j^{(\nu)}. \tag{4.3}$$

Letting $m \to \infty$ and using $|\Re(z_\nu)| > 1$ yields

$$b_i^{(\nu)} = z_\nu \sum_{j \in E} g_+(i,j) b_j^{(\nu)} .$$

Solving for the $g_+(i,j)$, the result follows in the case $\kappa'(0) < 0$.

The only difference when $\kappa'(0) \geq 0$ is that we also need to deal with the case $|\Re(z_\nu)| = 1$ in (4.3). However, $r_i(m;j) \to 0$, $m \to \infty$, when $\kappa'(0) \geq 0$. □

Corollary 4.5 *Consider a $GI/M/1$ type Markov chain with $\kappa'(0) = \boldsymbol{\nu\mu} < 0$, assume that there exist q complex numbers $z_1, \ldots, z_q$ with $|\Re(z_\nu)| > 1$ such that 1 is an eigenvalue of $\widehat{\boldsymbol{F}}[z_\nu]$ and let $\boldsymbol{\eta}^{(\nu)}$ be the corresponding left eigenvector. Then*

$$\boldsymbol{R} = \begin{pmatrix} z_1 \boldsymbol{\eta}^{(1)} \\ \vdots \\ z_q \boldsymbol{\eta}^{(q)} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\eta}^{(1)} \\ \vdots \\ \boldsymbol{\eta}^{(q)} \end{pmatrix} .$$

Proof. The analogue of $\widehat{\boldsymbol{F}}[z]$ for the time-reversed MAP is $\widehat{\boldsymbol{F}}^*[z] = \boldsymbol{\Delta}_{\boldsymbol{\nu}}^{-1} \widehat{\boldsymbol{F}}[z]^{\mathsf{T}} \boldsymbol{\Delta}_{\boldsymbol{\nu}}$. Thus the roots of $\det(\hat{\boldsymbol{F}}[z] - \boldsymbol{I}) = 1$ and $\det(\widehat{\boldsymbol{F}}^*[z] - \boldsymbol{I}) = 1$ are the same, and if $\boldsymbol{b}^{(\nu)^*}$ is the right eigenvector of $\widehat{\boldsymbol{F}}^*[z_\nu]$, then we may take $\boldsymbol{b}^{(\nu)^*} = \boldsymbol{\Delta}_{\boldsymbol{\nu}}^{-1} \boldsymbol{\eta}^{(\nu)^{\mathsf{T}}}$. The result then follows after simple algebra by computing $\boldsymbol{G}_+^*$ via Theorem 4.4 and using (4.1). □

Corollary 4.6 *Consider an $M/G/1$ type Markov chain with $\kappa'(0) = \boldsymbol{\nu\mu} \leq 0$, let $\widehat{\boldsymbol{F}}[z] = \sum_{-1}^{\infty} z^k \boldsymbol{F}(k)$, assume that there exist q complex numbers $z_1, \ldots, z_q$ with $|\Re(z_\nu)| \geq 1$ such that 1 is an eigenvalue of $\hat{\boldsymbol{F}}[z_\nu^{-1}]$ and let $\boldsymbol{b}^{(\nu)}$ be the corresponding right eigenvector. Then*

$$\boldsymbol{G} = \left(\boldsymbol{b}^{(1)} \ldots \boldsymbol{b}^{(q)}\right) \left(z_1 \boldsymbol{b}^{(1)} \ldots z_q \boldsymbol{b}^{(q)}\right)^{-1} .$$

Proof. By sign–reversion in the last part of Theorem 4.4. □

4b Iterative Solutions

We first complete:

Proof of Theorem 3.12. Recall that

$$\varphi(\boldsymbol{G}) = \boldsymbol{F}(-1) + \boldsymbol{F}(0)\boldsymbol{G} + \boldsymbol{F}(1)\boldsymbol{G}^2 + \cdots, \quad \boldsymbol{G}^{(0)} = \boldsymbol{0},\ \boldsymbol{G}^{(n+1)} = \varphi\bigl(\boldsymbol{G}^{(n)}\bigr).$$

Here $\boldsymbol{F}(-1)$ is the contribution to $\boldsymbol{G}$ coming from the event $\{S_1 = -1\}$. Similarly, if $S_1 = k > -1$, then to reach level -1 the process has to move down $k+1$ steps. The transitions in each step are governed by $\boldsymbol{G}$, so that the contribution to $\boldsymbol{G}$ coming from the event $\{S_1 = k\}$ is $\boldsymbol{F}(k)\boldsymbol{G}^{k+1}$. Summing over k gives $\varphi(\boldsymbol{G}) = \boldsymbol{G}$.

Clearly, $\boldsymbol{G}^{(1)} = \boldsymbol{F}(-1) \geq \boldsymbol{0} = \boldsymbol{G}^{(0)}$ and $\boldsymbol{G} \geq \boldsymbol{0} = \boldsymbol{G}^{(0)}$. Hence by induction,

$$\begin{aligned}\boldsymbol{G}^{(n+1)} - \boldsymbol{G}^{(n)} &= \sum_{k=-1}^{\infty} \boldsymbol{F}(k) \left({\boldsymbol{G}^{(n)}}^{k+1} - {\boldsymbol{G}^{(n-1)}}^{k+1} \right) \geq \boldsymbol{0}, \\ \boldsymbol{G} - \boldsymbol{G}^{(n)} &= \sum_{k=-1}^{\infty} \boldsymbol{F}(k) \left(\boldsymbol{G}^{k+1} - {\boldsymbol{G}^{(n-1)}}^{k+1} \right) \geq \boldsymbol{0}\end{aligned}$$

so that $\boldsymbol{G}^{(n)}$ is increasing in n and a limit, say $\boldsymbol{L}$, must exist and satisfy $\boldsymbol{L} \leq \boldsymbol{G}$, $\varphi(\boldsymbol{L}) = \boldsymbol{L}$. Repeating the argument with another solution instead of $\boldsymbol{G}$ shows that $\boldsymbol{L}$ must be the minimal solution, and it only remains to show that $\boldsymbol{L} \geq \boldsymbol{G}$.

Define F_n as the event that $\{S_n\}$ has at most n transitions of the form $\ell \to \{\ell, \ell+1, \ldots\}$ before τ_-^{s} and $\widetilde{\boldsymbol{G}}^{(n)} = \mathbb{P}_i(J_{\tau_-^{\mathrm{s}}} = j;\, F_n)$. Then $\widetilde{\boldsymbol{G}}^{(n+1)} \leq \varphi\big(\widetilde{\boldsymbol{G}}^{(n)}\big)$. Indeed, if F_{n+1} occurs and $S_1 = k \geq 0$, then in the $k+1$ ladder steps needed to go to level -1 there can be at most n transitions of the type considered since there was already one in the first step. Since $\widetilde{\boldsymbol{G}}^{(0)} = \boldsymbol{F}(-1) = \boldsymbol{G}^{(1)}$, it follows by induction that $\widetilde{\boldsymbol{G}}^{(n+1)} \leq \varphi\big(\boldsymbol{G}^{(n)}\big) \leq \varphi\big(\boldsymbol{G}^{(n+1)}\big) = \boldsymbol{G}^{(n+2)}$. Since $\widetilde{\boldsymbol{G}}^{(n)} \uparrow \boldsymbol{G}$, $n \to \infty$, it follows that $\boldsymbol{G} \leq \boldsymbol{L}$. □

Proof of $\boldsymbol{R}^{(n)} \to \boldsymbol{R}$ in Proposition 3.7. Note that ${\boldsymbol{G}_+^{(n)}}^* \to \boldsymbol{G}_+^*$ by Theorem 3.12 and appeal to (4.1). □

Turning to *QBD*'s, Proposition 3.14 suggests the following alternative algorithm (sometimes called the *linear progression algorithm*) for computing $\boldsymbol{G}$ (and thereby $\boldsymbol{R}$ and the stationary distribution):

Proposition 4.7 *Define $\boldsymbol{G}^{(0)} = \boldsymbol{0}$, $\boldsymbol{U}^{(k)} = \boldsymbol{F}(0) + \boldsymbol{F}(1)\boldsymbol{G}^{(k-1)}$, $\boldsymbol{G}^{(k)} = \big(\boldsymbol{I} - \boldsymbol{U}^{(k)}\big)^{-1}\boldsymbol{F}(-1)$. Then $\boldsymbol{G}^{(k)} \uparrow \boldsymbol{G}$, $k \to \infty$. Further in the positive recurrent case $\kappa'(0) = \boldsymbol{\nu\mu} < 0$, $\boldsymbol{G} - \boldsymbol{G}^{(k)}$ is of order* $\mathrm{e}^{-\gamma k}$, *where γ is the strictly positive solution of $\kappa(\gamma) = 0$. That is, $\gamma = \log z$ where $z > 1$ solves* $\mathrm{spr}\big(z^{-1}\boldsymbol{F}(-1) + \boldsymbol{F}(0) + z\boldsymbol{F}(1)\big) = 1$.

Proof. Define $\theta(k) = \inf\{n > 0 : S_n = k\}$, $k \in \mathbb{Z}$. It is readily checked that

$$\begin{aligned} u_{ij}^{(k)} &= \mathbb{P}_{i,0}\big(J_{\theta(0)} = j, \theta(0) < \theta(-1) \wedge \theta(k)\big), \\ g_{ij}^{(k)} &= \mathbb{P}_{i,0}\big(J_{\theta(-1)} = j, \theta(-1) < \theta(k)\big) \end{aligned}$$

for $k = 1$, and the validity for $k > 1$ is then a straightforward induction argument. Letting $k \to \infty$ yields $\boldsymbol{G}^{(k)} \uparrow \boldsymbol{G}$.

For the second part, just note that $(\boldsymbol{G} - \boldsymbol{G}^{(k)})_{ij} = \mathbb{P}_{i,0}\big(J_{\theta(-1)} = j, \theta(k) < \theta(-1)\big)$ and that $\mathbb{P}_{i,0}\big(\theta(k) < \theta(-1)\big)$ is shown to be of order $\mathrm{e}^{-\gamma k}$ in XIII.8.3. □

The following remarkable algorithm reduces improves the convergence rate from $\mathrm{O}(\mathrm{e}^{-\gamma k})$ to $\mathrm{O}(\mathrm{e}^{-\gamma 2^k})$ and is therefore referred to as the *logarithmic reduction algorithm* (requiring only of order $\log_2 N$ iterations to obtain the same accuracy as the linear progression algorithm with N iterations):

Proposition 4.8 *Define* $\boldsymbol{\Pi}^{[0]} = \boldsymbol{I}$, $\boldsymbol{G}^{[0]} = \boldsymbol{L}^{[0]}$, $\boldsymbol{H}^{[0]} = \big(\boldsymbol{I} - \boldsymbol{F}(0)\big)^{-1}\boldsymbol{F}(1)$, $\boldsymbol{L}^{[0]} = \big(\boldsymbol{I} - \boldsymbol{F}(0)\big)^{-1}\boldsymbol{F}(-1)$, *and, for* $k \geq 0$,

$$\begin{aligned}
\boldsymbol{U}^{[k]} &= \boldsymbol{H}^{[k]}\boldsymbol{L}^{[k]} + \boldsymbol{L}^{[k]}\boldsymbol{H}^{[k]}, && (4.4)\\
\boldsymbol{H}^{[k+1]} &= \big(\boldsymbol{I} - \boldsymbol{U}^{[k]}\big)^{-1}\boldsymbol{H}^{[k]2}, && (4.5)\\
\boldsymbol{L}^{[k+1]} &= \big(\boldsymbol{I} - \boldsymbol{U}^{[k]}\big)^{-1}\boldsymbol{L}^{[k]2}, && (4.6)\\
\boldsymbol{\Pi}^{[k+1]} &= \boldsymbol{\Pi}^{[k]}\boldsymbol{H}^{[k]}, && (4.7)\\
\boldsymbol{G}^{[k+1]} &= \boldsymbol{G}^{[k]} + \boldsymbol{\Pi}^{[k+1]}\boldsymbol{L}^{[k+1]} && (4.8)
\end{aligned}$$

Then $\boldsymbol{G}^{[k]} \uparrow \boldsymbol{G}$, $k \to \infty$, *and in the positive recurrent case* $\boldsymbol{G} - \boldsymbol{G}^{(k)}$ *is of order* $\mathrm{e}^{-\gamma 2^k}$.

Proof. The key idea is to let $M = \max_{n<\theta(-1)} S_n$ and decompose $\boldsymbol{G}$ according to the possibilities $M = 0$, $1 \leq M < 3$, $3 \leq M < 7, \ldots$, and to get expressions for the corresponding terms in terms of matrices having probabilistic interpretations for the MAP's $\big\{(J_n^{[k]}, S_n^{[k]})\big\}$ where the kth is defined as $\{(J_n, S_n)\}$ restricted to levels that are multiples of 2^k (i.e. values of $\{(J_n, S_n)\}$ where S_n is not such a multiple are cancelled).

We will first show that the interpretation of the matrices $\boldsymbol{U}^{[k]}, \boldsymbol{L}^{[k]}, \boldsymbol{H}^{[k]}$ is

$$\begin{aligned}
h_{ij}^{[k]} &= \mathbb{P}_{i,0}\big(J_{\theta(2^k)} = j, \theta(2^k) < \theta(-2^k)\big), && (4.9)\\
\ell_{ij}^{[k]} &= \mathbb{P}_{i,0}\big(J_{\theta(-2^k)} = j, \theta(-2^k) < \theta(2^k)\big), && (4.10)\\
u_{ij}^{[k]} &= \mathbb{P}_{i,0}\big(J_{\theta(0)} = j, \theta(2^k) < \theta(0) < \theta(2^{k+1}) \text{ or}\\
&\qquad\qquad \theta(-2^k) < \theta(0) < \theta(-2^{k+1})\big). && (4.11)
\end{aligned}$$

Indeed, given that (4.9) and (4.10) have been shown, (4.11) follows by noting that the first term in (4.4) corresponds to $\theta(2^k) < \theta(0) < \theta(2^{k+1})$ and the second to $\theta(-2^k) < \theta(0) < \theta(-2^{k+1})$. To get (4.9) and (4.10), we use induction and note that the validity in the case $k = 0$ is seen by arguments similar to the identification of $\boldsymbol{U}$ in the proof of Proposition 3.14. If (4.9) and (4.10) and hence (4.11) have been shown for k, then in order for $\theta(2^{k+1}) < \theta(-2^{k+1})$ to occur, there must first be a number of tours of the level taking us to either 2^k and then back to 0 before 2^{k+1}, or to -2^k and then back to 0 before -2^{k+1}. By (4.11), these change the phase according to $\big(\boldsymbol{I} - \boldsymbol{U}^{[k]}\big)^{-1}$, and we must then first go to 2^k before -2^k and from there to 2^{k+1} before 0, which gives a phase change governed by $\boldsymbol{H}^{[k]}$

in both steps. This shows that (4.9) is the same as (4.5), and the case of (4.10) is symmetric.

The contribution to $\boldsymbol{G}$ from $M = 0$ is $\boldsymbol{L}^{[0]}$. If $1 \leq M < 3$, we must first go to 1 before -1 (phase change governed by $\boldsymbol{H}^{[0]}$) and next to -1 before 3 (phase change governed by $\boldsymbol{L}^{[1]}$), giving the contribution $\boldsymbol{H}^{[0]}\boldsymbol{L}^{[1]}$ to G. If $3 \leq M < 7$, we must first go to 1 before -1, then to 3 before -1 (phase change governed by $\boldsymbol{H}^{[1]}$) and next to -1 before 7 (phase change governed by $\boldsymbol{L}^{[2]}$), giving the contribution $\boldsymbol{H}^{[0]}\boldsymbol{H}^{[1]}\boldsymbol{L}^{[1]}$ to G. Continuing in this manner we obtain

$$\boldsymbol{G} \;=\; \boldsymbol{L}^{[0]} + \boldsymbol{H}^{[0]}\boldsymbol{L}^{[1]} + \boldsymbol{H}^{[0]}\boldsymbol{H}^{[1]}\boldsymbol{L}^{[1]} + \cdots \;=\; \sum_{m=0}^{\infty} \boldsymbol{\Pi}^{[m]}\boldsymbol{L}^{[m]},$$

and here $\sum_0^k$ is just $\boldsymbol{G}^{[k]}$. The rate estimate then follows from $\mathbb{P}_{i,0}(M \geq 2^k) = \mathbb{P}_{i,0}(\theta(2^k) < \theta(-1)) = \mathrm{O}(\mathrm{e}^{-\gamma 2^k})$; cf. again XIII.8.3. □

We now turn to the fluid case. Here we work with the (terminating) E_+–valued Markov process $\{I_x\}$ where $I_x = J_{\tau(x)}$, the phase when the fluid level for the first time reaches level x. Let $\boldsymbol{U}$ be the intensity matrix. Then (recall the definition of $\boldsymbol{\Sigma}$ and the blocks $\boldsymbol{\Sigma}_{\pm\pm}$ from Remark 4.3):

Proposition 4.9 *$\psi_i(x) \;=\; \boldsymbol{\alpha}^{(i)}\mathrm{e}^{\boldsymbol{U}x}\boldsymbol{1}$, $i \in E$, where $\boldsymbol{\alpha}^{(i)} = \boldsymbol{1}_i'$, $i \in E_+$, $\boldsymbol{\alpha}^{(i)} = \boldsymbol{1}_i'\boldsymbol{P}_{-+}$, $i \in E_-$, where $\boldsymbol{P}_{-+}$ is the $E_- \times E_+$ matrix*

$$\boldsymbol{P}_{-+} \;=\; -\int_0^\infty \mathrm{e}^{-\boldsymbol{\Sigma}_{--}t}\boldsymbol{\Sigma}_{-+}\mathrm{e}^{\boldsymbol{U}t}\,\mathrm{d}t.$$

Proof. Define $\boldsymbol{\Theta}(t)$ as the $E_- \times E_-$ matrix whose ijth element is the probability that, starting from $J_0 = i \in E_-$, the fluid level will reach $-t$ in phase $j \in E_-$ without $\{J_t\}$ entering states in E_+. From $-t$, we reach $-t-h$ in $-h/r(k)$ time units when the current state of $\{J_t\}$ is k (note that only $k \in E_-$ occurs so that $r(k) < 0$!) so that there is a probability $-h\lambda_{kj}/r(k) + \mathrm{o}(h)$ of making a transition to j (an E_-– or E_+–state) before $-t-h$ is reached. This shows easily that $\boldsymbol{\Theta}'(t) = -\boldsymbol{\Sigma}_{--}$ which together with $\boldsymbol{\Theta}(0) = \boldsymbol{I}$ gives $\boldsymbol{\Theta}(t) = \mathrm{e}^{-\Sigma_{--}t}$.

The assertion of the proposition is clear for $i \in E_+$ where it just reflects the fact that the lifetime of $\{I_x\}$ is phase–type with phase generator $\boldsymbol{U}$ and initial vector $\boldsymbol{1}_i'$. For $i \in E_-$, the assertion also becomes clear once we have verified that the ijth element of $\boldsymbol{P}_{-+}$ coincides with $\mathbb{P}_i(J_\omega = j)$ where $\omega = \inf\{t > 0 : S_t = 0\}$. This follows since, as noted above, the ikth element of $\mathrm{e}^{-\boldsymbol{\Sigma}_{--}t}$ gives the probability that starting from $J_0 = i$, the fluid level will reach $-t$ in phase k without $\{J_t\}$ entering states in E_+. W.p. $-h\boldsymbol{1}_k'\boldsymbol{\Sigma}_{-+}\boldsymbol{1}_\ell + \mathrm{o}(h)$, there will then be a transition to $\ell \in E_+$ before the fluid level reaches $-t-h$, and then a process $\{\widetilde{I}_x\}$ distributed as $\{I_x\}$ starts and has $\widetilde{I}_t = J_\omega$. Summing over k, ℓ and integrating t out, the result follows. □

Theorem 4.10 $\boldsymbol{U} = \Phi(\boldsymbol{U})$ *where*

$$\Phi(\boldsymbol{U}) \;=\; \boldsymbol{\Sigma}_{++} - \boldsymbol{\Sigma}_{+-}\int_0^\infty \mathrm{e}^{-\boldsymbol{\Sigma}_{--}t}\boldsymbol{\Sigma}_{-+}\mathrm{e}^{\boldsymbol{U}t}\,\mathrm{d}t.$$

Further, $\boldsymbol{U}$ *is the minimal solution of this equation satisfying* $\boldsymbol{U} \ge \boldsymbol{\Sigma}_{++}$ *and can be computed as* $\lim_{n\to\infty}\Phi\bigl(\boldsymbol{U}^{(n)}\bigr)$ *where* $\boldsymbol{U}^{(0)} = \boldsymbol{\Sigma}_{++}$*,* $\boldsymbol{U}^{(n+1)} = \Phi\bigl(\boldsymbol{U}^{(n)}\bigr)$*.*

Proof. As in the proof of Proposition 4.9, $\boldsymbol{\Sigma}_{++}$ describes the phase changes of $\{I_x\}$ while the fluid level is increasing without phase changes of $\{J_t\}$ to states in E_-. Such changes are described by the matrix $\boldsymbol{\Sigma}_{+-}$, and if there is such a transition at fluid level x, say from $i \in E_+$ to $j \in E_-$, we will have $I_x = k$ with a probability that is the ijth element of $\boldsymbol{P}_{-+}$. Adding these two contributions and appealing to the explicit form of $\boldsymbol{P}_{-+}$ found in Proposition 4.9 shows $\boldsymbol{U} = \Phi(\boldsymbol{U})$. The rest of the proof is rather similar to the proof of Theorem 3.12 and omitted. □

4c The Spectral Method

The martingale calculations in Section 4a provided one way to find $\boldsymbol{G}_+$ in diagonal form. The spectral method provides another one, using the Wiener–Hopf factorization from Section 2f. We will see (not surprisingly) that exactly the same eigenvalues and eigenvectors are required as when using martingales. As an example, we give an alternative:

Proof of Theorem 4.4. In terms of generating functions, $\boldsymbol{I}-\boldsymbol{F} = (\boldsymbol{I}-{}^{\#}\boldsymbol{G}_-)*(\boldsymbol{I}-\boldsymbol{G}_+)$ means in the upward skip–free lattice case that

$$\boldsymbol{I}-\widehat{\boldsymbol{G}}[z] \;=\; \bigl(\boldsymbol{I}-{}^{\#}\widehat{\boldsymbol{G}}_-[z]\bigr)\bigl(\boldsymbol{I}-z\boldsymbol{G}_+\bigr). \tag{4.12}$$

Now $\boldsymbol{G}_-^*$ is stochastic since $\kappa'(0) < 0$. Thus if $|\Re(z)| > 1$, $\widehat{\boldsymbol{G}}_-^*[z]$ is bounded in absolute value by a substochastic matrix which implies that the absolute value of any eigenvalue is strictly smaller than 1, and hence the same is true for ${}^{\#}\widehat{\boldsymbol{G}}_-[z]$. For any ν, it follows by multiplying (4.12) by $\boldsymbol{b}^{(\nu)}$ to the right that $(\boldsymbol{I}-{}^{\#}\widehat{\boldsymbol{G}}_-[z])(\boldsymbol{I}-z\boldsymbol{G}_+)\boldsymbol{b}^{(\nu)} = \boldsymbol{0}$. This implies $(\boldsymbol{I}-z\boldsymbol{G}_+)\boldsymbol{b}^{(\nu)} = \boldsymbol{0}$, since otherwise ${}^{\#}\widehat{\boldsymbol{G}}_-[z]$ would have the eigenvalue 1. Thus $1/z$ is an eigenvalue of $\boldsymbol{G}_+$ with eigenvector $\boldsymbol{b}^{(\nu)}$, and the truth of this for all $\nu = 1,\ldots,q$ immediately implies that $\boldsymbol{G}_+$ has the stated form. □

In the fluid case, similar arguments yield:

Theorem 4.11 *Consider a Markovian fluid with* $\kappa'(0) < 0$ *and let* s_ν*,* $\boldsymbol{b}^{(\nu)}$*,* $\nu = 1,\ldots,q_+$*, etc. be as in Theorem 4.2. Then*

$$\boldsymbol{U} \;=\; \Bigl(s_1\boldsymbol{d}^{(1)}\cdots\; s_{q_+}\boldsymbol{d}^{(q_+)}\Bigr)\Bigl(\boldsymbol{d}^{(1)}\cdots\;\boldsymbol{d}^{(q_+)}\Bigr)^{-1}.$$

Notes The logarithmic reduction algorithm is due to Latouche and Ramaswami (1993). The spectral method for $GI/M/1$ and $M/G/1$ type models is discussed in Gail *et al.* (1996).

5 The Ross Conjecture and Other Ordering Results

A common folklore states that "adding variation decreases performance". For example, if a Markov–modulated $M/G/1$ queue has i.i.d. service times $U_1, U_2, \ldots$, then this leads to expecting that the mean waiting time is larger than the one in a standard $M/G/1$ queue where the service times are the same but the constant arrival rate is $\beta^* = \sum \beta_i \pi_i$, the long–run arrival rate in the Markov–modulated queue.

We will prove below (Corollary 5.4) a version of this result, which covers both a more general model and contains a more general concept of performance degradation than just an increased mean, namely a majorization in the increasing convex ordering $\leq_{\mathrm{icx}}$; see A4 for basic definitions and discussion of why comparisons in increasing convex ordering are also comparisons of performance. However, we start with a comparison of two queueing systems in the discrete time setting of Loynes's lemma; cf. IX.2. That is, we have given two input sequences $\bigl\{X_n^{(1)}\bigr\}_{n\in\mathbb{Z}}$, $\bigl\{X_n^{(2)}\bigr\}_{n\in\mathbb{Z}}$, and define $W_{n+1}^{(i)} = \bigl(W_n^{(i)} + X_n^{(i)}\bigr)^+$, $i = 1, 2$. Assuming that finite limits $W_\infty^{(1)}, W_\infty^{(2)}$ in distributions exist, we look for criteria for $W_\infty^{(1)} \leq_{\mathrm{icx}} W_\infty^{(2)}$. Let $S_n^{(i)} = X_1^{(i)} + \cdots + X_n^{(i)}$ and let $\leq_{\mathrm{cx}}$ denote the convex ordering; see again A4.

Proposition 5.1 *Assume that $S_n^{(i)} \overset{\mathrm{a.s.}}{\to} -\infty$, $i = 1, 2$ and that $\boldsymbol{X}_n^{(1)} \leq_{\mathrm{icx}} \boldsymbol{X}_n^{(2)}$ for all n where $\boldsymbol{X}_n^{(i)} = \bigl(X_1^{(i)}, \ldots, X_n^{(i)}\bigr)$. Then $W_\infty^{(1)} \leq_{\mathrm{icx}} W_\infty^{(2)}$.*

Proof. Define $w_1(x_1) = x_1^+$, $w_2(x_1, x_2) = w_1(w_1(x_1) + x_2)$, $w_3(x_1, x_2, x_3) = w_1(w_2(x_1, x_2) + x_3)$ and so on. Since w_1 is convex and sums of convex functions are convex, it follows easily by induction that $w_n(x_1, \ldots, x_n)$ is convex and increasing on $\mathbb{R}^n$.

By Strassen's theorem (cf. A4), we may assume w.l.o.g. that $\boldsymbol{X}_n^{(1)} \leq \mathbb{E}\bigl[\boldsymbol{X}_n^{(2)}\,\big|\boldsymbol{X}_n^{(1)}\bigr]$. Therefore by monotonicity, Jensen's inequality and the convexity of w_n,

$$\begin{aligned} W_n^{(1)} &= w_n\bigl(\boldsymbol{X}_n^{(1)}\bigr) \leq w_n\bigl(\mathbb{E}\bigl[\boldsymbol{X}_n^{(2)}\,\big|\boldsymbol{X}_n^{(1)}\bigr]\bigr) \\ &\leq \mathbb{E}\bigl[w_n(\boldsymbol{X}_n^{(2)})\,\big|\,\boldsymbol{X}_n^{(1)}\bigr] = \mathbb{E}\bigl[W_n^{(2)}\,\big|\,\boldsymbol{X}_n^{(1)}\bigr]. \end{aligned}$$

Therefore, if f is increasing and convex,

$$\mathbb{E}f\bigl(W_n^{(1)}\bigr) \leq \mathbb{E}f\bigl(\mathbb{E}\bigl[W_n^{(2)}\,\big|\,\boldsymbol{X}_n^{(1)}\bigr]\bigr) \leq \mathbb{E}\mathbb{E}\bigl[f(W_n^{(2)})\,\big|\,\boldsymbol{X}_n^{(1)}\bigr] = \mathbb{E}f\bigl(W_n^{(2)}\bigr),$$

where we used $f\uparrow$ in the first step and Jensen's inequality in the next. Letting $n\uparrow\infty$ and using monotone convergence justified by

$$W_{n+1}^{(i)}\ \overset{\mathscr{D}}{=}\ \max_{k\le n+1}X_k^{(i)}\ \ge\ \max_{k\le n}X_k^{(i)}\ \overset{\mathscr{D}}{=}\ W_n^{(i)}$$

completes the proof. □

Corollary 5.2 *Let $W_\infty^{(1)},W_\infty^{(2)}$ be the steady–state waiting times in two $GI/G/1$ queues with generic interarival times $T^{(1)},T^{(2)}$ and generic service times $U^{(1)},U^{(2)}$. If $U^{(1)}\le_{\rm icx}U^{(2)}$ and $T^{(1)}\le_{\rm cx}T^{(2)}$, then $W_\infty^{(1)}\le_{\rm icx}W_\infty^{(2)}$.*

Proof. Since $-T^{(1)}\le_{\rm cx}-T^{(2)}$ (implying $-T^{(1)}\le_{\rm icx}-T^{(2)}$), and $\le_{\rm icx}$ is closed under convolution, we have

$$X^{(1)}\ =\ U^{(1)}-T^{(1)}\ \le_{\rm icx}\ U^{(2)}-T^{(2)}\ =\ X^{(2)},$$

which is easily seen to imply $\boldsymbol{X}_n^{(1)}\le_{\rm icx}\boldsymbol{X}_n^{(2)}$. □

Corollary 5.3 (i) *Consider the class of stable $GI/G/1$ queues with fixed interarrival distribution A and fixed mean $\mu=\mu_B$ of the service time distribution B. Then the steady–state waiting time W is minimized in increasing convex ordering for $GI/D/1$, i.e. by taking the service time distribution degenerate at μ.*

(ii) *Consider the class of stable $GI/G/1$ queues with fixed service time distribution B and fixed mean $\nu=\mu_A$ of the service time distribution B. Then the steady–state waiting time W is minimized in increasing convex ordering for $D/G/1$, i.e. by taking the interarrival time distribution degenerate at ν.*

Proof. Just note for (i) that any r.v. U (a generic service time) satisfies $U\ge_{\rm cx}\mathbb{E}U$ and appeal to Corollary 5.2. The proof of (ii) is similar. □

We now return to the continuous–time problem discussed in the beginning of this section. Consider the $M/G/1$ workload process $\{V_t^{(2)}\}_{t\ge0}$ in a random environment, where the arrival process is a Cox process (cf. A3): the arrival intensity at time t is $\beta(t)$ for some stationary ergodic process $\{\beta(t)\}$ independent of the sequence $U_1,U_2,\ldots$ of service times (assumed i.i.d.). Define $\beta^*=\mathbb{E}\beta(t)$ and let $\{V_t^{(1)}\}_{t\ge0}$ be the standard $M/G/1$ workload process with arrival intensity β^* and the same service time distribution. Similarly, let the waiting time sequences in the two systems be $\{W_n^{(1)}\}_{n\in\mathbb{N}}$, $\{W_n^{(2)}\}_{n\in\mathbb{N}}$, where for the random environment waiting time sequence it is assumed that the arrival process is the Palm version (with governing probability measure $\mathbb{P}_0$) of $\{\beta_t\}$ (see VII.6).

Corollary 5.4 *Assume $\rho=\beta^*\mathbb{E}U_1<1$. Then:*

(a) *$W_n^{(1)}$, $W_n^{(2)}$ have limits $W_\infty^{(1)}$, resp. $W_\infty^{(2)}$, in distribution, and $W_\infty^{(1)}\le_{\rm icx}W_\infty^{(2)}$;*

(b) $V_t^{(1)}$, $V_t^{(2)}$ *have limits* $V_\infty^{(1)}$, *resp.* $V_\infty^{(2)}$, *in distribution, and* $V_\infty^{(1)} \leq_{\mathrm{icx}} V_\infty^{(2)}$.

Proof. In the setting of Loynes's lemma, $\{V_t^{(1)}\}$, $\{V_t^{(2)}\}$ are reflected versions of

$$S_t^{(1)} = \sum_{i=1}^{N_t^*} U_i - t, \quad S_t^{(2)} = \sum_{i=1}^{N_t} U_i - t,$$

where $\{N_t\}$ is Poisson with stochastic intensity $\beta(t)$ and independent of $U_1, U_2, \ldots$, and $\{N_t^*\}$ is standard Poisson with intensity β^* and independent of $U_1, U_2, \ldots$. By the ergodic theorem, $N_t/t \overset{\text{a.s.}}{\to} \beta^*$, and therefore by the LLN,

$$\frac{S_t^{(2)}}{t} = \frac{N_t}{t}\sum_{i=1}^{N_t} U_i/N_t - 1 \overset{\text{a.s.}}{\to} \beta^* \mathbb{E} U_i - 1 = \rho - 1.$$

Similarly but easier, $S_t^{(1)}/t \overset{\text{a.s.}}{\to} \rho - 1$. Since $\rho < 1$, we therefore have $S_t^{(i)} \overset{\text{a.s.}}{\to} -\infty$, $i = 1, 2$, and the existence of limits is immediate from IX.2c. A similar argument applies to the waiting times.

For the increasing convex ordering, we first consider (a). Denote the interarrival times in the two systems by $T_n^{(1)}, T_n^{(2)}$. We may assume that the $T_n^{(1)}, T_n^{(2)}$ have ben constructed from i.i.d. standard exponentials $T_n^{(0)}$ as $T_n^{(1)} = T_n^{(0)}/\beta^*$ and

$$T_n^{(2)} = \Gamma\Bigl(\sum_{k=1}^{n} T_k^{(0)}\Bigr) - \Gamma\Bigl(\sum_{k=1}^{n-1} T_k^{(0)}\Bigr),$$

where $\Gamma(t) = \inf\{v > 0 : B(v) \geq t\}$ with $B(v) = \int_0^v \beta(s)\,\mathrm{d}s$ (see A3 and VII.6.5 where it was also shown that $\beta^* \mathbb{E}_0 \Gamma(t) = t$ for all t). Letting $X_n^{(i)} = U_k - T_k^{(i)}$ and

$$\mathscr{G}_n = \sigma\bigl(U_1, T_1^{(0)}, \ldots, U_n, T_n^{(0)}\bigr), \quad \mathscr{G}_n = \sigma\bigl(T_1^{(0)}, \ldots, T_n^{(0)}\bigr),$$

it follows that

$$\begin{aligned} \mathbb{E}_0\bigl[X_m^{(2)} \,\big|\, \mathscr{G}_n\bigr] &= U_m - \mathbb{E}_0\bigl[T_m^{(2)} \,\big|\, \mathscr{G}_n\bigr] \\ &= U_m - \Bigl(\sum_{k=1}^{m} T_k^{(0)}/\beta^* - \sum_{k=1}^{m-1} T_k^{(0)}/\beta^*\Bigr) = U_m - T_m^{(0)}/\beta^* = X_m^{(1)}, \end{aligned}$$

$m = 1, \ldots, n$. Thus the convex ordering condition of Proposition 5.1 holds.

For (b), just appeal to the relation X.(3.3) between the stationary workload and waiting time distribution (the proof carries over the present setting without changes); the two systems have the same traffic intensity ρ and U^*, and $W_\infty^{(1)} + U^* \leq_{\mathrm{icx}} W_\infty^{(2)} + U^*$ by independence. □

Notes The classical reference on ordering of queues is Stoyan (1983) (see also Müller and Stoyan, 2002); there is also much material in Baccelli and Brémaud (2002), Chen and Yao (2001) and Szekli (1995). The special case $\mathbb{E}V_\infty^{(1)} \le \mathbb{E}V_\infty^{(2)}$ of Corollary 5.4 goes under the name *the Ross conjecture* and there is an extensive literature on this and related problems. The conjecture in the more general form of Corollary 5.4 was proved by Rolski (1981); for a recent paper in the area and references, see Hordijk (2001).

XII
Many–Server Queues

1 Comparisons with $GI/G/1$

Many–server queues present some of the most intricate problems in queueing theory and provide examples of how models, which are simple and well motivated from practical situations, may lead to substantial mathematical difficulties. Not only are the steady–state characteristics more difficult to evaluate than in the single–server case, but also even just to show existence of (unique) limits presents major difficulties when pursuing the model in its greatest generality.

We consider the standard $GI/G/s$ queue with customers $n = 0, 1, 2, \ldots$, service times $U_1, U_2, \ldots$ (governed by B), interarrival times $T_0, T_1, \ldots$ (governed by A) and FCFS queue discipline, meaning that the customers join service in the order they arrive (for $s = 1$, this implies the FIFO property of a similar ordering of the departures but this is not the case for $s > 1$). The model may be represented in various ways, one of the most obvious being that the customers form one line in the order of arrival and the customer in front joins the first server to become idle. However, most often we think of each server as having his own waiting line and the arriving customer joining the line that is the first to become available, i.e. that has the least residual work. For mathematical purposes, we order the residual work in the various lines at time t and thus obtain a vector $\boldsymbol{V}_t = \bigl(V_t^{(1)} \cdots V_t^{(s)}\bigr)^{\mathsf{T}}$ satisfying $V_t^{(1)} \le V_t^{(2)} \le \cdots \le V_t^{(s)}$. It is of particular interest to observe $\boldsymbol{V}_t$ just before the arrival instants $\tau(n) = T_0 + \cdots + T_{n-1}$ and we write $\boldsymbol{W}_n = \bigl(W_n^{(1)} \cdots W_n^{(s)}\bigr)^{\mathsf{T}} = \boldsymbol{V}_{\tau(n)-}$. Thus the FCFS discipline implies that $W_n^{(1)}$

is the waiting time of the nth customer. As generalization of the Lindley recursion $W_{n+1} = (W_n + U_n - T_n)^+$ we get

$$\boldsymbol{W}_{n+1} = \mathscr{R}\Bigl(\bigl(W_n^{(1)} + U_n - T_n\bigr)^+, \bigl(W_n^{(2)} - T_n\bigr)^+, \ldots, \bigl(W_n^{(s)} - T_n\bigr)^+\Bigr)^{\mathsf{T}} \qquad (1.1)$$

where $\mathscr{R}$ is the operator on $\mathbb{R}^s$ which orders the coordinates in ascending order (one immediate implication is that $\{\boldsymbol{W}_n\}$ is a Markov chain); (1.1) is commonly referred to as the *Kiefer–Wolfowitz recursion* and $\boldsymbol{W}_n$ is denoted the *Kiefer–Wolfowitz vector.* Finally, we let Q_t denote the queue length (number of customers in the system) at time t and write

$$|\boldsymbol{v}| = \bigl|(v^{(1)} \cdots v^{(s)})^{\mathsf{T}}\bigr| = v^{(1)} + \cdots + v^{(s)} \quad \text{when all } v^{(s)} \geq 0.$$

Thus, for example, $|\boldsymbol{V}_t|$ is the workload (remaining work) at time t.

By good luck, many problems in the theory of the $GI/G/s$ queue that are difficult to approach directly may be reduced to the case $s = 1$ by obtaining suitable bounds in terms of single–server systems with the same traffic intensity. For the present applications, it suffices to consider an initial empty queue, and this will be done for the sake of simplicity. Starting with the lower (and easier) bound, let $\{W_n^*\}$, $\{V_t^*\}$, etc. refer to a $GI/G/1$ queue with the same interarrival times $T_0, T_1, \ldots$ and service times $U_0/s, U_1/s, \ldots$. The server in this system reduces workload at the same rate as when all s servers are busy in the $GI/G/s$ system, i.e. as when this system is working at its highest capacity, and in fact we have:

Theorem 1.1 *If* $sV_{0-}^* \leq |\boldsymbol{V}_{0-}|$, *then* $sV_t^* \leq |\boldsymbol{V}_t|$ *for all* t.

Proof. We first note that $sV_t^* \leq |\boldsymbol{V}_t|$ for $0 \leq t < T_0$. This follows simply by using the inequality $(x+y)^+ \leq x^+ + y^+$ s times to obtain

$$\begin{aligned} sV_t^* &= s\bigl(V_{0-}^* + U_0/s - t\bigr)^+ \leq \Bigl(V_{0-}^{(1)} + U_0 - t + \sum_{i=2}^{s}\bigl(V_{0-}^{(i)} - t\bigr)\Bigr)^+ \\ &\leq \bigl(V_{0-}^{(1)} + U_0 - t\bigr)^+ + \sum_{i=2}^{s}\bigl(V_{0-}^{(i)} - t\bigr)^+ = |\boldsymbol{V}_t| \end{aligned}$$

for $t < T_0 = \tau(1)$. In particular, $sV_{\tau(1)-}^* \leq |\boldsymbol{V}_{\tau(1)-}|$ so that repeating the argument yields $sV_t^* \leq |\boldsymbol{V}_t|$, $\tau(1) \leq t < \tau(2)$, and the desired conclusion follows by iteration. □

Note that we cannot infer similar bounds for the waiting times themselves: If say $U_0 > sT_0$, then $W_1^{(1)} = 0$ but $W_1^* = (U_0/s - T_0)^+ > 0$.

As upper bound, we shall consider an s–server queue with the same interarrival times and service times, but with a different allocation of the customers to the servers (by an allocation we mean a $\{1, \ldots, s\}$–valued adapted function $\sigma(n)$ telling which server customer n joins). In addition to the $GI/G/s$ FCFS allocation rule, corresponding to a customer joining the server with the lowest workload, a main example is the *cyclic discipline*

$\sigma(ks+i) = i$, where every sth customer goes to server i. Then the queue in front of server i is simply a $GI/G/1$ queue with interarrival time distribution A^{*s} and service time distribution B (these s queues are highly dependent due to the dependence between their interarrival times). In the following, we will denote the system with a possibly non–FCFS allocation rule by tildes.

Intuitively, one feels that the given $GI/G/s$ rule should be optimal, and indeed, we will show:

Theorem 1.2 *For any (possibly non–FCFS) allocation rule, it holds for initially empty systems, that $Q_t \le_{\rm so} \widetilde{Q}_t$ and $|\boldsymbol{V}_t| \le_{\rm so} |\widetilde{\boldsymbol{V}}_t|$ for all $t \ge 0$. Similarly $|\boldsymbol{W}_n| \le_{\rm so} |\widetilde{\boldsymbol{W}}_n|$ for all n.*

It is tempting to assert that the result holds also in the sense of sample paths. This is, however, false (Problem 1.1) but can be achieved using a suitable coupling, cf. Lemma 1.3 below. Let

$$0 \le J_0 \le J_1 \le \ldots, \quad 0 \le \widetilde{J}_0 \le \widetilde{J}_1 \le \ldots$$

be the ordered epochs of initiation of service in the two systems (the FCFS rule then simply means that J_n is the instant where customer n initiates service), but with the modification that in the modified system we allocate the service times $U_0, U_1, \ldots$ to the customers not according to their order of arrival, but rather according to the order in which they join service. Thus the service time of a particular customer is chosen from the sequence $\{U_n\}$ in a way that depends on $\{T_n\}$ and the service times of other customers. However, by independence, distributional properties remain the same (e.g. in the cyclical case, any server still faces a $GI/G/1$ system). Similar remarks apply to modifications of the procedure as in the proof of Theorem 1.2 below.

Letting $\min^{(k)}$ denote the operation that returns the kth order statistics from a finite or infinite set of r.v.'s, it follows that the ordered departure times from the two systems are

$$D_k = \min{}^{(k)}\{J_0+U_0, J_1+U_1, \ldots\}, \quad \widetilde{D}_k = \min{}^{(k)}\{\widetilde{J}_0+U_0, \widetilde{J}_1+U_1, \ldots\}.$$

Lemma 1.3 *For initially empty systems coupled as above, it holds that $D_k \le \widetilde{D}_k$ for all k. In particular, $Q_t \le \widetilde{Q}_t$ for all $t \ge 0$.*

Proof. The crux is to establish the relations

$$\begin{aligned} \widetilde{J}_n &\ge \max\{\tau(n), \widetilde{D}_{n-s}\}, \quad J_n = \max\{\tau(n), D_{n-s}\} && (1.2)\\ \widetilde{D}_k &= \min{}^{(k)}\{\widetilde{J}_n + U_n : 0 \le n < k+s\} && (1.3)\\ D_k &= \min{}^{(k)}\{J_n + U_n : 0 \le n < k+s\} && (1.4) \end{aligned}$$

where in (1.2) we let $D_k = \widetilde{D}_k = 0$ for $k < 0$. In the first half of (1.2) we have obviously $\widetilde{J}_n \ge \tau(n)$ and thus the assertion is true for $n = 0, \ldots, s-1$. Also, for $n \ge s$ it would follow from $\widetilde{J}_n < \widetilde{D}_{n-s}$ that at time $\widetilde{J}_n$ at least

$n+1-(n-s) = s+1$ customers were receiving service, which is impossible. This shows the first half of (1.2). For $n \geq k+s$ we then get $\widetilde{D}_k \leq \widetilde{J}_{k+s} \leq \widetilde{J}_n \leq \widetilde{J}_n + U_n$ and (1.3) follows by combining with the definition of $\widetilde{D}_k$ (of course, (1.4) is just a special case of (1.3)).

For the second half of (1.2), note that if $W_n^{(1)} = 0$ (i.e. customer n does not wait), then $J_n = \tau(n)$, whereas otherwise $J_n = D_{n-s}$. The claim thus follows by noting that $W_n^{(1)} = 0$ if and only if $D_{n-s} \leq \tau(n)$.

It now follows from (1.2)–(1.4) that $\widetilde{J}_n \geq J_n$ for all n. Indeed, this is obvious for $n \leq s-1$, and if $\widetilde{J}_n \geq J_n$ for all $n \leq N$, then

$$\begin{aligned}
D_{N+1-s} &= \min{}^{(N+1-s)}\{J_n + U_n : n \leq N\} \\
&\leq \min{}^{(N+1-s)}\bigl\{\widetilde{J}_n + U_n : n \leq N\bigr\} \;=\; \widetilde{D}_{N+1-s}, \\
\widetilde{J}_{N+1} &\geq \max\bigl\{\tau(N+1), \widetilde{D}_{N+1-s}\bigr\} \;\geq\; \max\{\tau(N+1), D_{N+1-s}\} \\
&= J_{N+1}.
\end{aligned}$$

It then follows from (1.3)–(1.4) that $\widetilde{D}_n \geq D_n$ for all n, and since the arrival epochs in the modified system and the $GI/G/s$ queue are the same, it is immediate that $\widetilde{Q}_t \geq Q_t$ for all $t \geq 0$. □

Proof of Theorem 1.2. That $Q_t \leq_{\text{so}} \widetilde{Q}_t$ is immediate from Lemma 1.2 since the coupling does not change the distribution. For $\boldsymbol{V}_t$ and $\boldsymbol{W}_n$, we proceed by a modification of the coupling. Fixing t and letting M denote the number of arrivals before t, the idea is to treat the $m = \inf\{n : J_n > t\}$, resp. $\widetilde{m} = \inf\bigl\{n : \widetilde{J}_n > t\bigr\}$, of the customers who join service before t in the same way as before, but to allocate service times to the remaining $M-m$, respectively $\widetilde{M} - m$, in the natural order. Thus the service times of customers arriving before time t are permutations (different for the $GI/G/s$ system and the alternative rule for allocation to servers) of $U_0, \ldots, U_{M-1}$. The modified systems obtained this way are denoted by superscripts$^{\#}$, and obviously $\boldsymbol{V}_t^{\#} \stackrel{\mathscr{D}}{=} \boldsymbol{V}_t$, $\widetilde{\boldsymbol{V}}_t^{\#} \stackrel{\mathscr{D}}{=} \widetilde{\boldsymbol{V}}_t$ so that it suffices to show $|\boldsymbol{V}_t^{\#}| \leq \bigl|\widetilde{\boldsymbol{V}}_t^{\#}\bigr|$. But because of $J_n \leq \widetilde{J}_n$, we have $m \geq \widetilde{m}$, and hence (using $J_n \leq t$)

$$\begin{aligned}
\bigl|\boldsymbol{V}_t^{\#}\bigr| &= \sum_{n=0}^{m-1}(J_n + U_n - t)^+ + \sum_{n=m}^{M-1} U_n \\
&\leq \sum_{n=0}^{\widetilde{m}-1}\bigl(\widetilde{J}_n + U_n - t\bigr)^+ + \sum_{n=\widetilde{m}}^{m-1} U_n + \sum_{n=m}^{M-1} U_n \;=\; \bigl|\widetilde{\boldsymbol{V}}_t^{\#}\bigr|.
\end{aligned}$$

The similar inequality for the $\boldsymbol{W}_n$ follows by just the same argument. □

Here is a marked difference between single– and many–server queues. Define $\rho = \mathbb{E}U/s\mathbb{E}T$. Then (recall from X.2 that $\mathbb{E}W < \infty$ if and only if $\mathbb{E}U^2 < \infty$ when $s=1$, $\rho < 1$):

Theorem 1.4 $\mathbb{E}W^{(1)} < \infty$ *provided* $\rho < 1 - 1/s$ *and* $\mathbb{E}U^{3/2} < \infty$.

Proof. We give the proof for $s = 2$ only. The crux is the recursion $W_{n+1}^{(1)} = \big(W_n^{(1)} + V_n - T_n\big)^+$ where $V_n = U_n \wedge \big(W_n^{(2)} - W_n^{(1)}\big)$, as follows easily from (1.1). Of course, this is not a simple Lindley recursion since the V_n are highly dependent on the past, but after having shown $\mathbb{E}_e V_0^2 < \infty$ it indeed allows to infer $\mathbb{E}W^{(1)} < \infty$ in a rather similar way as in X.2.

Let $\mathbb{P}_e$ be the governing probability measure for a doubly infinite stationary version of $\big\{\big(U_n, T_n, W_n^{(1)}\big)\big\}$. Then $\mathbb{E}_e|\boldsymbol{W}_0|^{1/2} < \infty$ by Theorem 1.2 and X.2.1, and hence, since $\boldsymbol{W}_0$ and U_0 are independent,

$$\mathbb{E}_e V_0^2 \le \mathbb{E}_e\big[U_0 \wedge |\boldsymbol{W}_0|\big]^2 \le \mathbb{E}_e\big[U_0^{3/2}|\boldsymbol{W}_0|^{1/2}\big] = \mathbb{E}_e U_0^{3/2} \cdot \mathbb{E}_e|\boldsymbol{W}_0|^{1/2} < \infty.$$

Similarly, bounding V_0 by U_0 gives $\mathbb{E}_e[V_0 - T_0 \,|\, \mathscr{F}_0] \le -\delta$ where $\mathscr{F}_0 = \sigma(U_n, T_n : n < 0)$ and $-\delta = \mathbb{E}(U_n - T_n) < 0$. Now just appeal to a small modification of the proof of X.2.1 given in Problem 1.3. □

Problems

1.1 Let $s = 2$, $T_0 = 1/2$, $T_1 = 3/4$, $U_0 = 3/2$, $U_1 = 100$, $U_2 = 1/2$, and suppose that in the modified system customers 0,1 are allocated to server 1, customer 2 to server 2. Show that $|\boldsymbol{V}_1| = 100$, $|\widetilde{\boldsymbol{V}}_1| = 1$.
1.2 Show in the setting of Theorem 1.2 that $\mathbb{E}W_n^{(1)} \le \mathbb{E}\widetilde{W}_n^{(1)}$.
1.3 Let $\{X_n\}_{n\in\mathbb{Z}}$ be a stationary ergodic sequence such that $\mathbb{E}[X_0 \,|\, \mathscr{F}_0] \le -\delta$ where $\mathscr{F}_0 = \sigma(X_n : n < 0)$ and $\delta > 0$. Let as in IX.2 $\{(X_n, W_n)\}_{n\in\mathbb{Z}}$ be stationary with $W_{n+1} = (W_n + X_n)^+$. Show that $\mathbb{E}X_0^2 < \infty$ implies $\mathbb{E}W_0 < \infty$.

Notes Theorem 1.2 is from Wolff (1987); see also Foss (1980). A recent paper in the area is Foss and Chernova (2001), and further references can be found there.

Theorem 1.4 is from Scheller–Wolf and Sigman (1997). The intuition behind that the conditions for (say) $s = 2$ are weaker than for $s = 1$ is that if $s = 2$, $\rho < 1/2$, then the system would be stable if only one of the servers was operating so that the second server represents an extra capacity of the system. Sowewhat similar dependencies on ρ show up in heavy–tailed asymptotics; see Whitt (2000). Further conditions (of quite intricate form!) for finiteness of moments of $W^{(1)}$ are in Scheller–Wolf (2003).

2 Regeneration and Existence of Limits

Motivated from the single–server case, our aim is to show that waiting–time vectors, queue length processes and so on have limits for $\rho = \mathbb{E}U/s\mathbb{E}T < 1$ but not for $\rho \ge 1$.

The case $\rho \ge 1$ is by far the easiest. In fact, letting W_n^*, V_t^*, etc. refer to the $GI/G/1$ system in Theorem 1.1 (which has the same traffic intensity as the given $GI/G/s$ system), we have $W_n^* \xrightarrow{\mathscr{D}} \infty$, $V_t^* \xrightarrow{\mathscr{D}} \infty$ and thus it is clear from Theorem 1.1 that $|\boldsymbol{W}_n| \xrightarrow{\mathscr{D}} \infty$, $|\boldsymbol{V}_n| \xrightarrow{\mathscr{D}} \infty$. We give a slightly stronger result below (Corollary 2.6) and pass right on to the more interesting and difficult case $\rho < 1$.

The straightforward generalization of the $GI/G/1$ methodology would be to base the analysis on the sequence $\{\sigma(k)\}$ of customers arriving at an empty system. These are obviously regeneration points for say $\{\boldsymbol{W}_n\}$, but (perhaps somewhat unexpectedly), it turns out that $\rho < 1$ alone is not enough to ensure that the renewal process $\{\sigma(k)\}$ will be nonterminating. Define

$$\begin{aligned}\alpha_+ &= \operatorname{ess\,sup} A \;=\; \sup\bigl\{x : \overline{A}(x) > 0\bigr\},\\ \beta_- &= \operatorname{ess\,inf} B \;=\; \sup\{x : B(x) = 0\}\,.\end{aligned}$$

Example 2.1 Suppose that $\beta_- > \alpha_+$ (e.g. for $s = 2$, A may be uniform on $(1/2, 1)$ and B degenerate at $5/4$ so that $\rho = 5/6 < 1$). Then $U_n > T_n$ which means that customer n is still present in the system when $n + 1$ arrives. That is, *the system never becomes empty.* □

It turns out, in fact, that $\beta_- < \alpha_+$ (or equivalently $\mathbb{P}(U < T) > 0$) ensures that $\{\sigma(k)\}$ will be nonterminating. This assumption does not appear terribly restrictive from the point of view of applications (where typically either A or B has support on the whole of $(0, \infty)$ so that $\beta_- < \alpha_+$ is automatic). Nevertheless, we shall persue the general case, which presents a classical problem and for which many ingenious and interesting ideas have been developed.

It was noted in Section 1 that $\{\boldsymbol{W}_n\}$ is a Markov chain on $E = \bigl\{\boldsymbol{w} : 0 \le w^{(1)} \le \cdots \le w^{(s)}\bigr\}$, and we shall show:

Theorem 2.2 *If $\rho < 1$, then $\{\boldsymbol{W}_n\}$ is Harris ergodic on E. Therefore an E–valued random variable $\boldsymbol{W}$ exists such that $\boldsymbol{W}_n \to \boldsymbol{W}$ in total variation. In particular, the waiting times converge, $W_n^{(1)} \to W^{(1)}$ in t.v.*

The proof rests on two lemmas:

Lemma 2.3 *If $\rho < 1$, then $\{\boldsymbol{W}_n\}$* (or equivalently $\{|\boldsymbol{W}_n|\}$) *is tight.*

Proof. This follows simply by comparison with single–server queues: If the modified system of Section 1 corresponds to the cyclical allocation rule, then $\bigl\{|\widetilde{\boldsymbol{W}}_n|\bigr\}$ is the sum of s (dependent) waiting time sequences in $GI/G/1$ queues with $\rho < 1$. Hence $\bigl\{|\widetilde{\boldsymbol{W}}_n|\bigr\}$ is tight, and therefore $\{|\boldsymbol{W}_n|\}$ is so according to Theorem 1.2. □

Lemma 2.4 *If $\rho < 1$, then the set $R_K = \bigl\{\boldsymbol{w} \in E : w^{(s)} \le K\bigr\}$ is a regeneration set for $\{\boldsymbol{W}_n\}$ in the sense of VII.3 for all sufficiently large K.*

Proof. Recalling that $\alpha_+ = \operatorname{ess\,sup} A$, $\beta_- = \operatorname{ess\,inf} B$, it follows from $\rho < 1$ that $\beta_- < s\alpha_+$. Hence we can find $\eta, \epsilon > 0$ such that the event $F_k = \{U_k < s\eta - \epsilon, T_k > \eta\}$ has positive probability, say δ. Let r be an integer so large that $r > sK/\epsilon$ and define $F = F_0 \cdots F_{r-1}$. Loosely speaking, a main idea of the proof is that each occurrence of a F_k decreases residual work (at least when $W_k^{(i)} > \eta$ for all i) and that r has been chosen so large that the dependence on the particular value of $\boldsymbol{W}_0 = \boldsymbol{w} \in R_K$ becomes unimportant after r steps. To make this more precise, consider again the (coupled) cyclical system of Theorem 1.2 and let $\boldsymbol{W}_0 = \boldsymbol{w} \in R_K$. It is then easy to check (say from the expression III.6.3 for the $GI/G/1$ waiting time) that customer n does not have to wait provided that $r - 2s \le n \le r$ and that F occurs. Hence the queue length at the $(r-s)$th arrival is at most $s-1$ in the (coupled) cyclical system, therefore also in the $GI/G/s$ system. This means that customers $r-s, \ldots, r-1$ enter service immediately. Thus with $\lambda(A) = \mathbb{P}_{\boldsymbol{0}}(\boldsymbol{W}_s \in A \,|\, F_0 \ldots F_{s-1})$, we have

$$\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{W}_r \in A) \ \ge\ \delta^r \lambda(A), \quad \boldsymbol{w} \in R_K, \tag{2.1}$$

and it only remains to show that R_K is recurrent for some (and then necessarily all larger) K. But let $L = s\eta$, $G_K = \{W_n \in R_K \text{ i.o.}\}$, $G = \left\{\underline{\lim}\, W_n^{(s)} < \infty\right\}$. Then λ in (2.1) is concentrated on R_L and hence (say by the conditional Borel–Cantelli lemma) $\mathbb{P}G_L \ge \mathbb{P}G_K$ for all K. Thus also $\mathbb{P}G_L \ge \mathbb{P}G$ since $G_K \uparrow G$. But $\mathbb{P}G = 1$ by tightness, hence $\mathbb{P}G_L = 1$. □

Proof of Theorem 2.2. It follows from Lemma 2.4 that $\{\boldsymbol{W}_n\}$ is Harris recurrent. Also, aperiodicity follows since (2.1) holds for all sufficiently large r (cf. Problem VII.3.2) and we only have to show $|\pi| < \infty$, where π is the stationary measure. Now a perusal of the construction of π in VII.3 easily shows that R_K being a regeneration set implies that $\pi(R_K)$, being the expectation of the geometrically bounded number of visits to K during a cycle, is finite. Hence if $|\pi| = \infty$, VII.3.9 would yield $\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{W}_n \in R_K) \to 0$ for all K, i.e. $W_n^{(s)} \xrightarrow{\mathscr{D}} \infty$. But this contradicts Lemma 2.3. □

In continuation of Example 2.1, we also get:

Corollary 2.5 *If $\rho < 1$ and $\beta_- < \alpha_+$, then the sequence $\{\sigma(k)\}$ of customers entering an empty system* (i.e. satisfying $\boldsymbol{W}_{\sigma(k)} = \boldsymbol{0}$) *is an aperiodic nonterminating renewal process with finite mean interarrival time.*

Proof. The condition $\beta_- < \alpha_+$ ensures $\mathbb{P}F_k' = \delta' > 0$, where $F_k' = \{U_k < \eta' - \epsilon < \eta' < T_k\}$. Just as in the proof of Lemma 2.4 (even easier!) it then follows that $\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{W}_r = \boldsymbol{0}) \ge \delta'^r$, $\boldsymbol{w} \in R_K$, for all sufficiently large r. Since R_K is recurrent, a geometrical trial argument then shows that $\boldsymbol{0}$ is so too, and the rest of the argument is much the same as before. □

Returning to the case $\rho \ge 1$ for a brief remark, we shall show:

Corollary 2.6 *If $\rho \ge 1$, then the waiting time process $\left\{W_n^{(1)}\right\}$ satisfies $W_n^{(1)} \xrightarrow{\mathscr{D}} \infty$.*

This follows simply by combining the estimate $\left|W_n^{(1)}\right| \xrightarrow{\mathscr{D}} \infty$ observed earlier with the following bound on the dispersion of the servers:

Lemma 2.7 *Define* $Z_n^{(i)} = W_n^{(s)} - W_n^{(i)}$. *Then the sequence* $\{\boldsymbol{Z}_n\}$ *is tight.*

Proof. Since all $Z_n^{(i)} \geq 0$, it suffices to show that $\{|\boldsymbol{Z}_n|\}$ is tight. Now

$$|\boldsymbol{Z}_{n+1}| \;=\; W_{n+1}^{(s)} \;-\; (W_n^{(1)} + U_n - T_n)^+ \;+\; \sum_{i=2}^{s} \bigl[W_{n+1}^{(s)} - (W_n^{(i)} - T_n)^+\bigr].$$

Define $H = (W_n^{(1)} + U_n - T_n)^+$, $K = (W_n^{(s)} - T_n)^+$. If $H \geq K$, we have $W_{n+1}^{(s)} = H$ and get

$$\begin{aligned} |\boldsymbol{Z}_{n+1}| \;&=\; \sum_{i=2}^{s} \bigl[H - (W_n^{(i)} - T_n)^+\bigr] \;\leq\; \sum_{i=2}^{s} \bigl[W_n^{(1)} + U_n - W_n^{(i)}\bigr] \\ &\leq\; (s-1)U_n, \end{aligned}$$

using the inequality $h^+ - j^+ \leq h - j$ valid for $h \geq j$. If $H < K$, we have $W_{n+1}^{(s)} = K$ and get similarly

$$\begin{aligned} |\boldsymbol{Z}_{n+1}| \;&=\; K - H \;+\; \sum_{i=2}^{s} \bigl[K - (W_n^{(i)} - T_n)^+\bigr] \\ &\leq\; W_n^{(s)} \;-\; \bigl(W_n^{(1)} + U_n\bigr) \;+\; \sum_{i=2}^{s} \bigl[W_{n+1}^{(s)} - W_n^{(i)}\bigr] \;=\; |\boldsymbol{Z}_n| - U_n. \end{aligned}$$

Thus

$$\begin{aligned} |\boldsymbol{Z}_{n+1}| \;&\leq\; \max\Bigl\{|\boldsymbol{Z}_n| - U_n, (s-1)U_n\Bigr\} \;\leq\; \cdots \\ &\leq\; \max\Bigl\{|\boldsymbol{Z}_0| - \sum_{i=0}^{n} U_i,\; (s-1)U_k - \sum_{i=k+1}^{n} U_i \,:\, 0 \leq k \leq n\Bigr\} \\ &\overset{\mathscr{D}}{=}\; \max\Bigl\{|\boldsymbol{Z}_0| - \sum_{i=0}^{n} U_i,\; (s-1)U_k - \sum_{i=0}^{k-1} U_i \,:\, 0 \leq k \leq n\Bigr\} \\ &\overset{\mathscr{D}}{\to}\; \max_{0 \leq k < \infty} \Bigl\{(s-1)U_k - \sum_{i=0}^{k-1} U_i\Bigr\}. \end{aligned}$$

This limit is finite a.s. since $\mathbb{E}U < \infty$ implies $U_k/k \overset{\text{a.s.}}{\to} 0$ and $\sum_0^{k-1} U_i \sim k\mathbb{E}U$. This shows that $\{|\boldsymbol{Z}_n|\}$ is tight. □

We proceed to continuous time.

Corollary 2.8 *Suppose* $\rho < 1$. *If* (a) A *is nonlattice and* $\mathbb{P}(U < T) > 0$, *then* $Q = \lim_{t\to\infty} Q_t$ *exist as a t.v. limit and* $\boldsymbol{V} = \lim_{t\to\infty} \boldsymbol{V}_t$ *as a weak limit. If instead* (b) A *is spread out, then both limits exist in t.v.*

Proof. In case (a), the system regenerates in the usual sense at the instants $C(k) = T_0 + \cdots + T_{\sigma(k)-1}$ of arrivals at an empty system. The cycle length distribution is the $\mathbb{P}_{\boldsymbol{0}}$–distribution of $C(1)$. This is nonlattice by X.3.2 and the mean $\mathbb{E}_{\mathbf{0}}\tau(1)$ is $\mathbb{E}_{\mathbf{0}}\sigma(1)\mathbb{E}T < \infty$. Thus case (a) is just a standard application of regenerative processes and Corollary 2.5.

In case (b), we define

$$\boldsymbol{Q}_t^* = \bigl(Q_t, A_t, R_t^{(1)}, \ldots, R_t^{(s)}\bigr) \in \mathbb{N} \times [0,\infty)^{s+1},$$

where A_t is the backward recurrence time of the arrival process and the $R_t^{(i)}$ are the ordered residual service times just before time t (the residual service time at an empty channel is defined as 0). Then $\{\boldsymbol{Q}_t^*\}$ is a Markov process, and we shall carry out the proof by a slightly tricky application of ideas from VII.2–3. We let ψ be the distribution of $\boldsymbol{Q}^*_{T_0+\cdots+T_{s-1}}$ conditionally upon an initially empty queue and the events $F_0, \ldots, F_{s-1}$ of the proof of Lemma 2.4. Arguing as in the proof of Lemma 2.4, we can find a stopping time τ with $\mathbb{E}\tau < \infty$ such that $\boldsymbol{Q}_t^*$ is distributed according to ψ. Hence, along similar lines as in VII.3.2 or VII.6, it follows that $\{\boldsymbol{Q}_t^*\}$ has a stationary version $\bigl\{\widetilde{\boldsymbol{Q}}_t^*\bigr\}$, and we shall complete the proof by constructing a coupling of $\{\boldsymbol{Q}_t^*\}$ to $\bigl\{\widetilde{\boldsymbol{Q}}_t^*\bigr\}$. First, it follows from VII.2.7 that we can construct a coupling epoch S for $\{A_t\}$, $\bigl\{\widetilde{A}_t\bigr\}$. This means that there exists $r, \widetilde{r}$ such that $S = T_0 + \cdots + T_{r-1} = \widetilde{T}_0 + \cdots + \widetilde{T}_{r-1}$ and $T_{r+k} = \widetilde{T}_{\widetilde{r}+k}$, $k = 0, 1 \ldots$. Then

$$\bigl\{\boldsymbol{V}_{S+T_r+\cdots+T_{r+k-1}}\bigr\}_{k\in\mathbb{N}}, \quad \bigl\{\widetilde{\boldsymbol{V}}_{S+T_r+\cdots+T_{r+k-1}}\bigr\}_{k\in\mathbb{N}} \tag{2.2}$$

are both versions of $\{\boldsymbol{W}_n\}$, and by Harris ergodicity and VII.3.13, there exists a coupling epoch K such that the chains (2.2) at time K have at least one component equal to zero. But this implies that $\{\boldsymbol{Q}_t^*\}$ and $\bigl\{\widetilde{\boldsymbol{Q}}_t^*\bigr\}$ coincide at $S + T_r + \cdots + T_{r+K-1}$ which hence may be taken as the desired coupling epoch. The t.v. convergence of $\{\boldsymbol{Q}_t^*\}$ then immediately implies that of $\{Q_t\}$ and the t.v. convergence of $\{\boldsymbol{V}_t\}$ follows since the distribution of $\boldsymbol{V}_t$ is a function of the distribution of $\boldsymbol{Q}_t^*$. □

Notes The discrete time theory goes back to a remarkable *tour de force* paper by Kiefer and Wolfowitz (1955). The key step here, the construction of regeneration points in Lemma 2.4, can be found in Gnedenko and Kovalenko (1968), whereas Harris ergodicity was proved by Charlot *et al.* (1978). As for the topics of Section 1, the literature contains quite a few mistakes and gaps and should be read with care.

The existence of limits in continuous time under minimal conditions is more difficult; see Foss and Kalashnikov (1991) and Asmussen and Foss (1993).

3 The $GI/M/s$ Queue

The $GI/M/s$ case corresponds to B being exponential, $\overline{B}(x) = \mathrm{e}^{-\delta x}$. As for single–server queues, this is the simplest example and it is notable that for $s > 1$, the $M/G/s$ case does not appear substantially simpler than the $GI/M/s$ case.

Let Y_n denote the queue length just before the arrival of customer n. Then:

Proposition 3.1 *$\{Y_n\}$ is a Markov chain on $\mathbb{N}$. The transition matrix is of the form*

$$\left(\begin{array}{cccccc|cccc}
p_{00} & p_{01} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\
p_{10} & p_{11} & p_{12} & 0 & \cdots & 0 & 0 & 0 & 0 & \\
\vdots & & & & & \vdots & \vdots & & & \vdots \\
p_{(s-2)0} & p_{(s-2)1} & p_{(s-2)2} & p_{(s-2)3} & \cdots & p_{(s-2)(s-1)} & 0 & 0 & 0 & \cdots \\
\hline
p_{(s-1)0} & p_{(s-1)1} & p_{(s-1)2} & p_{(s-1)3} & \cdots & p_{(s-1)(s-1)} & q_0 & 0 & 0 & \cdots \\
p_{s0} & p_{s1} & p_{s2} & p_{s3} & \cdots & p_{s(s-1)} & q_1 & q_0 & 0 & \cdots \\
p_{(s+1)0} & p_{(s+1)1} & p_{(s+1)1} & p_{(s+1)3} & \cdots & p_{(s+1)(s-1)} & q_2 & q_1 & q_0 & \\
\vdots & & & & & & & & & \ddots
\end{array}\right)$$

with elements given by

$$\begin{aligned}
q_k &= \int_0^\infty \mathrm{e}^{-\delta st}\frac{(\delta st)^k}{k!}\,A(\mathrm{d}t), \quad k = 0,1,2,\ldots, && (3.1)\\
p_{ij} &= \int_0^\infty b_{i+1-j}\big(i+1, 1-\mathrm{e}^{-\delta t}\big)\,A(\mathrm{d}t),\ i \le s-1,\ j \le i+1, && (3.2)\\
p_{ij} &= \int_0^\infty A(\mathrm{d}t)\int_0^t b_{s-j}\big(s, 1-\mathrm{e}^{-\delta(t-y)}\big)E_{i+1-s}(\mathrm{d}y),\ i \ge s > j, && (3.3)
\end{aligned}$$

where E_k denotes the Erlang distribution with k stages and intensity δs and $b_k(n,p)$ is the binomial probability $\binom{n}{k}p^k(1-p)^{n-k}$.

Proof. This is seen by arguments that are similar to the case $s = 1$ in III.6.2 but more elaborate. With K_n the number of customers being served between the arrival of customers n and $n+1$, we have $Y_{n+1} = (Y_n + 1 - K_n)^+$ so clearly $p_{ij} = \mathbb{P}_i(K_0 = i+1-j)$. In the following let $Y_0 = i$. Then from $K_0 \ge 0$ it is clear that $p_{ij} = 0$ when $j > i+1$. Also if $i \ge s-1$, $j \ge s$, then $Y_1 = j$ mean that all servers were busy and performed a total of $k = i+1-j$ services before $T_0 = t$. The probability of this is obviously q_k and hence $p_{ij} = q_{i+1-j}$. For (3.2), note that if $i \le s-1$, then all $i+1$ customers present at time 0 receive service immediately. Thus conditionally upon $T_0 = t$, the distribution of K_0 is binomial with parameters $(i+1, 1-\mathrm{e}^{-\delta t})$ and (3.2) follows. Finally for (3.3), note first that if $i+1 > s$, then in order for $Y_1 = j < s$ the waiting line must disappear at some time $y < T_0 = t$, say. This is equivalent to $S = y < T_0 = t$ where S has distribution E_{i+1-s} and is independent of T_0. After time y, the s servers then need to complete $s-j$ services in $(y,t]$. Since the probability of this obviously is $b_{s-j}\big(s, 1-\mathrm{e}^{-\delta(t-y)}\big)$, (3.3) follows. □

The next step is to derive the stationary distribution of $\{Y_n\}$:

Theorem 3.2 *The Markov chain $\{Y_n\}$ has a stationary distribution π if and only if $\rho = (s\delta\mu_A)^{-1} < 1$. In that case, π may be computed by $\pi_i = C\nu_i$ where:*
(i) $\nu_i = \theta^i$, $i \geq s-1$, *with θ the unique solution in $(0,1)$ of the equation*

$$\theta \;=\; \int_0^\infty \mathrm{e}^{-\delta s(1-\theta)y}\,A(\mathrm{d}y); \tag{3.4}$$

(ii) $\nu_{s-2}, \nu_{s-3}, \ldots, \nu_0$ *are recursively determined by*

$$\nu_j \;=\; \frac{1}{p_{j(j+1)}}\Bigl\{\nu_{j+1}(1-p_{(j+1)(j+1)}) - \sum_{i=j+2}^{\infty}\nu_i p_{i(j+1)}\Bigr\}, \quad j = s-2,\ldots,0; \tag{3.5}$$

(iii) $C^{-1} \;=\; \nu_0 + \cdots + \nu_{s-2} + (1-\theta)^{-1}\theta^{s-1}$.

An equivalent formulation of (i) is that the number of customers awaiting service is geometrically distributed (given that there are such customers).

Proof of Theorem 3.2. Define $E = \{s-1, s, s+1, \ldots\}$. Then once $\{0,1,\ldots,s-2\}$ is entered, the next visit of $\{Y_n\}$ to E occurs necessarily at state $s-1$. Letting $\boldsymbol{Q}$ be the transition matrix of the Markov chain obtained by restricting $\{Y_n\}$ to E, it follows that

$$\boldsymbol{Q} \;=\; \begin{pmatrix} r_0 & q_0 & 0 & 0 & \cdots \\ r_1 & q_1 & q_0 & 0 & \cdots \\ r_2 & q_2 & q_1 & q_0 & \cdots \\ \vdots & & & & \ddots \end{pmatrix},$$

where $r_n = 1 - q_0 - \cdots - q_n$. This is as the same form as in III.6.2 and X.5, and we may infer immediately that $\boldsymbol{Q}$ is not positive recurrent when $\rho \geq 1$. Hence $\boldsymbol{P}$ cannot be so either. Conversely, when $\rho < 1$ a (finite) stationary measure $\boldsymbol{\nu}$ for $\boldsymbol{Q}$ exists and may be taken of the form in (i). By I.3.9, $\boldsymbol{\nu}$ has a unique extension to a stationary measure for $\boldsymbol{P}$, and considering the $(j+1)$th entry of the equation $\boldsymbol{\nu P} = \boldsymbol{\nu}$ for $j \leq s-2$ yields

$$\nu_{j+1} \;=\; \sum_{i=0}^{\infty}\nu_i p_{i(j+1)} \;=\; \sum_{i=j}^{\infty}\nu_i p_{i(j+1)},$$

which implies (3.5). Since $\nu_{s-1}, \nu_s, \ldots$ are known, (3.5) can then be solved for ν_{s-2} and we may repeat the argument to get $\nu_{s-3}, \ldots, \nu_0$. Finally, it is clear that for $\boldsymbol{\pi} = C\boldsymbol{\nu}$ to be a stationary distribution we simply have to let $C^{-1} = |\boldsymbol{\nu}|$, and this is equivalent to (iii). □

The remaining steady–state characteristics can now easily be found. Consider first the queue length Q at an arbitrary point of time:

Corollary 3.3 *If $\rho < 1$ and A is nonlattice, then $\pi_k^* = \lim_{t\to\infty} \mathbb{P}(Q_t = k)$ exists and is given by*

$$\pi_k^* = \begin{cases} \dfrac{\pi_{k-1}}{k\delta\mu_A} & k = 1,\dots,s \\ \rho\pi_{k-1} & k \geq s \end{cases},$$

$$\pi_0^* = 1 - \pi_1^* - \pi_2^* - \cdots = 1 - \rho - \frac{1}{\delta\mu_A}\sum_{k=1}^{s-1}\pi_{k-1}\Bigl(\frac{1}{k} - \frac{1}{s}\Bigr).$$

Proof. Existence follows immediately from Corollary 2.8(a). To derive the form of π_k^*, we use rate conservation exactly as in X.5. Let $k > 0$ and $X_t = I(Q_t \geq k)$. The upcrossing rate is the arrival rate μ_A^{-1} times the probability π_{k-1} that the state just before the arrival is $k-1$. The downcrossing rate is $s\delta\mathbb{P}(Q = k)$ when $k \geq s$ and $k\delta\mathbb{P}(Q = k)$ when $k \leq s$. Equating these two expressions immeditely yields the stated expression for π_k^* when $k \geq 1$, and the case $k = 0$ follows from this by easy algebra. □

Corollary 3.4 *If $\rho < 1$, then the waiting time process $\bigl\{W_n^{(1)}\bigr\}_{n\in\mathbb{N}}$ has a t.v. limit which is a mixture with weights $\zeta = \pi_0 + \cdots + \pi_{s-1}$ and $1 - \zeta$ of an atom at 0 and the exponential distribution with intensity $\eta = s\delta(1-\theta)$.*

Proof. An arriving customer has to wait if and only if he sees $Y = s$ or more customers upon arrival. Thus the atom at 0 has obviously weight $\mathbb{P}(Y \leq s-1) = \pi_0 + \cdots + \pi_{s-1}$ in steady state. If $Y \geq s$, the customer has to wait until $Y - s + 1$ services are completed, i.e. an Erlang time with $Y - s + 1$ stages and intensity $s\delta$. Since the distribution of $Y - s$ given $Y \geq s$ is geometric with parameter θ, the distribution of $W^{(1)}$ given $W^{(1)} > 0$ can therefore be evaluated as for the case $s = 1$ as a geometric mixture of Erlangs, and this immediately leads to the conclusion of the corollary. □

Problems

3.1 Explain that queue lengths just after departure times in $M/G/s$ do not form a Markov chain when $s > 1$.

Notes For algorithmic solutions of $GI/PH/s$ queues, see Neuts (1981) for queue length distributions and de Smit (1995) and Asmussen and Møller (2001) for waiting time distributions.

XIII
Exponential Change of Measure

1 Exponential Families

We consider i.i.d. r.v.'s $X_1, X_2, \ldots$ with common distribution F and c.g.f. (cumulant generating function, cf. the Notation and Conventions and A9)

$$\kappa(\alpha) \;=\; \log \mathbb{E}\mathrm{e}^{\alpha X_1} \;=\; \log \int_{-\infty}^{\infty} \mathrm{e}^{\alpha x}\, F(\mathrm{d}x).$$

For each $\theta \in \Theta = \{\theta \in \mathbb{R} : \kappa(\theta) < \infty\}$, we denote by F_θ the probability distribution with density $\mathrm{e}^{\theta x - \kappa(\theta)}$ w.r.t. F. In standard statistical terminology, $(F_\theta)_{\theta \in \Theta}$ is the *exponential family generated by* F. Similarly, $\mathbb{P}_\theta$ denotes the probability measure w.r.t. which $X_1, X_2, \ldots$ are i.i.d. with common distribution F_θ.

Proposition 1.1 *Let* $\kappa_\theta(\alpha) = \log \mathbb{E}_\theta \mathrm{e}^{\alpha X_1}$ *be the c.g.f. of* F_θ*. Then*

$$\kappa_\theta(\alpha) = \kappa(\alpha + \theta) - \kappa(\theta), \quad \mathbb{E}_\theta X_1 = \kappa'(\theta), \quad \mathbb{V}ar_\theta X_1 = \kappa''(\theta).$$

Proof. The formula for $\kappa_\theta(\alpha)$ follows from

$$\mathrm{e}^{\kappa_\theta(\alpha)} \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{\alpha x} F_\theta(\mathrm{d}x) \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{(\alpha+\theta)x - \kappa(\theta)} F(\mathrm{d}x) \;=\; \mathrm{e}^{\kappa(\alpha+\theta) - \kappa(\theta)}.$$

We then get $\mathbb{E}_\theta X_1 = \kappa_\theta'(0) = \kappa'(\theta)$, $\mathbb{V}ar_\theta X_1 = \kappa_\theta''(0) = \kappa''(\theta)$. □

Example 1.2 Let F be the normal distribution with mean μ and variance σ^2. Then $\Theta = \mathbb{R}$, $\kappa(\alpha) = \mu\alpha + \sigma^2\alpha^2/2$ so that

$$\kappa_\theta(\alpha) \;=\; \mu(\alpha+\theta) + \sigma^2(\alpha+\theta)^2/2 \;-\; \mu\theta - \sigma^2\theta^2/2 \;=\; (\mu + \theta\sigma^2)\alpha + \sigma^2\alpha^2/2$$

which shows that F_θ is the normal distribution with mean $\mu+\theta\sigma^2$ and the same variance σ^2. □

Example 1.3 Let F be the exponential distribution with rate λ. Then $\Theta = (-\infty,\lambda)$, $\kappa(\alpha)=\log\lambda-\log(\lambda-\alpha)$ so that $\kappa_\theta(\alpha)=\log(\lambda-\theta)-\log(\lambda-\theta-\alpha)$. That is, F_θ is the exponential distribution with rate $\lambda-\theta$. □

If $\mathbb{E}_\theta$ is the expectation operator corresponding to $\mathbb{P}_\theta$, then for any fixed n

$$\mathbb{E}_\theta f(X_1,\ldots,X_n) \quad=\quad \mathbb{E}\big[\mathrm{e}^{\theta S_n-n\kappa(\theta)}f(X_1,\ldots,X_n)\big] \tag{1.1}$$

for all measurable $f:\mathbb{R}^n\to\mathbb{R}$ which are bounded or nonnegative, where $S_n=X_1+\cdots+X_n$. This follows since the l.h.s. of (1.1) is

$$\begin{aligned}
&\int\cdots\int_{\mathbb{R}^n} f(x_1,\ldots,x_n)\,F_\theta(\mathrm{d}x_1)\ldots F_\theta(\mathrm{d}x_n)\\
&=\quad \int\cdots\int_{\mathbb{R}^n} f(x_1,\ldots,x_n)\,\mathrm{e}^{\theta x_1-\kappa(\theta)}F(\mathrm{d}x_1)\ldots\mathrm{e}^{\theta x_n-\kappa(\theta)}F(\mathrm{d}x_n)\\
&=\quad \int\cdots\int_{\mathbb{R}^n} f(x_1,\ldots,x_n)\,\mathrm{e}^{\theta s_n-n\kappa(\theta)}F(\mathrm{d}x_1)\cdots F(\mathrm{d}x_n)
\end{aligned}$$

which is the same as the r.h.s. (here $s_n = x_1+\cdots+x_n$). Replacing first $f(x_1,\ldots,x_n)$ by $e^{-\theta s_n+n\kappa(\theta)}f(x_1,\ldots,x_n)$ and specializing next to an indicator function of a Borel set $A\subseteq\mathbb{R}^n$, we get

$$\mathbb{E}f(X_1,\ldots,X_n) \quad=\quad \mathbb{E}_\theta\big[\mathrm{e}^{-\theta S_n+n\kappa(\theta)}f(X_1,\ldots,X_n)\big], \tag{1.2}$$

$$\mathbb{P}(A) \quad=\quad \mathbb{E}_\theta\big[\mathrm{e}^{-\theta S_n+n\kappa(\theta)};\,A\big]. \tag{1.3}$$

Thus we have expressed $\mathbb{P}(A)$ as an expectation corresponding to an i.i.d. sum with a changed increment distribution. At a first sight this appears to be a complicated way of evaluating $\mathbb{P}(A)$. The point is that in a number of cases the F_θ–distribution has more convenient properties than F, as will be demonstrated by a number of applications throughout the chapter.

Example 1.4 In the theory of the $GI/G/1$ queue, a fundamental distribution is $F(x)=\mathbb{P}(U-T\le x)$ with U,T being independent and having the service time distribution B, resp. the interarrival distribution A. Letting $\kappa^{(A)}(\cdot)$, $\kappa^{(B)}(\cdot)$, $\kappa(\cdot)$ be the c.g.f.'s of A, B, resp. F, we have $\kappa(\alpha)=\kappa^{(A)}(-\alpha)+\kappa^{(B)}(\alpha)$. If we define A_θ, B_θ by

$$A_\theta(\mathrm{d}x)=\mathrm{e}^{-\theta x-\kappa^{(A)}(-\theta)}A(\mathrm{d}x),\quad B_\theta(\mathrm{d}x)=\mathrm{e}^{\theta x-\kappa^{(B)}(\theta)}B(\mathrm{d}x),$$

it follows immediately that (in obvious notation) $\kappa_\theta(\alpha)=\kappa_\theta^{(A)}(-\alpha)+\kappa_\theta^{(B)}(\alpha)$. Thus, $\mathbb{P}_\theta$ corresponds to a $GI/G/1$ queue with interarrival distribution A_θ and service time distribution B_θ. Furthermore, it is straightforward to extend (1.3) to $2n$–dimensional sets (the first n coordinates specifying the first n interarrival times and the last the first n service times). □

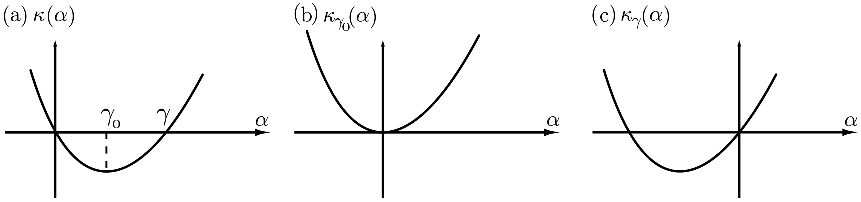

Figure 1.1

The basic case for queueing theory is F having negative mean μ, but not being concentrated on $(-\infty, 0]$. The typical shape of $\kappa(\cdot)$ is illustrated in Fig. 1.1(a). The slope at 0 is $\mu < 0$ and κ is strictly convex. Since $\mathrm{supp}(F)\cap(0,\infty) \neq \emptyset$, $\kappa(\alpha) \to \infty$ as $\alpha \to \infty$. Quite often (say the exponential or Gamma distribution) $\kappa(\cdot)$ has a finite radius of convergence, and in the case of heavy tails we even have $\kappa(\alpha) = \infty$ for all $\alpha > 0$ (this situation is therefore excluded from the analysis of most of the chapter). However, for light–tailed distributions the picture is typically as in Fig. 1.1(a). Of particular importance for the following are γ_0, the solution of $\kappa'(\gamma_0) = 0$, and γ, the positive solution of the *Lundberg equation* $\kappa(\gamma) = 0$. The shape of $\kappa_{\gamma_0}(\cdot)$ and $\kappa_\gamma(\cdot)$ is indicated in Fig. 1.1(b) and (c). Note in particular the simple geometric relation to $\kappa(\cdot)$, and that $\mathbb{E}_{\gamma_0}X_1 = \kappa'(\gamma_0) = 0$ and (by convexity) $\mathbb{E}_\gamma X_1 = \kappa'(\gamma) > 0$. □

Example 1.5 Consider $M/M/1$ with A, B having intensities $\beta < \delta$. Then A_θ, B_θ are exponential with densities $\beta_\theta = \beta+\theta$, $\delta_\theta = \delta-\theta$. The Lundberg equation is

$$\log\frac{\beta}{\beta+\gamma} + \log\frac{\delta}{\delta-\gamma} \;=\; 0$$

and the solution $\gamma > 0$ is $\gamma = \delta - \beta$. Thus $\beta_\gamma = \delta$, $\delta_\gamma = \beta$, and $\mathbb{P}_\gamma$ corresponds to interchanging β and δ. □

Problems

1.1 Verify Examples 1.2, 1.3 using densities rather than c.g.f.'s.
1.2 Show that for the $M/G/1$ queue, the Lundberg equation is equivalent to $\beta(\widehat{B}[\gamma]-1)-\gamma = 0$. [See further VII.7.8.]
1.3 Show that F_θ is lattice if and only if F is lattice, spread out if and only if F is spread out and satisfies Cramér's condition (C) if and only if F does so [hint: if $|\widehat{F}[t_k]| \to 1$, then $\cos t_k x \to 1$ for a.a. x w.r.t. to the symmetrized distribution with ch.f. $\widehat{F}^*[t] = \widehat{F}[t]\widehat{F}[-t]$. Hence $\widehat{F}^*_\theta[t_k] \to 1$ and $|\widehat{F}_\theta[t_k]| \to 1$].

Notes Exponential families occur widely in statistics (usually the given object is the whole family (F_θ) and not as here the single member $F = F_0$); see e.g. Barndorff–Nielsen (1978).

2 Large Deviations, Saddlepoints and the Relaxation Time

Define the *convex conjugate* $\kappa^*(\cdot)$ by $\kappa^*(x) = \theta(x)x - \kappa\bigl(\theta(x)\bigr)$ where $\theta(x)$ is the solution of

$$x \;=\; \kappa'\bigl(\theta(x)\bigr) \;=\; \mathbb{E}_{\theta(x)}X_1 \tag{2.1}$$

($\kappa^*(x) = \infty$ if no solution exists).

Theorem 2.1 *Let $x > \kappa'(0) = \mathbb{E}X_1$ and assume that* (2.1) *has a solution $\theta = \theta(x)$. Then*

$$\begin{aligned}
\mathbb{P}(S_n > nx) \;&\leq\; \mathrm{e}^{-n\kappa^*(x)}, && (2.2)\\
\frac{1}{n}\log \mathbb{P}(S_n > nx) \;&\to\; -\kappa^*(x), \quad x \to \infty, && (2.3)\\
\mathbb{P}(S_n > nx) \;&\sim\; \frac{1}{\theta\sqrt{2\pi\sigma_\theta^2 n}}\mathrm{e}^{-n\kappa^*(x)}, \quad x \to \infty, && (2.4)
\end{aligned}$$

provided in addition for (2.3) *that $\sigma_\theta^2 = \kappa''(\theta) < \infty$ and for* (2.4) *that $|\kappa'''(\theta)| < \infty$ and that F satisfies Cramér's condition* (C).

Proof. Taking $A = \{S_n > nx\}$, $\theta = \theta(x)$ in (1.3), we get

$$\mathbb{P}(S_n > nx) \;=\; \mathrm{e}^{-n\kappa^*(x)}\mathbb{E}_\theta\bigl[\mathrm{e}^{-\theta(S_n - nx)};\, S_n > nx\bigr]. \tag{2.5}$$

Since $\theta > 0$ and $S_n - nx > 0$ on A, (2.2) follows by bounding the indicator of A by 1.

When $\sigma_\theta^2 < \infty$, the CLT yields

$$\mathbb{P}_\theta\bigl(nx < S_n < nx + 1.96\sigma_\theta\sqrt{n}\bigr) \;\to\; \Phi(1.96) - \Phi(0) \;=\; 0.425.$$

Hence for large n,

$$\begin{aligned}
\mathbb{P}(S_n/n > x) \;&\geq\; \mathrm{e}^{-n\kappa^*(x)}\mathbb{E}_\theta\Bigl[\mathrm{e}^{-\theta(S_n - nx)};\, nx < S_n < nx + 1.96\sigma_\theta\sqrt{n}\Bigr]\\
&\geq\; \mathrm{e}^{-n\kappa^*(x)} \cdot 0.4\mathrm{e}^{1.96\sigma_\theta\sqrt{n}}.
\end{aligned}$$

Combining with (2.2) and taking logarithms, (2.3) follows.

Let $H_n(y) = \mathbb{P}_\theta(S_n - nx \leq y\sigma_\theta\sqrt{n})$. Then H_n satisfies (C), cf. Problem 1.3, and the third moment exists because of $|\kappa'''(\theta)| < \infty$. Hence (Bhattacharya and Rao, 1976, Theorem 20.1) we may write

$$H_n(\mathrm{d}y) \;=\; \Phi(\mathrm{d}y) + n^{-1/2}f'(y)\mathrm{d}y + G_n(\mathrm{d}y) \tag{2.6}$$

where $f(y) = \eta(1-y^2)\mathrm{e}^{-y^2/2}$ for some constant η and G_n is a (signed) measure with $\|G_n\| = \mathrm{o}(n^{-1/2})$. Thus with $g(y) = \mathrm{e}^{-\theta\sigma_\theta y}$ we have

$$\mathbb{E}_\theta\bigl[\mathrm{e}^{-\theta(S_n - nx)};\, S_n > nx\bigr] \;=\; \int_0^\infty g(yn^{1/2})\,H_n(\mathrm{d}y) \tag{2.7}$$

$$= \;\int_0^\infty g(yn^{1/2})\,\Phi(\mathrm{d}y) \;+\; n^{-1/2}\int_0^\infty g(yn^{1/2})f'(y)\,\mathrm{d}y \;+\; \mathrm{O}(\|G_n\|).$$

Here $g(yn^{1/2}) \to 0$ so that by dominated convergence $\int gf' = \mathrm{o}(1)$. Hence, using dominated convergence once more, (2.7) becomes

$$\begin{aligned}
&\int_0^\infty g(yn^{1/2})\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2}\,\mathrm{d}y \;+\; \mathrm{o}(n^{1/2}) \\
&= \;\frac{1}{\sqrt{2\pi n}}\int_0^\infty \mathrm{e}^{-\theta\sigma_\theta y}\mathrm{e}^{-y^2/2n}\,\mathrm{d}y \;+\; \mathrm{o}(n^{1/2}) \\
&= \;\frac{1}{\sqrt{2\pi n}}\Bigl(\int_0^\infty \mathrm{e}^{-\theta\sigma_\theta y}\,\mathrm{d}y \;+\; \mathrm{o}(1)\Bigr) \;=\; \frac{(\theta\sigma_\theta)^{-1}}{\sqrt{2\pi n}}\bigl(1+o(1)\bigr).
\end{aligned}$$

□

All three parts of Theorem 2.1 are classical in probability theory; see the Notes. A relatively small variant of the proof produces a highly relevant queueing result. Recall the definition $\kappa'(\gamma_0) = 0$ of γ_0.

Theorem 2.2 *Consider the waiting time process $W_0, W_1, \ldots$ of a stable ($\rho < 1$) $GI/G/1$ queue and let $\delta = \mathrm{e}^{\kappa(\gamma_0)}$, $\sigma^2 = \kappa''(\gamma_0)$. Then if $F(x) = \mathbb{P}(U - T \le x)$ satisfies the assumptions for* (2.4),

$$\mathbb{E}W - \mathbb{E}W_{N-1} \;\sim\; \frac{\delta^N}{N^{3/2}}\cdot\frac{1}{\gamma_0^2(1-\delta)\sqrt{2\pi\sigma^2}}, \quad N \to \infty. \tag{2.8}$$

Proof. It follows from VIII.4.5 that $\mathbb{E}W - \mathbb{E}W_{N-1} = \sum_N^\infty \mathbb{E}S_n^+/n$. Here exactly as above, with $x = 0$, $\theta = \gamma_0$, we have $\mathbb{E}S_n^+ = \delta^n J_n$ where

$$J_n \;=\; \mathbb{E}_{\gamma_0}\bigl[S_n\mathrm{e}^{-\gamma_0 S_n};\, S_n > 0\bigr] \;=\; \int_0^\infty h(xn^{1/2})\,F_n(\mathrm{d}x)$$

where $h(x) = \sigma x\mathrm{e}^{-\gamma_0\sigma x}$ is bounded with $h(xn^{1/2}) \to 0$, $n \to \infty$. Thus exactly as above, we have up to $\mathrm{o}(n^{-1/2})$ terms that

$$\begin{aligned}
J_n \;&=\; \int_0^\infty h(xn^{1/2})\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x \;=\; \frac{\sigma}{\sqrt{2\pi n}}\int_0^\infty y\mathrm{e}^{-\gamma_0\sigma y}\mathrm{e}^{-y^2/2n}\,\mathrm{d}y \\
&=\; \frac{\sigma}{\sqrt{2\pi n}}\int_0^\infty y\mathrm{e}^{-\gamma_0\sigma y}\,\mathrm{d}y \;=\; \frac{1}{\gamma_0^2\sqrt{2\pi\sigma^2 n}}\,,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}W - \mathbb{E}W_{N-1} \;&=\; \sum_{n=N}^\infty \frac{\delta^n J_n}{n} \;\sim\; \frac{1}{\gamma_0^2\sqrt{2\pi\sigma^2}}\sum_{n=N}^\infty \frac{\delta^n}{n^{3/2}} \\
&\sim\; \frac{\delta^N}{N^{3/2}}\cdot\frac{1}{\gamma_0^2(1-\delta)\sqrt{2\pi\sigma^2}}\,.
\end{aligned}$$

□

Obviously, (2.8) is of the same form and spirit as the relaxation time approximation for $M/M/1$ in III.8e (also, the proof in III.8e is based upon similar methods; cf. in particular III.8.11). It is natural to ask whether there is an analogue for the distribution of W_N rather than its mean, and indeed:

Theorem 2.3 *Under the assumptions of Theorem 2.2,*

$$\mathbb{P}(W_N \le x) - \mathbb{P}(W \le x) \;\sim\; \frac{\delta^N}{N^{3/2}} \mathrm{e}^{-\gamma_0 x} U_+^{(0)}(x) c, \quad N \to \infty, \tag{2.9}$$

where $U_+^{(0)}$ is the ascending ladder height renewal measure w.r.t. $\mathbb{P}_{\gamma_0}$ and

$$c \;=\; \frac{\delta^{3/2}\bigl[1 - \mathbb{E}_{\gamma_0} e^{-\gamma_0 S_{\tau_+}}\bigr]}{(1-\delta)\gamma_0\sqrt{2\pi\sigma^2}}\,.$$

The proof is substantially more involved and will not be given here.

Problems

2.1 Show that under the conditions of Theorem 2.1, the limiting distribution of S_n given $S_n > 0$ is exponential. [*Hint:* Laplace transforms.]

Notes We will not attempt to trace the history of Theorem 2.1. However, (2.2) goes under the name the *Chernoff bound*, (2.3) is of *large deviations* (LD) type, and (2.4) is the *saddlepoint approximation.*

Obviously, (2.4) is sharper than (2.3), and in particular, (2.3) does not allow to assert whether the exact asymptotic form of $\mathbb{P}(S_n > nx)$ is simply exponential, or contains a correction factor of slower variation like the $n^{-1/2}$ in (2.4) (other possibilities are not excluded by (2.3), for example $\mathbb{P}(S_n > nx) \sim \mathrm{e}^{-n\kappa^*(x) - n^\epsilon}$ with $\epsilon < 1$). This is a typical feature of LD theory, that the results are in logarithmic asymptotics rather than exact asymptotics form and therefore subject to refinement. However, often the logarithmic asymptotics gives sufficient information on the problem under study, and the advantage of LD theory is then the generality into which the subject has been pushed. Particularly notable results are the *Gärtner–Ellis theorem*, stating that the independence assumptions for (2.3) can be relaxed to the existence of $\kappa(\alpha) = \lim_{n\to\infty} \log \mathbb{E}\mathrm{e}^{\alpha S_n}/n$ (and some further weak regularity), and *Mogulskii'stheorem*, stating that if $S^{(n)}$ is the random function on $[0,1]$ given by $S_t^{(n)} = S_{\lfloor nt \rfloor}/n$, then for a smooth function f on $[0,1]$,

$$\frac{1}{n} \log \mathbb{P}(S^{(n)} \approx f) \;\to\; \int_0^1 \kappa^*(f'(t))\,\mathrm{d}t \tag{2.10}$$

with $\approx$ defined in a suitable sense. Of course, (2.10) is strongly suggested by (2.3): if f is linear with slope x in the interval $[t_0, t_0 + \epsilon)$, then (after trivial substitutions) (2.3) asserts that the probability that $S^{(n)}$ has slope x in the same interval is $\epsilon\kappa^*(x)$ in the logaritmic sense so that a simple Riemann approximation leads to (2.10).

Among many textbooks on LD, we mention in particular Dembo and Zeitouni (1998) and, for queueing applications, Shwartz and Weiss (1995). For saddlepoint approximations, see Jensen (1995) and references there.

Theorem 2.2 is from Heathcote (1967) and Heathcote and Winer (1969), Theorem 2.3 from Veraverbeke and Teugels (1975/76) (with the constants rewritten somewhat here). For continuous–time analogues, see Teugels (1977, 1982).

3 Change of Measure: General Theory

We consider stochastic processes $\{Z_t\}$ with a Polish state space E and sample paths in the Skorokhod space $D = D\bigl([0,\infty), E\bigr)$, which we equip with the natural filtration $\{\mathscr{F}_t\}_{t\geq 0}$ and the Borel σ–field $\mathscr{F}$ (the discrete–time case is entirely similar). Two such processes may be represented by probability measures $\mathbb{P}$, $\widetilde{\mathbb{P}}$ on $(D, \mathscr{F})$, and in analogy with the theory of measures on finite–dimensional spaces, one could study conditions for the Radon–Nikodym derivative $\mathrm{d}\widetilde{\mathbb{P}}/\mathrm{d}\mathbb{P}$ to exist. However, this set–up is too restrictive. Consider, e.g. a random walk $\{S_n\}$ with increment distributions $F, \widetilde{F}$ with means $\mu \neq \widetilde{\mu}$ and let $A = \{S_n/n \to \mu\}$, $\widetilde{A} = \{S_n/n \to \widetilde{\mu}\}$. Then $\mathbb{P}$ is concentrated on A and $\widetilde{\mathbb{P}}$ on $\widetilde{A}$ which are disjoint, excluding absolute continuity (this extends to all pairs of processes where the parameters of the two processes can be reconstructed from a single infinite path).

The interesting concept is therefore to look for absolute continuity only on finite time intervals (possibly random, cf. Theorem 3.2 below). That is, we look for a process $\{L_t\}$ (the likelihood ratio process) such that

$$\widetilde{\mathbb{P}}(A) = \mathbb{E}[L_t; A], \quad A \in \mathscr{F}_t, \tag{3.1}$$

(i.e. that the restriction of $\widetilde{\mathbb{P}}$ to $(D, \mathscr{F}_t)$ is absolutely continuous w.r.t. the restriction of $\mathbb{P}$ to $(D, \mathscr{F}_t)$). The following result gives the connection to martingales.

Proposition 3.1 *Let $\{\mathscr{F}_t\}_{t\geq 0}$ be the natural filtration on D, $\mathscr{F}$ the Borel σ–field and $\mathbb{P}$ a given probability measure on $(D, \mathscr{F})$.*
(i) *If $\{L_t\}_{t\geq 0}$ is a nonnegative martingale w.r.t. $\bigl(\{\mathscr{F}_t\}, \mathbb{P}\bigr)$ such that $\mathbb{E}L_t = 1$, then there exists a unique probability measure $\widetilde{\mathbb{P}}$ on $\mathscr{F}$ such that* (3.1) *holds.*
(ii) *Conversely, if for some probability measure $\widetilde{\mathbb{P}}$ and some $\{\mathscr{F}_t\}$–adapted process $\{L_t\}_{t\geq 0}$* (3.1) *holds, then $\{L_t\}$ is a nonnegative martingale w.r.t. $\bigl(\{\mathscr{F}_t\}, \mathbb{P}\bigr)$ such that $\mathbb{E}L_t = 1$.*

Proof. Under the assumptions of (i), define $\widetilde{\mathbb{P}}$ by $\widetilde{\mathbb{P}}_t(A) = \mathbb{E}[L_t; A]$, $A \in \mathscr{F}_t$. Then $L_t \geq 0$ and $\mathbb{E}L_t = 1$ ensure that $\widetilde{\mathbb{P}}_t$ is a probability measure on $(D, \mathscr{F}_t)$. Let $s < t$, $A \in \mathscr{F}_s$. Then

$$\begin{aligned}\widetilde{\mathbb{P}}_t(A) &= \mathbb{E}[L_t; A] = \mathbb{E}\,\mathbb{E}\bigl[L_t I(A)\big|\mathscr{F}_s\bigr] = \mathbb{E}\left[I(A)\mathbb{E}[L_t|\mathscr{F}_s]\right] \\ &= \mathbb{E}\,I(A)L_s = \widetilde{\mathbb{P}}_s(A),\end{aligned}$$

using the martingale property in the fourth step. Hence the family $\bigl\{\widetilde{\mathbb{P}}_t\bigr\}_{t\geq 0}$ is consistent in the Kolmogorov sense and hence extendable to a probability measure $\widetilde{\mathbb{P}}$ on $(D, \mathscr{F})$ such that $\widetilde{\mathbb{P}}(A) = \widetilde{\mathbb{P}}_t(A)$, $A \in \mathscr{F}_t$. This proves (i).

Conversely, under the assumptions of (ii) we have for $A \in \mathscr{F}_s$ and $s < t$ that $A \in \mathscr{F}_t$ as well and hence $\mathbb{E}[L_s; A] = \mathbb{E}[L_t; A]$. The truth of this for all $A \in \mathscr{F}_s$ implies that $\mathbb{E}[L_t|\mathscr{F}_s] = L_s$ and the martingale property. Finally,

$\mathbb{E}L_t = 1$ follows by taking $A = D$ in (3.1), and nonnegativity by letting $A = \{L_t < 0\}$. Then $\widetilde{\mathbb{P}}(A) = \mathbb{E}[L_t; L_t < 0]$ can only be nonnegative if $\mathbb{P}(A) = 0$. □

The following likelihood ratio identity is a fundamental tool throughout this chapter:

Theorem 3.2 *Let $\{L_t\}$, $\widetilde{\mathbb{P}}$ be as in Proposition 3.1(i). If τ is a stopping time and $G \in \mathscr{F}_\tau$, $G \subseteq \{\tau < \infty\}$, then*

$$\mathbb{P}(G) \;=\; \widetilde{\mathbb{E}}\Big[\frac{1}{L_\tau}; G\Big]. \tag{3.2}$$

More generally, if $W \geq 0$ is $\mathscr{F}_\tau$–measurable, then $\mathbb{E}[W; \tau < \infty] = \widetilde{\mathbb{E}}[W/L_\tau; \tau < \infty]$.

Proof. Assume first $G \subseteq \{\tau \leq T\}$ for some fixed deterministic $T < \infty$. By the martingale property, we have $\mathbb{E}[L_T|\mathscr{F}_\tau] = L_\tau$ on $\{\tau \leq T\}$. Hence

$$\begin{aligned}\widetilde{\mathbb{E}}\Big[\frac{1}{L_\tau}; G\Big] \;&=\; \mathbb{E}\Big[\frac{L_T}{L_\tau}; G\Big] \;=\; \mathbb{E}\Big[\frac{1}{L_\tau}I(G)\,\mathbb{E}[L_T|\mathscr{F}_\tau]\Big] \\ &=\; \mathbb{E}\Big[\frac{1}{L_\tau}I(G)L_\tau\Big] \;=\; \mathbb{P}(G).\end{aligned} \tag{3.3}$$

In the general case, applying (3.3) to $G \cap \{\tau \leq T\}$ we get

$$\mathbb{P}\big(G \cap \{\tau \leq T\}\big) \;=\; \widetilde{\mathbb{E}}\Big[\frac{1}{L_\tau}; G \cap \{\tau \leq T\}\Big].$$

Since everything is nonnegative, both sides are increasing in T, and letting $T \to \infty$, (3.2) follows by monotone convergence. The last statement follows by standard measure theory. □

A crucial step in specific cases is to obtain information on the process evolving according to $\widetilde{\mathbb{P}}$. First we ask when the Markov property is preserved. To this end, we need the concept of a *multiplicative functional.* Assume that $\{Z_t\}$ is Markov w.r.t. the natural filtration $\{\mathscr{F}_t\}$ on D and define $\{L_t\}$ to be a multiplicative functional if $\{L_t\}$ is adapted to $\{\mathscr{F}_t\}$ and

$$L_{t+s} \;=\; L_t \cdot (L_s \circ \theta_t) \tag{3.4}$$

$\mathbb{P}_x$–a.s. for all x, s, t, where θ_t is the shift operator. The precise meaning of this is the following: being $\mathscr{F}_t$–measurable, L_t has the form $L_t = \varphi_t\big(\{Z_u\}_{0 \leq u \leq t}\big)$ for some mapping $\varphi_t : D[0,t] \to [0,\infty)$, and then $L_s \circ \theta_t = \varphi_s\big(\{Z_{t+u}\}_{0 \leq u \leq s}\big)$.

Theorem 3.3 *Let $\{Z_t\}$ be Markov w.r.t. the natural filtration $\{\mathscr{F}_t\}$ on D, let $\{L_t\}$ be a nonnegative martingale with $\mathbb{E}_x L_t = 1$ for all x, t and let $\widetilde{\mathbb{P}}_x$ be the probability measure given by $\widetilde{\mathbb{P}}_x(A) = \mathbb{E}_x[L_t; A]$. Then the family $\big\{\widetilde{\mathbb{P}}_x\big\}_{x \in E}$ defines a time–homogeneous Markov process if and only if $\{L_t\}$ is a multiplicative functional.*

Proof. Since both sides of (3.4) are $\mathscr{F}_{t+s}$ measurable, (3.4) is equivalent to

$$\mathbb{E}_x[L_{t+s}V_{t+s}] \;=\; \mathbb{E}_x\bigl[L_t\cdot(L_s\circ\theta_t)V_{t+s}\bigr] \tag{3.5}$$

for any $\mathscr{F}_{t+s}$–measurable r.v. V_{t+s}, which in turn (by standard measure theory) is the same as

$$\mathbb{E}_x\bigl[L_{t+s}W_t\cdot(Y_s\circ\theta_t)\bigr] \;=\; \mathbb{E}_x\bigl[L_t\cdot(L_s\circ\theta_t)W_t\cdot(Y_s\circ\theta_t)\bigr] \tag{3.6}$$

for any $\mathscr{F}_t$–measurable W_t and any $\mathscr{F}_s$–measurable Y_s.

Similarly, the Markov property can be written

$$\widetilde{\mathbb{E}}_x\bigl[Y_s\circ\theta_t\bigm|\mathscr{F}_t\bigr] \;=\; \widetilde{\mathbb{E}}_{Z_t}Y_s,\quad t<s,$$

for any $\mathscr{F}_s$–measurable r.v. Y_s, which is the same as

$$\widetilde{\mathbb{E}}_x\bigl[W_t(Y_s\circ\theta_t)\bigr] \;=\; \widetilde{\mathbb{E}}_x\bigl[W_t\widetilde{\mathbb{E}}_{Z_t}Y_s\bigr]$$

for any $\mathscr{F}_t$–measurable r.v. W_t. By definition of $\widetilde{\mathbb{P}}_x$, this in turn means

$$\mathbb{E}_x\bigl[L_{t+s}W_t(Y_s\circ\theta_t)\bigr] \;=\; \mathbb{E}_x\bigl[L_tW_t\mathbb{E}_{Z_t}[L_sY_s]\bigr],$$

or, since $\mathbb{E}_{Z_t}[L_sY_s]=\mathbb{E}\bigl[(Y_s\circ\theta_t)(L_s\circ\theta_t)\bigm|\mathscr{F}_t\bigr]$,

$$\mathbb{E}_x\bigl[L_{t+s}W_t(Y_s\circ\theta_t)\bigr] \;=\; \mathbb{E}_x\bigl[L_tW_t(Y_s\circ\theta_t)(L_s\circ\theta_t)\bigr], \tag{3.7}$$

which is the same as (3.6). □

For a random walk, a Markovian change of measure as in Theorem 3.3 does not necessarily lead to a random walk. The necessary and sufficient condition is given in Problem 3.2, but in the rest of the chapter we consider only the exponential change of measure introduced in Section 1, corresponding to $Z_n = S_n$, $L_n = \mathrm{e}^{\theta S_n - n\kappa(\theta)}$ and $\widetilde{\mathbb{P}} = \mathbb{P}_\theta$. The analogue for a Lévy process is as follows:

Theorem 3.4 *Let $\{S_t\}$ be a Lévy process with characteristic triplet (μ,σ^2,ν) in IX.(1.4), i.e.*

$$\kappa(\alpha) \;=\; \alpha\mu+\alpha^2\sigma^2/2+\int_{-\infty}^{\infty}\bigl[\mathrm{e}^{\alpha x}-1-\alpha I(|x|\le 1)\bigr]\nu(\mathrm{d}x) \tag{3.8}$$

and $L_t = \mathrm{e}^{\theta S_t - t\kappa(\theta)}$. Then $\{L_t\}$ satisfies the assumptions of Theorem 3.3, and $\mathbb{P}_\theta = \widetilde{\mathbb{P}}$ corresponds a new Lévy process, such that the changed parameters are

$$\widetilde{\mu}=\mu+\theta\sigma^2+\int_{-1}^{1}[\mathrm{e}^{\theta x}-1]\nu(\mathrm{d}x),\quad \widetilde{\sigma}^2=\sigma^2,\quad \widetilde{\nu}(\mathrm{d}x)=\mathrm{e}^{\theta x}\nu(\mathrm{d}x). \tag{3.9}$$

In the finite variation case, the changed parameters in the representation

$$\kappa(\alpha) \;=\; \alpha\mu+\alpha^2\sigma^2/2+\int_{-\infty}^{\infty}[\mathrm{e}^{\alpha x}-1]\nu(\mathrm{d}x) \tag{3.10}$$

are the same expressions as in (3.9), only with the integral deleted.

Proof. The first part of the theorem is easy. For (3.9), define $\mathrm{e}^{\widetilde{\kappa}(\alpha)} = \widetilde{\mathbb{E}}_0 \mathrm{e}^{\alpha S_1}$. Then

$$\mathrm{e}^{\widetilde{\kappa}(\alpha)} \;=\; \mathbb{E}_0\bigl[L_1 \mathrm{e}^{\alpha X_1}\bigr] \;=\; \mathrm{e}^{-\kappa(\theta)}\mathbb{E}_0\bigl[\mathrm{e}^{(\alpha+\theta)X_1}\bigr] \;=\; \mathrm{e}^{\kappa(\alpha+\theta)-\kappa(\theta)}.$$

Thus $\widetilde{\kappa}(\alpha) = \kappa(\alpha+\theta) - \kappa(\theta)$ which after inserting (3.8) by straightforward algebra reduces to

$$\begin{aligned} & \alpha(\mu+\theta\sigma^2) + \alpha^2\sigma^2/2 + \int_{-\infty}^{\infty} \bigl[(\mathrm{e}^{\alpha x}-1)\mathrm{e}^{\theta x} - \alpha I(|x|\le 1)\bigr]\nu(\mathrm{d}x), \\ = \;& \alpha\widetilde{\mu} + \alpha^2\widetilde{\sigma}^2/2 + \int_{-\infty}^{\infty} \bigl[\mathrm{e}^{\alpha x} - 1 - \alpha I(|x|\le 1)\bigr]\widetilde{\nu}(\mathrm{d}x)\,. \end{aligned}$$

The finite variation case is similar. □

Problems

3.1 Construct two transition matrices $\boldsymbol{P}, \widetilde{\boldsymbol{P}}$ for a Markov chain, such that the Markov probabilities $\mathbb{P}, \widetilde{\mathbb{P}}$ corresponding to some initial state are equivalent on the whole of $E^{\mathbb{N}}$.

3.2 Assume in Theorem 3.3 that $\{X_n\}$ is a random walk, $X_n = Y_1 + \cdots + Y_n$. Show that $\widetilde{\mathbb{P}}$ corresponds to a new random walk if and only if $L_n = h(Y_1)\cdots h(Y_n)$ for some function h with $\mathbb{E}h(Y) = 1$, and that then the changed increment distribution is $\widetilde{F}(x) = \mathbb{E}[h(Y); Y \le x]$.

3.3 Let $\{Z_t\}$ be a diffusion with generator $\mathscr{A}h$, and for some h in the domain, define

$$L_t \;=\; \frac{h(Z_t)}{h(Z_0)} \exp\left\{ -\int_0^t \frac{\mathscr{A}h(Z_s)}{h(Z_s)}\,\mathrm{d}s \right\}.$$

Show, using Ito's formula, that $\{L_t\}$ is a local martingale, and identify $\widetilde{\mathbb{P}}$ when $\{L_t\}$ is a proper martingale.

3.4 In the Markov case, show that a multiplicative functional $\{L_t\}$ with $\mathbb{E}_x L_t = 1$ for all x, t is a martingale.

3.5 Let $X_t = \sum_1^{N_t} U_i$ be a compound Poisson process with Poisson rate β and distribution B of the U_i. Show that the likelihood ratio $L_t = \exp\{\theta X_t - t\kappa(\theta)\}$ provides a new compound Poisson process with parameters $\widetilde{\beta}$, $\widetilde{B}$ where $\widetilde{\beta} = \beta\mathbb{E}\mathrm{e}^{\theta U}$, $\widetilde{B}(\mathrm{d}x) = \mathrm{e}^{\theta x}B(\mathrm{d}x)/\mathbb{E}\mathrm{e}^{\theta U}$.

3.6 Let $X_t = \sum_1^{N_t} U_i$ be a compound Poisson process, such that the Poisson rate is β and the distribution B of the U_i is B w.r.t. $\mathbb{P}$, and $\widetilde{\beta}$, $\widetilde{B}$ w.r.t. $\widetilde{\mathbb{P}}$. Compute L_t.

Notes The results of the present section are standard; see e.g. Jacod and Shiryaev (1987), Küchler and Sørensen (1997) and Revuz and Yor (1999). The theory is sometimes summarized under the term *Girsanov's theorem*, though more commonly this refers to the special case of diffusions.

Generalizations of Problem 3.3 are in Palmowski and Rolski (2002), who also give much survey material.

4 First Applications

For a general martingale $\{M_t\}$, the criteria for optional stopping at τ (i.e. $\mathbb{E}M_\tau = \mathbb{E}M_0$) usually involve uniform integrability of the $M_{\tau\wedge t}$. For likelihood ratio martingales, a different sort of criterion is available:

Corollary 4.1 *Let $\{L_t\}$, $\widetilde{\mathbb{P}}$ be as in Proposition 3.1 and let τ be a stopping time with $\mathbb{P}(\tau < \infty) = 1$. Then a necessary and sufficient condition that $\mathbb{E}L_\tau = 1$ is that $\widetilde{\mathbb{P}}(\tau < \infty) = 1$.*

Proof. Taking $Z = L_\tau$ in Theorem 3.2 gives $\mathbb{E}L_\tau = \widetilde{\mathbb{P}}(\tau < \infty)$. □

Example 4.2 Consider a random walk or Lévy process and the Wald martingale $\bigl\{\mathrm{e}^{\theta S_t - t\kappa(\theta)}\bigr\}$. Corollary 4.1 then states that

$$\mathbb{E}\mathrm{e}^{\theta S_\tau - \tau\kappa(\theta)} = 1 \tag{4.1}$$

if and only if $\mathbb{P}_\theta(\tau < \infty) = 1$. Here are some main examples:

(a) $\tau = \inf\bigl\{t > 0 : S_t \notin [a, b]\bigr\}$ with $a < 0 < b$. Here $\tau < \infty$ for any random walk or Lévy process (except for the trivial case $S_t \equiv 0$), so that $\mathbb{P}_\theta(\tau < \infty) = 1$ for all θ and (4.1) *holds always.*

(b) $\tau = \inf\{t > 0 : S_t > u\}$ with $u > 0$. Here $\mathbb{P}(\tau < \infty) = 1$ if and only if the drift $\kappa'(0)$ is nonnegative. Thus *the necessary and sufficient condition that* $\mathbb{P}_\theta(\tau < \infty) = 1$ *and* (4.1) *holds is that* $\kappa'(\theta) \geq 0$. □

We next consider the case where $\{S_t\}_{t\geq 0}$ is Brownian motion with drift ξ and unit variance constant, and shall derive certain standard formulas involving the distributions of $M_T = \sup_{0\leq t\leq T} S_t$ and the first passage time $\tau(\xi, c) = \inf\{t > 0 : S_t \geq c\}$, $c > 0$. The distribution $G(\,\cdot\,; \xi, c)$ of $\tau(\xi, c)$ is known as the *inverse Gaussian distribution*; it has already been met in III.7, X.7 and is of importance later in this chapter as well. Note that $G(T; \xi, c) = \mathbb{P}(M_T \geq c)$.

Theorem 4.3 *$G(\,\cdot\,;\xi, c)$ is proper for $\xi \geq 0$ and defective for $\xi < 0$. Further $G(T;\xi, c) = 0$, $T < 0$, whereas for $T > 0$*

$$G(T;\xi,c) \;=\; 1 - \Phi\Bigl(\frac{c}{\sqrt{T}} - \xi\sqrt{T}\Bigr) + \mathrm{e}^{2\xi c}\Phi\Bigl(-\frac{c}{\sqrt{T}} - \xi\sqrt{T}\Bigr)\,. \tag{4.2}$$

In particular, $\|G(\cdot;\xi,c)\| = \mathrm{e}^{2\xi c}$ in the defective case $\xi < 0$. The density $g(T;\xi,c)$ and the c.g.f. $\lambda(\alpha;\xi,c)$ are

$$g(T;\xi,c) \;=\; \frac{c}{\sqrt{2\pi}T^{3/2}}\exp\left\{\xi c - \frac{1}{2}\left(\frac{c^2}{T} + \xi^2 T\right)\right\}, \quad T > 0, \tag{4.3}$$

$$\lambda(\alpha;\xi,c) \;=\; \log\int_0^\infty \mathrm{e}^{\alpha t} g(t;\xi,c)\,\mathrm{d}t \;=\; \xi c - c\sqrt{\xi^2 - 2\alpha}, \quad \alpha \leq \frac{\xi^2}{2}, \tag{4.4}$$

and for $\xi \geq 0$ the mean and variance are

$$\mathbb{E}\tau(\xi,c) \;=\; \frac{c}{\xi}, \quad \mathbb{V}ar\,\tau(\xi,c) \;=\; \frac{c}{\xi^3}. \tag{4.5}$$

Proof. We use exponential change of measure. We have $\kappa(\alpha) = \alpha\xi + \alpha^2/2$, and $\mathbb{P}_\theta$ makes $\{S_t\}$ Brownian motion with drift $\xi + \theta$ and unit variance; cf. Theorem 3.4. In particular, $\mathbb{P}_{-\xi}$ corresponds to zero drift.

In III.7.6, it was shown by the reflection principle that in the case of zero drift,

$$\mathbb{P}\big(\tau(0,c) \leq T\big) \;=\; \mathbb{P}_{-\xi}(M_T \geq c) \;=\; 2\Phi\Big(\frac{-c}{\sqrt{T}}\Big).$$

By straightforward differentiation, this gives the density, $\mathbb{P}(\tau(0,c) \in \mathrm{d}T) = c\varphi(-c/\sqrt{T})/T^{3/2}\,\mathrm{d}T$, and hence (using $\kappa(-\xi) = -\xi^2/2$ and $S_{\tau(0,c)} = c$)

$$\begin{aligned}\mathbb{P}\big(\tau(\xi,c) \in \mathrm{d}T\big) \;&=\; \mathbb{E}_{-\xi}\Big[\mathrm{e}^{\xi S_{\tau(0,c)}+\tau(0,c)\kappa(-\xi)};\, \tau(0,c) \in \mathrm{d}T\Big] \\ =\; e^{\xi c - T\xi^2/2}\mathbb{P}\big(\tau(0,c) \in \mathrm{d}T\big) \;&=\; \frac{c}{\sqrt{2\pi}}T^{-3/2}\mathrm{e}^{\xi c - T\xi^2/2}\mathrm{e}^{-c^2/2T}\,\mathrm{d}T\,,\end{aligned}$$

showing (4.3); (4.2) then follows by checking that the derivative of the r.h.s. is (4.3) and that the value at $T = 0$ is 0. The expression for $\|G(\,\cdot\,;\xi,c)\|$ follows by letting $T \to \infty$ in (4.2). For (4.4), note that $\kappa(\theta) = -\alpha$ has two solutions $\theta_\pm = \pm\sqrt{\xi^2 - 2\alpha} - \xi$ when $\alpha \leq \xi^2/2$ and that θ_+ satisfies $\kappa'(\theta_+) = \xi + \theta_+ \geq 0$. Thus by Example 4.2(b), we have for $\xi \geq 0$ that

$$1 \;=\; \mathbb{E}\mathrm{e}^{\theta_+ S_{\tau(\xi,c)} - \tau(\xi,c)\kappa(\theta_+)} \;=\; \mathbb{E}\mathrm{e}^{\theta_+ c + \alpha\tau(\xi,c)},$$

and (4.4) follows. We then get (4.5) either by differentiation of (4.4) or by optional stopping of $\{S_t - \xi t\}$, $\big\{(S_t - \xi t)^2 - t\big\}$.

It remains to show that (4.4) also holds when $\xi < 0$. But then by (4.3),

$$\begin{aligned}\int_0^\infty \mathrm{e}^{\alpha t} g(t;\xi,c)\,\mathrm{d}t \;&=\; \mathrm{e}^{2\xi c}\int_0^\infty e^{\alpha t} g(t;-\xi,c)\,\mathrm{d}t \\ =\; \exp\Big\{2\xi c + \big(-\xi c - \sqrt{\xi^2 - 2\alpha}\big)\Big\} \;&=\; \exp\Big\{\xi c - \sqrt{\xi^2 - 2\alpha}\Big\}.\end{aligned}$$

□

We formulate the last step of the proof as:

Corollary 4.4 *For $\xi < 0$, $\mathbb{P}(\tau(\xi,c) < \infty) = \mathrm{e}^{2\xi c}$, and the distribution of $\tau(\xi,c)$ given $\tau(\xi,c) < \infty$ is the same as the distribution of $\tau(-\xi,c)$.*

Some further noteworthy properties of the inverse Gaussian distribution:

Proposition 4.5 (i) *$G(T;\,\xi,c) \;=\; G(T/u^2;\,\xi u, c/u)$;* (ii) *$G(\cdot;\,\xi, c_1{+}c_2) = G(\cdot;\,\xi,c_1){*}G(\cdot;\,\xi,c_2)$;* (iii) *if $\xi > 0$, then $G(\cdot;\,\xi,c)$ is asymptotically normal as $c \to \infty$ with mean c/ξ and variance c/ξ^2.*

Proof. (i) is clear from (4.2). For (ii), note just that $\tau(\xi,c_1{+}c_2) \overset{\mathscr{D}}{=} \tau(\xi,c_1){+}V$ where V is the time of first passage from level c_1 to level $c_1 + c_2$. But by the strong Markov property, V is independent of $\tau(\xi, c_1)$ and distributed as $\tau(\xi, c_2)$. Finally, (iii) is an immediate consequence of (ii) and (4.5). □

Example 4.6 Some main applications of the techniques studied in the present chapter occur in sequential analysis. As a digression, we shall present some of the ideas in that setting, stressing the probability calculations rather that the statistical aspects (which are practical as well as touching upon questions in the foundations of statistical inference).

Suppose we are given an exponential family $(F_\theta)_{\theta\in\Theta}$ of distributions on $\mathbb{R}$ with $\mathbb{E}_0 X = \kappa_0'(0) = 0$ and want to test the hypothesis $H_0 : \theta \geq 0$ versus the alternative $H_1 : \theta < 0$. The traditional likelihood ratio test based upon a fixed sample size N then rejects for small values of S_N, say $S_N < a_N$. In the sequential setting, one instead prescribes two constants $a, b > 0$ and proceeds by sampling $X_1, X_2, \ldots$ one after another. If $S_N < -a$ at time N, H_0 is rejected, and similarly H_0 is accepted if $S_N > b$, whereas if $-a \leq S_N \leq b$ one continues by taking the extra observation X_{N+1}. That is, the sampling stops at time $\tau = \inf\bigl\{N \geq 1 : S_N \notin [-a, b]\bigr\}$ by rejecting H_0 if $S_\tau < -a$ and accepting H_0 if $S_\tau > b$.

The probability problem is to compute the so–called *operation characteristics* (two–sided ruin probability) $k(\theta) = \mathbb{P}_\theta(\text{accept } H_0) = \mathbb{P}_\theta(S_\tau > b)$, one purpose being to tell how a, b should be chosen to meet certain prescribed requirements on $k(\theta)$. Also, one wants to say something about the sample size τ, in particular to compute $\mathbb{E}_\theta\tau$. The idea is to observe that if, as will typically be the case, both a and b are large compared to the typical sizes of the X_i, then we may neglect the excess (overshoot) of S_τ over the boundaries $-a, b$, i.e. use the approximations $S_\tau \approx -a$ on $\{S_\tau < -a\}$, $S_\tau \approx b$ on $\{S_\tau > b\}$. Now the assumption $\kappa'(0) = 0$ ensures that to each $\theta \neq 0$ we can find θ_L of opposite sign satisfying $\kappa_0(\theta_L) = \kappa_0(\theta)$, i.e. $\kappa_\theta(\gamma_\theta) = 0$ where $\gamma_\theta = \theta_L - \theta$. Then by Example 4.2(a), $1 = \mathbb{E}_\theta \mathrm{e}^{\gamma_\theta S_\tau}$ which neglecting the excess over the boundary yields

$$
\begin{aligned}
1 &\approx \mathrm{e}^{-\gamma_\theta a}\mathbb{P}_\theta(S_\tau < -a) + \mathrm{e}^{\gamma_\theta b}\mathbb{P}_\theta(S_\tau > b) \\
&= k(\theta)[\mathrm{e}^{\gamma_\theta b} - \mathrm{e}^{-\gamma_\theta a}] + \mathrm{e}^{-\gamma_\theta a}, \\
k(\theta) &\approx \frac{1 - \mathrm{e}^{-\gamma_\theta a}}{\mathrm{e}^{\gamma_\theta b} - \mathrm{e}^{-\gamma_\theta a}}. \qquad (4.6)
\end{aligned}
$$

Similarly, Wald's identity $\kappa_\theta'(0)\mathbb{E}_\theta\tau = \mathbb{E}_\theta S_\tau$ yields $\kappa_\theta'(0)\mathbb{E}_\theta\tau \approx -a(1 - k(\theta)) + bk(\theta)$, from which an approximation for $\mathbb{E}_\theta\tau$ follows by inserting (4.6). □

Problems

4.1 Show that $\tau(0, c)$ is α–stable with $\alpha = 1/2$ [i.e., the distribution of n independent copies is the same as the distribution of $n^{1/\alpha}\tau(0, c)$].

4.2 Show that $\{\tau(\xi, c)\}_{c\geq 0}$ is a subordinator for $\xi \geq 0$, and that the Lévy measure has density $(2\pi)^{-1/2}x^{-3/2}\mathrm{e}^{-\xi^2 x/2}$, $x > 0$.

4.3 Consider the case $\theta = 0$ in Example 4.6. Show that $k(0) \approx a/(a+b)$, $\mathbb{E}_0\tau \approx ab/\kappa_0''(0)$.

Notes For the derivation of the inverse Gaussian distribution, see e.g. Harrison (1985) and Siegmund (1985); further properties of the distribution and related Brownian functionals are in Borodin and Salminen (1996). A main reference for sequential analysis is Siegmund (1985).

5 Cramér–Lundberg Theory

Let $\{S_t\}_{t\in\mathbb{T}}$ be a random walk or Lévy process with negative drift, $\kappa'(0) < 0$. The *Lundberg equation* is $\kappa(\gamma) = 0$ and we assume the existence of a solution $\gamma > 0$. We will exploit the exponential change of measure corresponding to $\theta = \gamma$ and write $\mathbb{P}_L$ instead of $\mathbb{P}_\gamma$; one notes right away that the fundamental likelihood ratio identity, Theorem 3.2, takes a particularly simple form in this case,

$$\mathbb{P}(G) \;=\; \mathbb{E}_L[\mathrm{e}^{-\gamma S_\tau};\, G] \tag{5.1}$$

when τ is a stopping time and $G \in \mathscr{F}_\tau$, $G \subseteq \{\tau < \infty\}$. As the main example of the use of this formula, we shall look at the distribution of $M = \sup_{t\in\mathbb{T}} S_t$. To this end, take $\tau = \tau(u) = \inf\{t > 0 : S_t > u\}$ and let $B(u) = S_\tau - u$ be the overshoot.

Theorem 5.1 $\mathbb{P}(M > u) \le \mathrm{e}^{-\gamma u}$ *for all* $u \ge 0$.

Proof. Let $G = \{\tau < \infty\}$. Then $\mathbb{P}_L(G) = 1$ because $\mathbb{E}_L S_1 = \kappa'(\gamma) > 0$ by convexity, and therefore $B(u) \ge 0$ yields

$$\mathbb{P}(M > u) \;=\; \mathbb{P}(G) \;=\; \mathbb{E}_L\mathrm{e}^{-\gamma S_\tau} \;=\; \mathrm{e}^{-\gamma u}\mathbb{E}_L\mathrm{e}^{-\gamma B(u)} \;\le\; \mathrm{e}^{-\gamma u}. \tag{5.2}$$

□

Theorem 5.1 is known as *Lundberg's inequality* in insurance risk (where $\mathbb{P}(M > u)$ is interpreted as the ruin probability) and has been reproved and refined in queueing theory (where $W \overset{\mathscr{D}}{=} M$) by various authors. The argument in (5.2) is apparently just to neglect the excess over the boundary. A small refinement produces the celebrated *Cramér–Lundberg approximation*:

Theorem 5.2 *If $B(u)$ converges in $\mathbb{P}_L$–distribution as $u \to \infty$, say to $B(\infty)$, then $\mathbb{P}(M > u) \sim C\mathrm{e}^{-\gamma u}$ where $C = \mathbb{E}_L\mathrm{e}^{-\gamma B(\infty)}$.*

Proof. Since $\mathrm{e}^{-\gamma x}$ is bounded and continuous, we just have to note that in (5.2) $\mathbb{E}_L\mathrm{e}^{-\gamma B(u)} \to C$ by general results on weak convergence. □

Theorem 5.3 *For a discrete–time nonlattice random walk, $B(\infty)$ exists w.r.t. $\mathbb{P}_L$ if $\kappa'(\gamma) < \infty$. In that case, C is given in terms of the ladder* *height distributions G_+, G_- by*

$$C \;=\; \frac{1-\|G_+\|}{\gamma\int_0^\infty x\mathrm{e}^{\gamma x}\,G_+(\mathrm{d}x)} \;=\; \frac{\mathbb{E}X}{\gamma\kappa'(\gamma)\mathbb{E}S_{\tau_-}}\left\{1-\int_{-\infty}^0 \mathrm{e}^{\gamma x}\,G_-(\mathrm{d}x)\right\}. \quad (5.3)$$

Proof. Since $\kappa'(\gamma) = \mathbb{E}_L X$, the existence of $B(\infty)$ was noted in VIII.2.1 together with the expression $\bigl(1-G_+^{(L)}(x)\bigr)/\mu_+^{(L)}$ for the $\mathbb{P}_L$–density where $G_+^{(L)}$ is the ascending ladder height distribution w.r.t. $\mathbb{P}_L$ and $\mu_+^{(L)}$ its mean. Now if we put $\tau = \tau_+$, $G = \bigl\{S_{\tau_+}\in A, \tau_+ < \infty\bigr\}$ in (5.1), we get

$$G_+(A) \;=\; \mathbb{P}(G) \;=\; \mathbb{E}_L\bigl[\mathrm{e}^{-\gamma S_{\tau_+}}; G\bigr] \;=\; \int_A \mathrm{e}^{-\gamma x}\,G_+^{(L)}(\mathrm{d}x) \qquad (5.4)$$

which shows that $G_+(\mathrm{d}x) = \mathrm{e}^{-\gamma x}G_+^{(L)}(\mathrm{d}x)$. Hence

$$\begin{aligned} C \;&=\; \mathbb{E}_L\mathrm{e}^{-\gamma B(\infty)} \;=\; \int_0^\infty \mathrm{e}^{-\gamma x}\bigl(1-G_+^{(L)}(x)\bigr)/\mu_+^{(L)}\,\mathrm{d}x \\ &=\; \frac{1}{\mu_+^{(L)}}\int_0^\infty \frac{1}{\gamma}(1-\mathrm{e}^{-\gamma y})\,G_+^{(L)}(\mathrm{d}y) \;=\; \frac{1}{\gamma\mu_+^{(L)}}(1-\|G_+\|) \end{aligned}$$

and since $\mu_+^{(L)} = \int_0^\infty x\mathrm{e}^{\gamma x}G_+(\mathrm{d}x)$, the first identity in (5.3) follows. For the second, note first that as in (5.4), $G_-(\mathrm{d}x) = \mathrm{e}^{-\gamma x}G_-^{(L)}(\mathrm{d}x)$, and thus $\{\ldots\}$ in (5.3) is just $1-\|G_-^{(L)}\|$. Now just note that (cf. VIII.2.3(c))

$$1-\|G_+\| \;=\; \frac{1}{\mathbb{E}\tau_-} \;=\; \frac{\mathbb{E}X}{\mathbb{E}S_{\tau_-}}, \quad \frac{1}{\mu_+^{(L)}} \;=\; \frac{1}{\mathbb{E}_L X\mathbb{E}_L\tau_+^{(L)}} \;=\; \frac{1-\|G_-^{(L)}\|}{\kappa'(\gamma)}.$$

□

Remark 5.4 In the lattice case of Theorem 5.3, we still have $\mathbb{P}(M>u)\sim C\mathrm{e}^{-\gamma u}$ provided $u\to\infty$ through values of the lattice span only (however, C comes out slightly different). The same applies to a compound Poisson process with lattice jumps. For all other types of Lévy processes (including, for example, a compound Poisson process with lattice jumps and an added drift or Brownian component) $B(\infty)$ is easily seen to exist assuming only $\kappa'(\gamma)<\infty$. See further Bertoin and Doney (1994a). □

In view of $W \overset{\mathscr{D}}{=} M$, the Cramér–Lundberg approximation states that under appropriate conditions, the tail of the waiting time W in a $GI/G/1$ queue is asymptotically exponential. This result is superficially similar to the heavy traffic approximation in X.7, but of course the range of parameters for which the two results give an exponential approximation for $\mathbb{P}(W>u)$ is not the same (neither are the constants equal).

Example 5.5 For $GI/M/1$, we have in the notation of VIII.5.8 that $\mathbb{P}(W>u) = \theta\mathrm{e}^{-\eta u}$. It follows that the Cramér–Lundberg approximation is exact in this case and that $\theta = C$, $\eta = \gamma$ (it is straightforward to check directly from the expression for G_+ in VIII.5.8(a) that indeed (5.3) reduces to θ). For $M/G/1$, $-S_{\tau_-}$ is exponential with intensity β whereas

$$\kappa(\alpha) = \log \widehat{B}[\alpha] + \log \frac{\beta}{\beta+\alpha}, \quad \kappa'(\gamma) \;=\; \frac{\widehat{B}'[\gamma]}{\widehat{B}[\gamma]} - \frac{1}{\beta+\gamma} \;=\; \frac{\beta\widehat{B}'[\gamma]-1}{\beta+\gamma},$$

and (5.3) becomes

$$\begin{aligned} C &= \frac{\mu_B - 1/\beta}{\gamma\kappa'(\gamma)(-1/\beta)}\left\{1-\beta\int_{-\infty}^{0} \mathrm{e}^{(\gamma+\beta)x}\,\mathrm{d}x\right\} \\ &= \frac{1-\rho}{\gamma\kappa'(\gamma)}\left\{1-\frac{\beta}{\gamma+\beta}\right\} \;=\; \frac{1-\rho}{\kappa'(\gamma)(\gamma+\beta)} \;=\; \frac{1-\rho}{\beta\widehat{B}'[\gamma]-1}\,. \end{aligned}$$

□

Just as the heavy traffic approximation has a time–dependent version X.7.4 (in terms of the inverse Gaussian distribution), so is the case for the Cramér–Lundberg approximation. This time the correction factor is normal:

Theorem 5.6 *Suppose in addition to the conditions of Theorem 5.2 that $\sigma_L^2 = \kappa''(\gamma) = \mathbb{V}ar_L S_1 < \infty$, and define $\mu_L = \mathbb{E}_L S_1 = \kappa'(\gamma)$, $\omega^2 = \sigma_L^2/\mu_L^3$, $x(u) = (T - u/\mu_L)(\omega u^{1/2})$. Then for $T \in \mathbb{T}$,*

$$\mathbb{P}(M_T > u) \;\sim\; C\mathrm{e}^{-\gamma u}\Phi\big(x(u)\big), \quad u \to \infty, \tag{5.5}$$

in the sense that if T varies with u in such a way that $x = \lim x(u)$ exists, then $\mathbb{P}(M_T > u) = C\mathrm{e}^{-\gamma u}\Phi(x) + \mathrm{o}(\mathrm{e}^{-\gamma u})$.

The proof rests on two lemmas:

Lemma 5.7 *As $u \to \infty$, it holds that:* (i) *$\tau(u)/u \xrightarrow{\mathbb{P}_L} \mu_L^{-1}$;* (ii) *$\mathbb{E}_L\tau(u)/u \to \mu_L^{-1}$;* (iii) *$\tau(u)$ is $\mathbb{P}_L$–asymptotically normal with mean u/μ_L and variance $\omega^2 u$.*

Proof. First note that the standard LLN and CLT assert that

$$\overline{S}_t = \frac{S_t}{t} \overset{\text{a.s.}}{\to} \mu_L, \quad Z_t \;=\; \frac{S_t - t\mu_L}{\sqrt{t}} \overset{\mathscr{D}}{\to} N(0,\sigma_L^2)$$

w.r.t. $\mathbb{P}_L$. Let $t = \tau$ and write as above $S_\tau = u + B(u)$. Since $\tau \to \infty$ and $B(u) \overset{\mathscr{D}}{\to} B(\infty)$ w.r.t. $\mathbb{P}_L$, we have $B(u)/\tau \overset{\mathbb{P}_L}{\to} 0$ and hence $\overline{S}_\tau \to \mu_L$ implies $u/\tau \overset{\mathbb{P}_L}{\to} \mu_L$ and (i). The proof of (ii) is the same as for the elementary renewal theorem in V.1 (or as in Problem VIII.2.3). By (i) and Anscombe's theorem, $Z_\tau \overset{\mathscr{D}}{\to} N(0,\sigma_L^2)$, and (iii) follows since $B(u)/\tau^{1/2} \overset{\mathbb{P}_L}{\to} 0$ implies

$$Z_\tau \overset{\mathcal{D}}{\sim} \frac{u - \tau\mu_L}{\tau^{1/2}} \overset{\mathcal{D}}{\sim} \mu_L^{3/2}\frac{\tau - u/\mu_L}{u^{1/2}}\,.$$

□

Lemma 5.8 *$B(u)$ and τ are asymptotically independent as $u \to \infty$. That is, for f, g bounded and continuous*

$$\mathbb{E}_L\bigl[f(B(u))\,g\bigl((\tau - u/\mu_L)/\omega u^{1/2}\bigr)\bigr] \;\to\; \mathbb{E}_L f(B(\infty)) \cdot \mathbb{E}g(U), \qquad (5.6)$$

where U is standard normal.

Proof. Define $u' = u - u^{1/4}$. Then the distribution of $\tau(u) - \tau(u')$ given $\mathscr{F}_{\tau(u')}$ is degenerate at 0 if $S_{\tau(u')} > u$ and otherwise that of $\tau(v)$ with $v = u - S_{\tau(u')} = u^{1/4} - B(u')$. Hence

$$\begin{aligned} \mathbb{E}_L[\tau(u) - \tau(u')] &= \mathbb{E}_L\bigl[\tau(u^{1/4} - B(u'));\, B(u') \le u^{1/4}\bigr] \\ &\le \mathbb{E}\tau(u^{1/4}) \;=\; \mathrm{O}(u^{1/4}), \end{aligned}$$

and thus in (5.6), we can replace $\tau(u)$ by $\tau(u')$. Defining $h(u) = \mathbb{E}_L f(B(u))$, so that $h(u) \to h(\infty) = \mathbb{E}_L f(B(\infty))$, it follows similarly that

$$\begin{aligned} &\mathbb{E}_L\bigl[f(B(u))\,\big|\,\mathscr{F}_{\tau(u')}\bigr] \\ &= h\bigl(u^{1/4} - B(u')\bigr)I\bigl(B(u') \le u^{1/4}\bigr) \;+\; f\bigl(B(u') - u^{1/4}\bigr)I\bigl(B(u') > u^{1/4}\bigr) \\ &\overset{\mathbb{P}_L}{\to} h(\infty)\cdot 1 + 0, \end{aligned}$$

using that $u^{1/4} - B(u') \overset{\mathbb{P}_L}{\to} \infty$ as follows from $B(u') \overset{\mathscr{D}}{\to} B(\infty)$. Hence

$$\begin{aligned} &\mathbb{E}_L\bigl[f(B(u))\,g\bigl((\tau(u') - u/\mu_L)/\omega u^{1/2}\bigr)\bigr] \\ &= \mathbb{E}_L\bigl[\mathbb{E}_L\bigl[f(B(u))\,|\,\mathscr{F}_{\tau(u')}\bigr]\,g\bigl((\tau(u') - u/\mu_L)/\omega u^{1/2}\bigr)\bigr] \\ &\sim h(\infty)\mathbb{E}_L g\bigl((\tau(u') - u/\mu_L)/\omega u^{1/2}\bigr) \;\sim\; h(\infty)\mathbb{E}g(U). \end{aligned}$$

□

Proof of Theorem 5.6. By (5.1), (5.6),

$$\begin{aligned} \mathbb{P}(M_T > u) &= \mathbb{P}(\tau(u) \le T) \;=\; \mathrm{e}^{-\gamma u}\mathbb{E}_L\bigl[\mathrm{e}^{-\gamma B(u)};\, \tau(u) \le T\bigr] \\ &= \mathrm{e}^{-\gamma u}\mathbb{E}_L\bigl[\mathrm{e}^{-\gamma B(u)};\, \tau(u) \le u/\mu_L + x\omega u^{1/2} + \mathrm{o}(u^{1/2})\bigr] \\ &= \mathrm{e}^{-\gamma u}\bigl\{\mathbb{E}_L\bigl[\mathrm{e}^{-\gamma B(\infty)}\Phi(x)\bigr] + \mathrm{o}(1)\bigr\}. \end{aligned}$$

□

We finally consider a variant of the Cramér–Lundberg approximation, the asymptotics of the tail of the cycle maximum in the $GI/G/1$ queue:

Corollary 5.9 *Consider a discrete–time random walk and define $M_- = \sup\{S_n : 0 \le n < \tau_-\}$. Then under the conditions of Theorem 5.2, $\mathbb{P}(M_- > u) \sim C_-\mathrm{e}^{-\gamma u}$ where $C_- = C\mathbb{P}_L(\tau_- = \infty) = \mathbb{E}_L\mathrm{e}^{-\gamma B(\infty)} \cdot \mathbb{P}_L(\tau_- = \infty)$.*

Proof. Since $\{\tau_- > \tau(u)\} \in \mathscr{F}_{\tau(u)}$, (5.1) yields

$$\begin{aligned} \mathbb{P}(M_- > u) &= \mathbb{P}(\tau_- > \tau(u)) \;=\; \mathrm{e}^{-\gamma u}\mathbb{E}_L\bigl[\mathrm{e}^{-\gamma B(u)};\, \tau_- > \tau(u)\bigr] \\ &\sim \mathrm{e}^{-\gamma u}\mathbb{E}_L\mathrm{e}^{-\gamma B(\infty)} \cdot \mathbb{P}_L(\tau_- = \infty), \end{aligned}$$

where the last step used $B(u) \to B(\infty)$, $\{\tau_- > \tau(u)\} \uparrow \{\tau_- = \infty\}$ in $\mathbb{P}_L$–distribution and a similar asymptotic independence argument as in Lemma 5.8. □

Proposition VI.4.9 now yields the existence of a Gumbel limit for the extremes of the $GI/G/1$ waiting times (recall that σ is the number of customers served in a busy cycle):

Corollary 5.10 *Consider a $GI/G/1$ queue such that the distribution of $X = U - T$ satisfies the conditions of Theorem 5.2, and let $\overline{W}_N = \max_{n=0,\ldots,N} W_n$ be the largest waiting time among customers $0,\ldots,N$. Then*

$$\mathbb{P}\bigl(\gamma \overline{W}_N \le x - \log N - \log(C_-/\mathbb{E}\sigma) \le x\bigr) \;\to\; \mathrm{e}^{-\mathrm{e}^{-x}}, \quad N \to \infty.$$

Notes General references on ruin probabilities are in Asmussen (2000); see also XIV.5–6. In the queueing setting, Lundberg's inequality was first proved by Kingman. The Lundberg parameter $\theta = \gamma$ occurs in a variety of settings, e.g. Feller (1971), Siegmund (1985), Neuts (1986) and Asmussen (2000). Lemma 5.8 is known in the literature as *Stam's lemma.*

Exponential tail approximations for queues is a large and active area. Exact asymptotics becomes quickly diffficult to derive beyond simple models (see, however, Sadowsky and Szpankowsky, 1995, for $GI/G/s$ and Section 8 for Markov additive models), but large deviations techniques (yielding logarithmic asymptotics) have proved successful in a number of cases. A particularly useful general result in this direction was given by Glynn and Whitt (1994) under conditions similar to those of the Gärtner–Ellis theorem (see the Notes to Section 2); a related reference is Duffield and O'Connell (1995). For networks, see the Notes to IV.6.

Further directions include heavy tails, see X.9 and the Notes there, and Gaussian processes where we mention for example Dębicki *et al.* (1998), Choe and Shroff (1999) and Dębicki (2002); when looking for asymptotics of $\mathbb{P}(\sup X_t > u)$, a useful first approximation is often $\mathbb{P}(X_{t(u)} > u)$ where $t = t(u)$ maximizes $\mathbb{P}(X_t > u)$. For the particular case of fractional Brownian motion, see Massoulie and Simonian (1999), Hüssler and Piterbarg (1999) and Piterbarg (2001).

6 Siegmund's Corrected Heavy Traffic Approximations

We now consider a discrete random walk heavy traffic situation similar to X.7 where the mean $\mu = \mathbb{E}X$ is smaller than but close to zero. It is convenient to define the exponential family in a slightly different way than in Section 1 by letting $\theta = 0$ correspond not to the given increment distribution F but to the mean zero case. We then have $F = F_{\theta_0}$ with $\theta_0 < 0$ and $\kappa_0'(0) = 0$. That is, given F we let $\theta_0 < 0$ be the solution of $\kappa'(-\theta_0) = 0$ and let F_θ be the distribution with c.g.f. $\kappa_\theta(\alpha) = \kappa(\alpha + \theta - \theta_0) - \kappa(\theta - \theta_0)$. We then represent the heavy traffic set–up by the limit $\theta_0 \uparrow 0$ in the exponential family $(F_\theta)_{\theta\in\Theta}$.

It is assumed that Θ contains a neighbourhood of zero and, for the ease of notation, that the scale is chosen such that $\mathbb{V}ar_0 X = \kappa_0''(0) = \kappa''(-\theta_0) = 1$. Then for small θ

$$\begin{aligned} \kappa_0(\theta) &= \frac{\theta^2}{2} + \frac{\theta^3}{3!}\mathbb{E}_0 X^3 + \cdots, && (6.1)\\ \mu_\theta &= \kappa_0'\theta) \;=\; \theta + \mathrm{O}(\theta^2), && (6.2)\\ \mathbb{V}ar_\theta X &= \kappa_0''(\theta) \;=\; 1 + \mathrm{O}(\theta), \quad \mathbb{E}_\theta X^2 \;=\; 1 + \mathrm{O}(\theta). && (6.3) \end{aligned}$$

Also, $\theta_L > 0$ connected to $\theta_0 < 0$ by means of $\kappa_0(\theta_0) = \kappa_0(\theta_L)$ is well defined, and by (6.1) we have in the limit $\theta_0 \uparrow 0$ that

$$\frac{\theta_L}{-\theta_0} \to 1, \quad \frac{\gamma}{-\theta_0} = \frac{\theta_L - \theta_0}{-\theta_0} \to 2. \tag{6.4}$$

We shall let u vary with θ_0 in such a way that any of the equivalent relations

$$u\theta_0 \to -\xi, \quad u\theta_L \to \xi \tag{6.5}$$

hold for some $\xi \ge 0$ and as previously let $\tau(u) = \inf\{n \ge 1 : S_n > u\}$. Some preliminary estimates follow immediately from X.7:

Proposition 6.1 *As* $\theta_0 \uparrow 0$, $\gamma\mathbb{E}_{\theta_0} M \to 1$ *and subject to* (6.5),

$$\begin{aligned} \mathbb{P}_{\theta_0}(M > u) &= \mathbb{P}_{\theta_0}(\tau(u) < \infty) \;\to\; \mathrm{e}^{-2\xi}, && (6.6)\\ \mathbb{P}_{\theta_0}\Bigl(\frac{\tau(u)}{u^2} \le T\Bigr) &\to\; G(T; -\xi, 1). && (6.7) \end{aligned}$$

Proof. The condition X.(7.1) is clear since clearly $F_{\theta_0} \xrightarrow{w} F_0$, and by (6.2) and (6.3) $\mu_{\theta_0} \to \mu_0 = 0$, $\mathbb{E}_{\theta_0} X^2 \to 1 = \mathbb{E}_0 X^2$. Hence, for example, X.7.1 implies $\mathbb{E}_{\theta_0}[-\mu_{\theta_0} M / \mathbb{E}_{\theta_0} X^2] \to 1/2$ which is equivalent to $\gamma\mathbb{E}_{\theta_0} M \to 1$ by (6.2)–(6.4). For (6.7), apply X.7.4 to get

$$\mathbb{P}_{\theta_0}\Bigl(\tau\Bigl(\frac{-y\mathbb{E}_{\theta_0} X^2}{\mu_{\theta_0}}\Bigr) \le \frac{t}{\mu_{\theta_0}^2}\Bigr) \;\to\; G(t; -1, y).$$

Letting $y = -u\theta_0$, $t = \xi^2 T$ and using (6.2), (6.3) and (6.5) this implies

$$\mathbb{P}_{\theta_0}(\tau(u) \le Tu^2) \;\to\; G(\xi^2 T; -1, \xi) \;=\; G(T; -\xi, 1),$$

cf. Proposition 4.5(i). Similar estimates yield (6.6). □

We shall now study improvements of these estimates, obtained in a 1979 paper by D. Siegmund. The idea is to estimate the excess over the boundary more carefully and thereby obtain correction terms of lower magnitude $\mathrm{O}(\gamma)$, $\mathrm{O}(\gamma^2)$, ... (in view of (6.2) or (6.4), we might as well have replaced γ by $-\theta_0$ or $-\mu_{\theta_0}$). Only the form of the first–order correction term will be derived rigorously, but the second term is included in the statement of the results because of their importance as approximations. Considering first (6.6), we have $\gamma u \sim 2\xi$. Hence $\mathbb{P}_{\theta_0}(M > u) \sim \mathrm{e}^{-\gamma u}$, and we have:

Theorem 6.2 *Suppose that the $\mathbb{P}_0$–distribution F_0 of X is spread out, and let $\beta = \mathbb{E}_0 B(\infty) = \mathbb{E}_0 S_{\tau_+}^2 / 2\mathbb{E}_0 S_{\tau_+}$. Then as $\theta_0 \uparrow 0$, $u\theta_0 \to -\xi$*

$$\mathbb{P}_{\theta_0}(M > u) \;=\; \mathrm{e}^{-\gamma(u+\beta)} + \mathrm{o}(\gamma^2)\,. \tag{6.8}$$

The proof is based on the relation $\mathbb{P}_{\theta_0}(M > u) = \mathrm{e}^{-\gamma u} C(u)$ with $C(u) = \mathbb{E}_{\theta_L} \mathrm{e}^{-\gamma B(u)}$, which was used in the proof of the Cramér–Lundberg approximation. However, $C(u)$ must now be estimated in a different manner since θ_0 is no longer fixed. We shall need some lemmas, in particular a variant (Lemma 6.4) of Lemma 5.8:

Lemma 6.3 $\mathbb{E}_0 \mathrm{e}^{\epsilon S_{\tau_+}} < \infty$ *for any $\epsilon > 0$ with $\kappa_0(\epsilon) < \infty$.*

Proof. This can be obtained either by an easy variant of the proof of X.2.1 or by Wiener–Hopf factorization; cf. Problem 6.1. □

Lemma 6.4 *$B(u)$ and $\tau(u)/u^2$ are asymptotically independent w.r.t. $\mathbb{P}_{\theta_L}$ as $\theta_0 u \to -\xi$, with the limiting distribution of $\tau(u)/u^2$ being $G(\cdot;\xi,1)$ and that of $B(u)$ the $\mathbb{P}_0$–distribution of $B(\infty)$. That is, for f, g bounded and continuous*

$$\mathbb{E}_{\theta_L} f\bigl(B(u)\bigr) g\bigl(\tau(u)/u^2\bigr) \;\to\; \mathbb{E}_0 f\bigl(B(\infty)\bigr) \int_0^\infty g(x)\, G(\mathrm{d}x;\xi,1)\,. \tag{6.9}$$

More generally, there is an $\epsilon > 0$ such that (6.9) *holds when f is continuous with $f(x) = \mathrm{O}(\mathrm{e}^{\epsilon x})$.*

Proof. The l.h.s. of (6.9) is

$$\mathbb{E}_0 \bigl[f\bigl(B(u)\bigr) g\bigl(\tau(u)/u^2\bigr) \exp\bigl\{\theta_L\bigl(u + B(u)\bigr) - \tau(u)\kappa_0(\theta_L)\bigr\}\bigr]\,, \tag{6.10}$$

and thus we have to inspect the $\mathbb{P}_0$–limit of $(B(u), \tau(u)/u^2)$. Clearly, $B(u) \xrightarrow{\mathscr{D}} B(\infty)$ and by Proposition 6.1, $\tau(u)/u^2 \xrightarrow{\mathscr{D}} G(\cdot;0,1)$. By a variant of the proof of Lemma 5.8, it is seen that we also have asymptotic independence: letting $u' = u - u^{1/4}$, the only new estimate needed is $\tau(u) - \tau(u')$ $= \mathrm{o}(u^2)$ which follows from the stochastical domination by $\tau(u^{1/4})$ and $\tau(u^{1/4})/u^{1/2} \xrightarrow{\mathscr{D}} G(\cdot;0,1)$. Also by Lemma 6.3 and renewal theory, we have $\mathbb{E}_0 \mathrm{e}^{\delta B(u)} \to \mathbb{E}_0 \mathrm{e}^{\delta B(\infty)} < \infty$, in particular $\mathbb{E}_0 \mathrm{e}^{\delta B(u)} \le c$ for all $u \ge 0$. For $\epsilon < \delta$, we then have that $\epsilon + \theta_L < \delta$ eventually, and using uniform integrability and $\kappa_0(\theta_L) \sim \xi^2/2u^2$, it follows that the limit of (6.10) exists for $f(x) = \mathrm{O}(\mathrm{e}^{\epsilon x})$ and is

$$\mathbb{E}_0 f\bigl(B(\infty)\bigr) \int_0^\infty g(x) \mathrm{e}^{\xi - \xi^2 x/2}\, G(\mathrm{d}x;0,1) \;=\; \mathbb{E}_0 f(B(\infty)) \int_0^\infty g(x)\, G(\mathrm{d}x;\xi,1)$$

(note that $\theta_L B(u) \xrightarrow{\mathbb{P}} 0$). □

To obtain the second–order correction terms, the following two lemmas are needed:

Lemma 6.5 *As $\theta_0 u \to -\xi$, it holds for some $\epsilon > 0$ that $\mathbb{E}_{\theta_L} B(u) = \mathbb{E}_{\theta_L} B(\infty) + \mathrm{O}(\mathrm{e}^{-\epsilon u})$.*

Lemma 6.6 $\mathbb{E}_{\theta_L} S_{\tau_+}^k \;=\; \mathbb{E}_0 S_{\tau_+}^k + k\theta_L/(k+1)\mathbb{E}_0 S_{\tau_+}^{k+1} + \mathrm{o}(\gamma)$, $k = 1, 2, \ldots$

It is seen that Lemma 6.5 is a uniform version of VII.2.10, and the proof may proceed by first showing that the $\mathbb{P}_{\theta_L}$–distributions of $B(0) = S_{\tau_+}$ have a common absolute continuous component, and next to check that the estimates in VII.2 hold uniformly in small θ_L. Also, the proof of Lemma 6.6 is not unreasonably complicated, but we omit the details of both proofs.

Proof of Theorem 6.2. By Taylor expansion

$$C(u) \;=\; \mathbb{E}_{\theta_L} \mathrm{e}^{-\gamma B(u)} \;=\; \mathbb{E}_{\theta_L}\Bigl[1 - \gamma B(u) + \frac{\gamma^2}{2} B(u)^2 + \gamma^3 \mathrm{O}\bigl(B(u)^3 \mathrm{e}^{\gamma B(u)}\bigr)\Bigr].$$

Here the last term is $\mathrm{O}(\gamma^3)$ by Lemma 6.4, while $\mathbb{E}_{\theta_L} B(u) \to \mathbb{E}_0 B(\infty) = \beta$. This is more than sufficient for $C(u) = 1 - \gamma\beta + o(\gamma)$, and thus that the remainder term in (6.8) is $\mathrm{o}(\gamma)$. To see that it is actually $\mathrm{o}(\gamma^2)$, note that by Lemmas 6.5 and 6.6

$$\begin{aligned}
\mathbb{E}_{\theta_L} B(u) \;&=\; \mathbb{E}_{\theta_L} B(\infty) + \mathrm{O}(\mathrm{e}^{-\epsilon u}) \;=\; \frac{\mathbb{E}_{\theta_L} S_{\tau_+}^2}{2\mathbb{E}_{\theta_L} S_{\tau_+}} \;+\; \mathrm{O}(\mathrm{e}^{-\epsilon u}) \\
&=\; \frac{\mathbb{E}_0 S_{\tau_+}^2 \;+\; 2\theta_L \mathbb{E}_0 S_{\tau_+}^3/3}{2\mathbb{E}_0 S_{\tau_+} \;+\; \theta_L \mathbb{E}_0 S_{\tau_+}^2} \;+\; \mathrm{o}(\gamma) \\
&=\; \beta \;+\; \theta_L \biggl(\frac{\mathbb{E}_0 S_{\tau_+}^3}{3\mathbb{E}_0 S_{\tau_+}^2} - \beta^2\biggr) \;+\; \mathrm{o}(\gamma) \\
&=\; \beta \;+\; \frac{\gamma}{2}\biggl(\frac{\mathbb{E}_0 S_{\tau_+}^3}{3\mathbb{E}_0 S_{\tau_+}} - \beta^2\biggr) \;+\; \mathrm{o}(\gamma), \\
\mathbb{E}_{\theta_L} B(u)^2 \;&=\; \mathbb{E}_0 B(\infty)^2 + \mathrm{o}(1) \;=\; \frac{\mathbb{E}_0 S_{\tau_+}^3}{3\mathbb{E}_0 S_{\tau_+}} + \mathrm{o}(1).
\end{aligned}$$

Combining these estimates, the terms involving $\mathbb{E}_0 S_{\tau_+}^3$ cancel and we get

$$C(u) \;=\; 1 - \gamma\beta + \frac{\gamma^2\beta^2}{2} + \mathrm{o}(\gamma^2) \;=\; \mathrm{e}^{-\gamma\beta} + \mathrm{o}(\gamma^2),$$

$$\mathbb{P}(M > u) \;=\; \mathrm{e}^{-\gamma u} C(u) \;=\; \mathrm{e}^{-\gamma(u+\beta)} \mathrm{o}(\gamma^2). \qquad \square$$

There is also a similar refinement of $\mathbb{E}_{\theta_0} M \sim \gamma^{-1}$:

Theorem 6.7 *As* $\theta_0 \uparrow 0$, $\mathbb{E}_{\theta_0} M \;=\; \dfrac{1}{\gamma} - \beta + \dfrac{\gamma}{2}\biggl(\dfrac{\mathbb{E}_0 S_{\tau_+}^3}{3\mathbb{E}_0 S_{\tau_+}} - \beta^2\biggr) + \mathrm{o}(\gamma).$

Proof. Using X.(2.3),

$$\mathbb{E}_{\theta_0} M \;=\; \frac{\mathbb{E}_{\theta_0}[S_{\tau_+};\, \tau_+ < \infty]}{1 - \mathbb{P}_{\theta_0}(\tau_+ < \infty)} \;=\; \frac{\mathbb{E}_{\theta_L}[S_{\tau_+} \mathrm{e}^{-\gamma S_{\tau_+}}]}{1 - \mathbb{E}_{\theta_L} \mathrm{e}^{-\gamma S_{\tau_+}}}$$

$$= \frac{\mathbb{E}_{\theta_L} S_{\tau_+} - \gamma \mathbb{E}_{\theta_L} S^2_{\tau_+} + \gamma^2 \mathbb{E}_{\theta_L} S^3_{\tau_+}/2 + \mathrm{O}(\gamma^3)}{\gamma \mathbb{E}_{\theta_L} S_{\tau_+} - \gamma^2 \mathbb{E}_{\theta_L} S^2_{\tau_+}/2 + \mathrm{O}(\gamma^3)}$$

from which the result follows by Lemma 6.6 after some elementary calculus. Again, Lemma 6.6 is not needed to prove rigorously that $\mathbb{E}_{\theta_0} M = \gamma^{-1} - \beta + \mathrm{o}(1)$. □

We mention also without proof the similar refinement of (6.7), i.e. a time–dependent version of the expansion (6.8) for $\mathbb{P}(M > u) = \mathbb{P}(\tau(u) < \infty)$,

$$\mathbb{P}(\tau(u) \le tu^2) \;\approx\; G\bigl(tu^2 + u\mathbb{E}_0 X^3/3;\, -\gamma/2, u + \beta\bigr) \qquad (6.11)$$

(by Proposition 4.5(i), this is the same as $G(t + \mathbb{E}_0 X^3/3u;\, -\gamma u/2, 1 + \beta/u)$). Numerical studies indicate that the above approximations are superior to all others known, not only for θ_0 close to zero but in fact in a remarkably wide range. A deficit is that the constants such as β can be cumbersome to evaluate. We mention in this connection the formula

$$\beta \;=\; \mathbb{E}_0 X^3/6 \;-\; \frac{1}{\pi}\int_0^\infty t^{-2}\Re\bigl(\log[2(1-\phi(t))/t^2]\bigr)\,\mathrm{d}t \qquad (6.12)$$

where $\phi(t) = \mathbb{E}_0 \mathrm{e}^{\mathrm{i}tX}$ which can be implemented by numerical integration. The proof is based upon Fourier inversion but omitted.

Problems

6.1 Let $F^{(k)}$, $S^{(k)}_{\tau_+}$, etc. correspond to $X^{(k)} = X_n \wedge k$. Show by Wiener–Hopf factorization that

$$\mathbb{E}_0 \mathrm{e}^{\alpha S^{(k)}_{\tau_+}} \;\le\; \frac{\mathrm{e}^{\kappa_0(\alpha)} - \mathbb{E}_0 \mathrm{e}^{\alpha S^{(k)}_{\tau_-}}}{1 - \mathbb{E}_0 \mathrm{e}^{\alpha S^{(k)}_{\tau_-}}}$$

when $\alpha > 0$, $\kappa_0(\alpha) < \infty$, and deduce that $\mathbb{E}_0 \mathrm{e}^{\alpha S_{\tau_+}} < \infty$.

6.2 Check that under the assumptions of Theorem 6.2 the constant C of the Cramér–Lundberg approximation satisfies $C = 1 - \gamma\beta + \mathrm{o}(\gamma)$.

Notes The results are from Siegmund (1979). See also Siegmund (1985) where in particular the approach to time–dependent formulas such as (6.11) is somewhat different.

7 Rare Events Simulation

We now consider some applications of exponential change of measure to simulation. Consider a probability $z = \mathbb{P}(A)$ which is not available analytically. As in VI.2d, the crude Monte Carlo (CMC) method then amounts to simulating i.i.d. replicates $Z_1, \ldots, Z_N$ of the r.v. $Z = I(A)$, estimating z by the empirical mean $\overline{z} = (Z_1 + \cdots + Z_N)/N$ and the variance of Z by the empirical variance $s^2 = (N-1)^{-1}\sum_1^N(Z_i-\overline{z})^2$. According to standard central limit theory, $\sqrt{N}(\overline{z}-z) \xrightarrow{\mathscr{D}} N(0,\sigma_Z^2)$, where $\sigma_Z^2 = \mathbb{V}ar(Z)$. Hence

$$\overline{z} \pm \frac{1.96\,s}{\sqrt{N}} \tag{7.1}$$

is an asymptotic 95% confidence interval, and this is the form in which the result of the simulation experiment is commonly reported.

Specific problems arise when the event A is rare, that is, when z is small, even if the CMC variance $\sigma_Z^2 = z(1-z)$ of course tends to zero as $z \downarrow 0$. However, the issue is not so much that the precision is good as that relative precision is bad:

$$\frac{\sigma_Z}{z} \;=\; \frac{\sqrt{z(1-z)}}{z} \;\sim\; \frac{1}{\sqrt{z}} \;\to\; \infty.$$

In other words, a confidence interval of width 10^{-4} may look small, but if the point estimate $\overline{z}$ is of the order 10^{-5}, it does not help telling whether z is of the magnitude 10^{-4}, 10^{-5} or even much smaller. Another way to illustrate the problem is in terms of the sample size N needed to acquire a given relative precision, say 10%, in terms of the half–width of the confidence interval. This leads to the equation $1.96\sigma_Z/(z\sqrt{N}) = 0.1$, i.e.

$$N \;=\; \frac{100\cdot 1.96^2 z(1-z)}{z^2} \;\sim\; \frac{384}{z}$$

which increases like z^{-1} as $z \downarrow 0$. Thus, if z is small, large sample sizes are required.

To improve the efficiency, a common tool in simulation is *importance sampling*, which means simulating from a probability measure $\widetilde{\mathbb{P}}$ different from the given probability measure $\mathbb{P}$ and having the property that there exists a r.v. L such that

$$z \;=\; \mathbb{P}(A) \;=\; \widetilde{\mathbb{E}}[L;\,A]. \tag{7.2}$$

One then employs the CMC method exactly as above, only taking $Z = LI(A)$ (simulated from $\widetilde{\mathbb{P}}$) rather than $Z = I(A)$ (simulated from $\mathbb{P}$).

We consider here two examples $A = A(n) = \{S_n > ny\}$, resp. $A = A(x) = \{\tau(x) < \infty\}$, which have been studied by exponential change of measure techniques in Section 2, resp. Section 5. From the analysis there, it seems an obvious idea to implement importance sampling with $\widetilde{\mathbb{P}}$ given by $\mathbb{P}_\theta$ (with $\theta = \theta(y)$ the saddlepoint), resp. $\widetilde{\mathbb{P}} = \mathbb{P}_L$. The corresponding estimators are

$$\begin{aligned} Z(n) \;&=\; \mathrm{e}^{-\theta S_n + n\theta} I(S_n > ny), && (7.3)\\ Z(x) \;&=\; \mathrm{e}^{-\gamma S_{\tau(x)}} I(\tau(x) < \infty) \;=\; \mathrm{e}^{-\gamma x}\mathrm{e}^{-\gamma B(x)} && (7.4)\end{aligned}$$

(for the last identity in (7.4), note that $\mathbb{P}_L(\tau(x) < \infty) = 1$).

To see that these two estimators indeed work extremely well, we shall employ two established efficiency criteria in rare events simulation, *bounded* *relative error* and *logarithmic efficiency*. To introduce these, assume that the rare event $A = A(x)$ depends on a parameter x as in (7.4) (in (7.3), n takes the role of x). Write $z(x) = \mathbb{P}(A(x))$, assume that the $A(x)$ are rare in the sense that $z(x) \to 0$, $x \to \infty$, and let $Z(x)$ be a Monte Carlo estimator of $z(x)$. We then say that $\{Z(x)\}$ has *bounded relative error* if $\mathbb{V}ar\big(Z(x)\big)/z(x)^2$ remains bounded as $x \to \infty$ (according to the above discussion, this means that the sample size $N = N_\epsilon(x)$ required to obtain a given fixed relative precision, say ϵ =10%, remains bounded). *Logarithmic efficiency* is defined by the slightly weaker requirement that one can get as close to the power 2 as desired: $\mathbb{V}ar\big(Z(x)\big)$ should go to 0 as least as fast as $z(x)^{2-\epsilon}$, i.e.

$$\limsup_{x\to\infty} \frac{\mathbb{V}ar\big(Z(x)\big)}{z(x)^{2-\epsilon}} < \infty \tag{7.5}$$

for any $\epsilon > 0$. This allows $\mathbb{V}ar\big(Z(x)\big)$ to decrease slightly slower than $z(x)^2$, so that $N_\epsilon(u)$ may go to infinity. However, the mathematical definition puts certain restrictions on this growth rate, and in practice, logarithmic efficiency is almost as good as bounded relative error. The term *logarithmic* comes from the equivalent form

$$\liminf_{x\to\infty} \frac{-\log \mathbb{V}ar\big(Z(x)\big)}{-\log z(x)} \;\geq\; 2 \tag{7.6}$$

of (7.5) which is inspired from large deviations theory.

Theorem 7.1 *The estimator* (7.4) *has bounded relative error as $x \to \infty$.*

Proof. By a minor variant of the proofs of Theorems 5.1, 5.2, we get

$$\mathbb{E}_L Z(x)^2 \;=\; \mathrm{e}^{-2\gamma x}\mathbb{E}_L \mathrm{e}^{-2\gamma B(x)} \;\sim\; \mathrm{e}^{-2\gamma x}\mathbb{E}_L \mathrm{e}^{-2\gamma B(\infty)}$$

so that

$$\overline{\lim_{x\to\infty}} \frac{\mathbb{V}ar\big(Z(x)\big)}{z(x)^2} \;\leq\; \overline{\lim_{x\to\infty}} \frac{\mathrm{e}^{-2\gamma x}\mathbb{E}_L \mathrm{e}^{-2\gamma B(\infty)}}{\big(\mathrm{e}^{-\gamma x}\mathbb{E}_L \mathrm{e}^{-\gamma B(\infty)}\big)^2} \;=\; \frac{\mathbb{E}_L \mathrm{e}^{-2\gamma B(\infty)}}{\big(\mathbb{E}_L \mathrm{e}^{-\gamma B(\infty)}\big)^2} \;<\; \infty.$$

□

Theorem 7.2 *The estimator* (7.3) *is logaritmically efficient as $n \to \infty$.*

Proof. By a minor variant of the proof of (2.2), we get

$$\mathbb{E}_\theta Z(n)^2 \;=\; \mathrm{e}^{-2n\kappa^*(y)}\mathbb{E}_\theta\big[\mathrm{e}^{-2\theta(S_n-ny)};\, S_n > ny\big] \;\leq\; \mathrm{e}^{-2n\kappa^*(y)}$$

so that (recall (2.3), (2.4))

$$\begin{aligned}\underline{\lim_{n\to\infty}} \frac{-\log \mathbb{V}ar\big(Z(n)\big)}{-\log z(n)} \;&\geq\; \underline{\lim_{n\to\infty}} \frac{-\log \mathbb{E}Z(n)^2}{-\log z(n)} \;=\; \underline{\lim_{n\to\infty}} \frac{-\log \mathbb{E}Z(n)^2}{n\kappa^*(x)} \\ &\geq\; \underline{\lim_{n\to\infty}} \frac{2n\kappa^*(y)}{n\kappa^*(y)} \;=\; 2.\end{aligned}$$

□

Notes Even if (given the earlier developments of this chapter) it is not surprising that the estimators (7.3), (7.4) work well and that the proofs of this are short, natural questions are whether indeed the particular changes of measures which were employed are the best possible ones, and why. For the first question, we refer to a short argument given in Asmussen and Rubinstein (1995), based upon the information inequality. For the second, specializing a general optimality criterion in importance sampling gives that the change of measure $\widetilde{\mathbb{P}} = \mathbb{P}(\,\cdot \mid A)$ for estimating z gives zero variance (this is of course trivial since the corresponding estimator $Z = z$ is constant). Now $\mathbb{P}(\,\cdot \mid A)$ is typically not known so simulating from this distribution meets difficulties and, more seriously, even if it could be done the importance sampling estimator $Z = z$ could not be evaluated since the whole point in using simulation is that z is not known. However, what is suggested is that taking $\widetilde{\mathbb{P}}$ close to $\mathbb{P}(\,\cdot \mid A)$ would give a small variance. Indeed, in the setting of Theorem 7.2 a classical result from statistical mechanics known as *Boltzmann's law* (see e.g. Khinchin, 1949, or Martin–Löf, 1979) states that given $S_n > ny$, the r.v.'s $X_1, \ldots, X_n$ are asymptotically i.i.d. with distribution F_θ, and similar results supporting Theorem 7.1 are in Asmussen (1982) (see also Anantharam, 1988). In general, large deviations theory and in particular Mogulskii's theorem will often identify the asymptotically most likely path leading to the rare event A and thereby suggest a change of measure. For discussion of these topics, surveys on rare events simulation and references, see Asmussen and Rubinstein (1995) and Heidelberger (1995).

8 Markov Additive Processes

We consider a (finite) Markov additive process $\{(J_t, S_t)\}_{t\in\mathbb{T}}$ in the notation of XI.2; when $\mathbb{T} = [0,\infty)$, we recall in particular the expression $\mathrm{e}^{t\boldsymbol{K}[\alpha]}$ for the matrix $\widehat{\boldsymbol{F}}_t[\alpha]$ with ijth element $\mathbb{E}_i[\mathrm{e}^{\alpha S_t};\, J_t = j]$ where

$$\boldsymbol{K}[\alpha] \;=\; \boldsymbol{\Lambda} + \bigl(\kappa^{(i)}(\alpha)\bigr)_{\text{diag}} + \bigl(\lambda_{ij} q_{ij} (\widehat{B}_{ij}[\alpha] - 1)\bigr).$$

For a fixed θ, write

$$L_t \;=\; \frac{h^{(\theta)}_{J_t}}{h^{(\theta)}_{J_0}} \mathrm{e}^{\theta S_t - t\kappa(\theta)}$$

This is just the Wald martingale normalized to have mean 1, and Proposition 3.1 immediately gives the existence of a probability measure $\widetilde{\mathbb{P}}_i$ such that

$$\widetilde{\mathbb{P}}_i(A) \;=\; \mathbb{E}_i\left[L_t;\, A\right], \quad A \in \mathscr{F}_t \;=\; \sigma\bigl((J_v, S_v) : v \le t\bigr). \qquad (8.1)$$

Theorem 8.1 *Consider the irreducible case with E finite. Then the family $\{\widetilde{\mathbb{P}}_i\}_{i\in E}$ defines a new MAP with parameters given by*

$$\widetilde{\boldsymbol{P}} = \mathrm{e}^{-\kappa(\theta)}\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\widehat{\boldsymbol{F}}[\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}, \quad \widetilde{H}_{ij}(\mathrm{d}x) = \frac{\mathrm{e}^{\theta x}}{\widehat{H}_{ij}[\theta]}H_{ij}(\mathrm{d}x)$$

in the discrete–time case, and by

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\boldsymbol{K}[\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}} - \kappa(\theta)\boldsymbol{I}, \;\; \widetilde{\mu}_i = \mu_i + \theta\sigma_i^2 - \int_{-1}^{1}[\mathrm{e}^{\theta x}-1]\nu_i(\mathrm{d}x), \;\; \widetilde{\sigma}_i^2 = \sigma_i^2,$$

$$\widetilde{\nu}_i(\mathrm{d}x) = e^{\theta x}\nu_i(\mathrm{d}x), \quad \widetilde{q}_{ij} = \frac{q_{ij}\widehat{B}_{ij}[\theta]}{1+q_{ij}(\widehat{B}_{ij}[\theta]-1)}, \quad \widetilde{B}_{ij}(dx) = \frac{e^{\theta x}}{\widehat{B}_{ij}[\theta]}B_{ij}(\mathrm{d}x)$$

in the continuous–time case. Here $\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}$ is the diagonal matrix with the $h_i^{(\theta)}$ on the diagonal. In particular, if $\nu_i(\mathrm{d}x)$ is compound Poisson, $\nu_i(\mathrm{d}x) = \beta_i B_i(dx)$ with $\beta_i < \infty$ and B_i a probability measure, then also $\widetilde{\nu}_i(\mathrm{d}x)$ is compound Poisson with

$$\widetilde{\beta}_i = \beta_i\widehat{B}_i[\theta], \quad \widetilde{B}_i(\mathrm{d}x) = \frac{\mathrm{e}^{\theta x}}{\widehat{B}_i[\theta]}B_i(\mathrm{d}x).$$

Remark 8.2 The expression for $\widetilde{\boldsymbol{\Lambda}}$ means

$$\widetilde{\lambda}_{ij} \;=\; \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\lambda_{ij}\big[1+q_{ij}(\widehat{B}_{ij}[\theta]-1)\big], \quad i \neq j. \tag{8.2}$$

In particular, this gives a direct verification that $\widetilde{\boldsymbol{\Lambda}}$ is an intensity matrix: the off–diagonal elements are nonnegative because $\lambda_{ij} \geq 0$, $0 \leq q_{ij} \leq 1$ and $\widehat{B}_{ij}[\theta] > 0$. That the rows sum to 0 follows from

$$\begin{aligned}\widetilde{\boldsymbol{\Lambda}}\mathbf{1} \;&=\; \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\boldsymbol{K}[\theta]\boldsymbol{h}^{(\theta)} - \kappa(\theta)\mathbf{1} \;=\; \kappa(\theta)\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\boldsymbol{h}^{(\theta)} - \kappa(\theta)\mathbf{1} \\ &=\; \kappa(\theta)\mathbf{1} - \kappa(\theta)\mathbf{1} \;=\; \mathbf{0}\;.\end{aligned}$$

That $0 \leq \widetilde{q}_{ij} \leq 1$ follows from the inequality

$$\frac{qb}{1+q(b-1)} \leq 1, \quad 0 \leq q \leq 1, \; 0 < b < \infty. \qquad \square$$

Proof of Theorem 8.1. First note that the ijth element of $\widehat{\widetilde{\boldsymbol{F}}}_t[\alpha]$ is

$$\widetilde{\mathbb{E}}_i[\mathrm{e}^{\alpha S_t}; J_t = j] \;=\; \mathbb{E}_i[L_t\mathrm{e}^{\alpha S_t}; J_t = j] \;=\; \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\mathrm{e}^{-t\kappa(\theta)}\mathbb{E}_i[\mathrm{e}^{(\alpha+\theta)S_t}; J_t = j].$$

In matrix notation, this means that

$$\widehat{\widetilde{\boldsymbol{F}}}_t[\alpha] \;=\; \mathrm{e}^{-t\kappa(\theta)}\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\widehat{\boldsymbol{F}}_t[\alpha+\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}\;. \tag{8.3}$$

Consider first the discrete–time case. Here the stated formula for $\widetilde{\boldsymbol{P}}$ follows immediately by letting $t=1$, $\alpha=0$ in (8.3). Further

$$\begin{aligned}\widetilde{F}_{ij}(\mathrm{d}x) &= \widetilde{\mathbb{P}}_i(Y_1\in \mathrm{d}x, J_1=j) \;=\; \mathbb{E}_i[L_1;\, Y_1\in\mathrm{d}x, J_1=j]\\ &= \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\mathrm{e}^{\theta x-\kappa(\theta)}\mathbb{P}_i(Y_1\in\mathrm{d}x, J_1=j) \;=\; \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\mathrm{e}^{\theta x-\kappa(\theta)}F_{ij}(\mathrm{d}x).\end{aligned}$$

This shows that $\widetilde{F}_{ij}$ is absolutely continuous w.r.t. F_{ij} with a density proportional to $\mathrm{e}^{\theta x}$. Hence the same is true for $\widetilde{H}_{ij}$ and H_{ij}; since $\widetilde{H}_{ij}$, H_{ij} are probability measures, it follows that indeed the normalizing constant is $\widehat{H}_{ij}[\theta]$.

Similarly, in continuous time (8.3) yields

$$\mathrm{e}^{t\widetilde{\boldsymbol{K}}[\alpha]} \;=\; \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\mathrm{e}^{t(\boldsymbol{K}[\alpha+\theta]-\kappa(\theta)\boldsymbol{I})}\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}.$$

By the general formula $\boldsymbol{\Delta}^{-1}\mathrm{e}^{\boldsymbol{A}}\boldsymbol{\Delta} = \mathrm{e}^{\boldsymbol{\Delta}^{-1}\boldsymbol{A}\boldsymbol{\Delta}}$ ($\boldsymbol{\Delta}$ diagonal), this implies

$$\widetilde{\boldsymbol{K}}[\alpha] = \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\bigl(\boldsymbol{K}[\alpha+\theta]-\kappa(\theta)\boldsymbol{I}\bigr)\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}} = \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}\boldsymbol{K}[\alpha+\theta]\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}} - \kappa(\theta)\boldsymbol{I}.$$

Letting $\alpha=0$ yields the stated expression for $\widetilde{\boldsymbol{\Lambda}}$.

Now we can write

$$\begin{aligned}\widetilde{\boldsymbol{K}}[\alpha] &= \widetilde{\boldsymbol{\Lambda}} + \boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}^{-1}(\boldsymbol{K}[\alpha+\theta]-\boldsymbol{K}[\theta])\boldsymbol{\Delta}_{\boldsymbol{h}^{(\theta)}}\\ &= \widetilde{\boldsymbol{\Lambda}} + \bigl(\kappa^{(i)}(\alpha+\theta)-\kappa^{(i)}(\theta)\bigr)_{\mathrm{diag}}\\ &\qquad + \left(\frac{h_j^{(\theta)}}{h_i^{(\theta)}}\lambda_{ij}q_{ij}\bigl(\widehat{B}_{ij}[\alpha+\theta]-\widehat{B}_{ij}[\theta]\bigr)\right).\end{aligned}$$

That $\kappa^{(i)}(\alpha+\theta)-\kappa^{(i)}(\theta)$ corresponds to the stated parameters $\widetilde{\mu}_i, \widetilde{\sigma}_i^2$, $\widetilde{\nu}_i$ of a Lévy process follows from Theorem 3.4. Finally note that by (8.2),

$$\begin{aligned}\frac{h_j^{(\theta)}}{h_i^{(\theta)}}\lambda_{ij}q_{ij}\bigl(\widehat{B}_{ij}[\alpha+\theta]-\widehat{B}_{ij}[\theta]\bigr) &= \frac{h_j^{(\theta)}}{h_i^{(\theta)}}\lambda_{ij}q_{ij}\widehat{B}_{ij}[\theta]\bigl(\widehat{\widetilde{B}}_{ij}[\alpha]-1\bigr)\\ &= \widetilde{\lambda}_{ij}\widetilde{q}_{ij}\bigl(\widehat{\widetilde{B}}_{ij}[\alpha]-1\bigr).\end{aligned}$$

□

As a main application, we give the analogue of the Cramér–Lundberg approximation and Corollary 5.9. Let $I(u) = J_{\tau(u)}$ and use obvious notation like $\tau(u) = \inf\{t>0: S_t>u\}$, $B(u) = S_{\tau(u)}-u$, $M = \sup_{t\ge 0} S_t$, $M_- = \sup_{0\le t<\tau_-} S_t$, etc. Write further $\mathbb{P}_{i,\theta}$ for the probability measure $\widetilde{\mathbb{P}}$ constructed above and conditioned to $J_0=i$.

Theorem 8.3 *Consider a MAP in discrete or continuous time with $\kappa'(0)<0$, and $\kappa(\gamma)=0$, $\kappa'(\gamma)<\infty$ for some $\gamma>0$. Assume further that $\bigl\{\bigl(I(x), B(x)\bigr)\bigr\}$ satisfies the nonlattice condition of VII.5.2(i) for the* *existence of a limit* $\bigl(I(\infty), B(\infty)\bigr)$. *Then*

$$\mathbb{P}_i(M > u) \ \sim\ h_i^{(\gamma)} C \mathrm{e}^{-\gamma u} \quad \textit{where} \quad C \ = \ \mathbb{E}_{i,\gamma}\left[\frac{\mathrm{e}^{-\gamma B(\infty)}}{h_{I(\infty)}^{(\gamma)}}\right].$$

Similarly, if $\mathbb{P}_i(\tau_- > 0) = 1$, *then*

$$\mathbb{P}_i(M_- > u) \ \sim\ h_i^{(\gamma)} C \mathbb{P}_{i,\gamma}(\tau_- = \infty)\mathrm{e}^{-\gamma u}\,.$$

Proof. By the stopping time version of (8.1),

$$\begin{aligned}\mathbb{P}_i(M > u) \ &= \ \mathbb{P}_i(\tau(u) < \infty) \ = \ \mathbb{E}_{i,\gamma}\left[\frac{h_{J_0}^{(\theta)}}{h_{J_{\theta(u)}}^{(\theta)}}\mathrm{e}^{-\gamma S_{\tau(u)}};\, \tau(u) < \infty\right] \\ &= \ \mathrm{e}^{-\gamma u}\mathbb{E}_{i,\gamma}\left[\frac{h_i^{(\theta)}}{h_{I(u)}^{(\theta)}}\mathrm{e}^{-\gamma B(u)}\right],\end{aligned}$$

where we used $\mathbb{P}_{i,\gamma}(\tau(u) < \infty)$, as follows from $\kappa'(\gamma) > 0$ (convexity!). The asymptotics for $\mathbb{P}_i(M > u)$ therefore immediately follows from $\bigl(I(u), B(u)\bigr) \xrightarrow{\mathscr{D}} \bigl(I(\infty), B(\infty)\bigr)$. The case of $\mathbb{P}_i(M_- > u)$ is a similar asymptotic independence argument as in the proof of Corollary 5.9. □

Notes As in Remark 5.4, the conditions for existence of $(I(\infty), B(\infty))$ in Theorem 8.3 are very mild. However, to find simple expressions for C in terms of model parameters is harder than in one dimension; examples can be found e.g. in Asmussen (2000) and Miyazawa (2002, 2004).

XIV
Dams, Inventories and Insurance Risk

1 Compound Poisson Dams with General Release Rule

This model originates from problems of storage of water in dams or reservoirs. Water flows in, say from a river or several creeks, according to an input process $\{A_t\}_{t\geq 0}$ and is released at a rate $r(x)$ depending on the present content x of the dam. We let X_t be the content at time t and shall be interested in the ergodicity problems for the process $\{X_t\}_{t\geq 0}$. From a practical point of view, the stationary distribution π is of importance for assessing values of quantities such as the proportion $\pi_0 = \pi(\{0\})$ of time the dam is empty and the average release rate $\int_0^\infty r(x)\,\pi(\mathrm{d}x)$. Some guidelines for the choice of r are thereby possibly also provided.

We shall assume that $\{A_t\}$ is a compound Poisson process,

$$A_t \;=\; \sum_{n=1}^{N_t} U_n, \tag{1.1}$$

where $\{N_t\}$ is a Poisson process with intensity β and $U_1, U_2, \ldots$ are i.i.d. with distribution B and independent of $\{N_t\}$ (here $U > 0$, i.e. $B(0) = 0$). In terms of water storage, this corresponds intuitively to the input to the dam being due mainly to rare large rainfalls. This assumption is acceptable for the dry climatic conditions for which the theory was initially developed, but certainly not always. Thus it would frequently be reasonable to add a drift term ct to (1.1), and also the effects of frequent small rainfalls may not be negligible, which would lead to $\{A_t\}$ being a Lévy process with only positive jumps. These cases will, however, not be discussed here.

The dam is taken to be infinitely high, i.e. the state space for $\{X_t\}$ is $[0,\infty)$. The release rate being $r(x)$ at content x means that in between jumps, $\{X_t\}$ should satisfy the differential equation

$$\dot{x} \;=\; -r(x), \tag{1.2}$$

where $\dot{x}$ means left derivative. We shall assume that r is in $D(0,\infty)$ with

$$0 < \inf_{\epsilon<x<\epsilon^{-1}} r(x), \qquad \sup_{0<x<\epsilon^{-1}} r(x) < \infty \tag{1.3}$$

for any $\epsilon \in (0,1)$, and extend $r(x)$ to $x = 0$ by letting $r(0) = 0$ (in applications, $r(x)$ will typically be nondecreasing). It is then easily seen that for each $x_0 > 0$ (1.2) has a unique solution x_t starting at x_0. In fact, if we let

$$\theta(x;y) \;=\; \int_y^x \frac{1}{r(v)}\,\mathrm{d}v, \quad x \geq y, \tag{1.4}$$

then $(\mathrm{d}/\mathrm{d}t)\theta(x_0;x_t) = 1$ and $\theta(x_0;x_0) = 0$ yields $\theta(x_0;x_t) = t$. That is, x_t is the inverse function of $\theta(x_0;\cdot)$ or equivalently, $\theta(x;y)$ is the time required for x_t to pass from x to y. Note that in view of $r(0) = 0$, x_t gets absorbed at 0 once 0 is hit, which happens if and only if $\theta(x;0) < \infty$ for some (and then all) $x > 0$.

The construction of the process is now obvious since we only have to start at say $X_0 = x$ and let the process move deterministically according to (1.2) until the first jump of $\{A_t\}$ where $\{X_t\}$ jumps the same amount. Then (1.2) governs the motion until the next jump and so on. These properties can be summarized in the so–called *storage equation*

$$X_t \;=\; x + A_t - \int_0^t r(X_s)\,\mathrm{d}s \tag{1.5}$$

and under the given set of assumptions, it is easy to see that there is a unique solution [(1.5) simply reflects that $\{X_t\}$ has the same upward jumps as $\{A_t\}$ and moves according to (1.2), with $r(0) = 0$, in between jumps]. It is also intuitively clear from the construction that $\{X_t\}$ is Markov, and this is readily checked as well as the strong Markov property; cf. Problem 1.1. An example of a sample path is in Fig. 1.1. Here $r(x) = 1+x$ corresponding to $x_t = (x_0+1)\mathrm{e}^{-t} - 1$, and $\theta(x;0) < \infty$ because of $\underline{\lim}_{x\to 0} r(x) > 0$. We note that if $r(x) \equiv 1$, then $\{X_t\}$ is simply the workload process of a $M/G/1$ queue; cf. Fig. III.1.4.

Preparing for the study of the ergodicity problem, let $\tau_+(u) = \inf\{t \geq 0 : X_t > u\}$, $\tau_-(u) = \inf\{t \geq 0 : X_t \leq u\}$ and $\tau(u) = \inf\{t \geq 0 : X_t = u\}$.

Lemma 1.1 *For any $u \in (0,\infty)$ and $T > 0$, there is an $\epsilon > 0$ such that $\mathbb{P}_x(\tau_+(u) \leq T) \geq \epsilon$ for all $x \leq u$.*

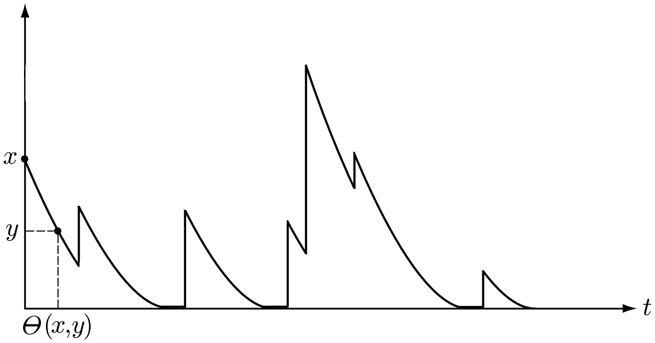

Figure 1.1

Proof. Define $m = \sup_{x \le u} r(x)$, $F = \{A_T > u + mT\}$. Then $\epsilon = \mathbb{P}F > 0$ since the distribution of A_T as compound Poisson has unbounded support. Furthermore, $\tau_+(u) \le T$ on F, since otherwise $X_t \le u$, $r(X_t) \le m$ for all $t \le T$, and the storage equation yields the contradiction

$$X_T \ > \ 0 + u + mT - \int_0^T r(X_t)\,\mathrm{d}t \ \ge \ u. \qquad \square$$

Proposition 1.2 *The process is either transient in the sense that* $\mathbb{P}_x(X_t \to \infty) = 1$ *for all* $x \ge 0$*, or recurrent in the sense that* $\mathbb{P}_x(\tau(u) < \infty) = 1$ *for all* $x \ge 0$*,* $u > 0$*.*

Proof. For $v \ge 0$, define $F = \{\underline{\lim}_{t\to\infty} X_t < v\}$. If $\mathbb{P}_x F = 0$ for all $x \ge 0$, $v > 0$, it is clear that $\mathbb{P}_x(X_t \to \infty) = 1$, so suppose $\mathbb{P}_x F > 0$ for some $x \ge 0$, $v > 0$. On F, there exists an increasing sequence $\{\sigma_k\}$ of stopping times with $\sigma_{k+1} - \sigma_k \ge 1$ (say), $X_{\sigma_k} \le v$. Then by Lemma 1.1 with $T = 1$, we have for $u > v$ that

$$\sum_{k=1}^{\infty} \mathbb{P}_x\big(X_t > u \text{ for some } t \in [\sigma_k, \sigma_{k+1}) \,\big|\, \mathscr{F}_{\sigma_k}\big)$$
$$\ge \ \sum_{k=1}^{\infty} \mathbb{P}_{X_{\sigma_k}}(\tau_+(u) \le 1) \ \ge \ \sum_{k=1}^{\infty} \epsilon \ = \ \infty$$

on F. Hence, by the conditional Borel–Cantelli lemma it occurs on F for infinitely many k that $X_t > u$ for some $t \in [\sigma_k, \sigma_{k+1})$. Since $X_{\sigma_{k+1}} \le v$, we have thus by the downward skip–free property of the paths that u is visited i.o. in between visits to $[0, v]$. This is only possible if $\mathbb{P}_u(\tau(u) < \infty) = 1$. Hence, starting from u, there are infinitely many returns to u, and since any state $y > u$ can be reached from u, we must have $\mathbb{P}_y(\tau(u) < \infty) = 1$. To get $\mathbb{P}_0(\tau(u) < \infty) = 1$, just condition on the first state y entered at the jump away from zero. $\square$

It follows easily from irreducibility properties of the process that in the recurrent case, $\mathbb{E}_u \tau(u)$ is either finite for all $u > 0$ or infinite for all $u > 0$.

For obvious reasons, we refer to the two possibilities as *positive recurrence* and *null recurrence*, respectively. In either case, it follows from VII.3 that $\{X_t\}$ has a stationary measure ν which is unique up to a constant, and which for any $u > 0$ may be written as $\nu = c_u \nu^{(u)}$ where

$$\int_0^\infty f(x)\nu^{(u)}(\mathrm{d}x) \;=\; \mathbb{E}_u \int_0^{\tau(u)} f(X_t)\,\mathrm{d}t. \tag{1.6}$$

Proposition 1.3 *For any $u > 0$, the function $z(t) = \mathbb{P}_u(X_t \le u, \tau(u) > t)$ tends to zero exponentially fast. In particular, the stationary measure ν is Radon* ($\nu(A) < \infty$ for A compact) *and in the null recurrent case $\mathbb{P}_x(X_t \le u) \to 0$ for all $x, u \ge 0$. That is, $X_t \to \infty$ in $\mathbb{P}_x$–distribution.*

Proof. An inspection of the paths shows that $z(t) \le \mathbb{P}_u(\tau_+(u) > t)$. Letting H_n be the conditional distribution of X_n given $\tau_+(u) > n$, Lemma 1.1 yields

$$\begin{aligned}
\mathbb{P}_u(\tau_+(u) > n+1) \;&=\; \mathbb{P}_u(\tau_+(u) > n)\int_0^u \mathbb{P}_x(\tau_+(u) > 1)\,H_n(\mathrm{d}x)\\
&\le\; (1-\epsilon)\mathbb{P}_u(\tau_+(u) > n)
\end{aligned}$$

and the exponential decay of $z(t)$ follows. Clearly z is measurable, thus Lebesgue integrable. Therefore $\nu^{(u)}[0,u] = \int z < \infty$, the truth of which for all u shows that ν is Radon. Since clearly the cycle length distribution is absolutely continuous with mean $\mathbb{E}_u\tau(u)$ which is infinite in the null recurrent case, VII.3.8(iv) then yields $\mathbb{P}_x(X_t \le u) \to 0$ for all x, u, so that $X_t \xrightarrow{\mathscr{D}} \infty$. □

Theorem 1.4 *The stationary measure ν has an atom $\nu_0 = \nu(\{0\}) > 0$ at zero if and only if $\theta(x;0) < \infty$ for some* (and then all) $x > 0$. *Also ν is absolutely continuous on* $(0,\infty)$ *and there exists a version g of the density satisfying*

$$\begin{aligned}
g(x) \;&=\; \frac{1}{r(x)}\beta\int_0^x \overline{B}(x-y)\,\nu(\mathrm{d}y) &(1.7)\\
&=\; \frac{\beta}{r(x)}\left\{\nu_0\overline{B}(x) + \int_0^x \overline{B}(x-y)g(y)\,\mathrm{d}y\right\}. &(1.8)
\end{aligned}$$

Proof. We take $\nu = \nu^{(u)}$ for a while. Starting from u, we reach zero with positive probability if and only if $\theta(u;0) < \infty$, and then have a nonzero sojourn time. From this the statement concerning ν_0 is clear. For the absolute continuity, note that a particle moving at a speed at least δ spends at most time $\delta^{-1}|A|T$ in the set A within T units of time. Hence, if $A \subseteq (\epsilon^{-1}, \epsilon)$ and $|A| = 0$, it follows by (1.3) that $\int_0^\infty I(X_t \in A)\,\mathrm{d}t = 0$, implying $\nu^{(u)}(A) = 0$ and absolute continuity.

For the proof of (1.7), we apply rate conservation to $Z_t = I(X_t > x)$. If no Poisson arrivals occur in $(t, t+h]$, there will be a downcrossing (negative jump) precisely when $X_t \in (x, x+z(h)]$ where $\theta(x+z(h);x) = h$. Since

$z(h) = r(x)h + o(h)$, the probability of this is

$$\mathrm{e}^{-\beta h}\mathbb{P}_e\bigl(X_t \in (x, x+z(h)]\bigr) \;=\; \mathrm{e}^{-\beta h}\int_x^{x+z(h)} g(y)\,\mathrm{d}y \;=\; r(x)g(x)h + o(h).$$

The probability of a downcrossing and a Poisson event in $(t, t{+}h]$ is bounded by

$$(1-\mathrm{e}^{-\beta h})\int_0^{x+z(h)} B\bigl(x-y, x-y+z(h)\bigr]\,\nu(\mathrm{d}y)$$

which is $o(h)$ by monotone convergence. It follows that the rate of downcrossings is $r(x)g(x)$. The rate of upcrossings (positive jumps) is $\beta\int_0^x \overline{B}(x-y)\,\nu(\mathrm{d}y)$, and (1.7) follows by equating these two rates. □

In the positive recurrent case $\|\nu\| < \infty$, we can define the unique stationary distribution π by $\pi = \nu/\|\nu\|$ and have by general results on regenerative processes that $X_t \to \pi$ in $\mathbb{P}_x$–distribution for all x. In that case, the solution g to (1.7) and (1.8) exists and is integrable.

We shall next show that conversely the existence of an integrable solution to (1.7) and (1.8) implies positive recurrence. To this end, let $\{Y_n\}$ be the content just before the nth jump. Then $\{Y_n\}$ is a Markov chain, and we have:

Lemma 1.5 (i) *Either* $\{Y_n\}$ *is transient in the sense that* $\mathbb{P}_x(Y_n \to \infty) = 1$ *for all* x*, or* $\{Y_n\}$ *is recurrent in the sense that* $\mathbb{P}_x(Y_n \le v \text{ i.o.}) = 1$ *for all* $x \ge 0$*,* $v > 0$*;* (ii) *in the recurrent case,* $\{Y_n\}$ *is Harris recurrent;* (iii) *if a distribution* π *has the property that* (1.7) *and* (1.8) *hold for* $\nu = \pi$*, then* π *is stationary for* $\{Y_n\}$*.*

Proof. Here (i) is shown similarly as in Proposition 1.2. For (ii), let $R = [0, v]$ for some arbitrary $v > 0$ and choose $0 < c < d < f < \infty$ such that $\delta_1 = \mathbb{P}\bigl(U \in (d, f]\bigr) > 0$. Now for $(a, b) \subseteq (c, d)$ and $x \in R$,

$$\mathbb{P}\bigl(Y_1 \in (a,b)\,\big|\,Y_0 = x\bigr) \;\ge\; \int_d^f \mathbb{P}\bigl(\theta(x+u;b) < T < \theta(x+u;a)\bigr)\,B(\mathrm{d}u)$$

$$=\; \int_d^f \mathrm{e}^{-\beta\theta(x+u;b)}\bigl(1-\mathrm{e}^{-\beta\theta(b;a)}\bigr)\,B(\mathrm{d}u) \;\ge\; \delta_1\mathrm{e}^{-\beta\theta(f+v;c)}\bigl(1-\mathrm{e}^{-\beta\theta(b;a)}\bigr)$$

where T is the time between the first (at $t = 0$) and second jump. But from (1.3), we can find $\delta_2 > 0$ such that $1-\mathrm{e}^{-\beta\theta(b;a)} \ge \delta_2(b-a)$ for $(a, b) \subseteq (c, d)$. Hence

$$\mathbb{P}\bigl(Y_1 \in (a,b)\,\big|\,Y_0 = x\bigr) \;\ge\; \delta_3(b-a), \quad x \in R,$$

so that the minorization condition VII.(3.1) holds if we take λ to be the uniform distribution on (c, d). This proves (ii). For (iii), let $\mathbb{P}_{\pi-}$ refer to the initial condition where the first jump occurs at time zero and $X_{0-} = Y_0$ has distribution π. Then X_0 has distribution $\pi * B$, and $Y_1 > z$ will occur if $X_0 = y > z$ and the first jump occurs before $\theta(y;z)$. Hence

$$\begin{aligned}\mathbb{P}_{\pi-}(Y_1 > z) &= \int_z^\infty \bigl(1 - \mathrm{e}^{-\beta\theta(y;z)}\bigr)\,(\pi * B)(\mathrm{d}y) \\ &= \int_z^\infty \frac{\beta}{r(y)}\mathrm{e}^{-\beta\theta(y;z)}\overline{(\pi * B)}(y)\,\mathrm{d}y. \qquad (1.9)\end{aligned}$$

Now clearly by (1.7),

$$\overline{(\pi * B)}(y) = \overline{\pi}(y) + \int_0^y \overline{B}(y-x)\,\pi(\mathrm{d}x) = \overline{\pi}(y) + \frac{1}{\beta}g(y)r(y)$$

and hence (1.9) becomes

$$\begin{aligned}&\int_z^\infty \frac{\beta}{r(y)}\mathrm{e}^{-\beta\theta(y;z)}\overline{\pi}(y)\,\mathrm{d}y + \int_z^\infty \mathrm{e}^{-\beta\theta(y;z)}g(y)\,\mathrm{d}y \\ &\quad = \int_z^\infty \bigl(1 - \mathrm{e}^{-\beta\theta(y;z)}\bigr)\,\pi(\mathrm{d}y) + \int_z^\infty \mathrm{e}^{-\beta\theta(y;z)}\,\pi(\mathrm{d}y) = \overline{\pi}(z),\end{aligned}$$

proving stationarity of π for $\{Y_n\}$. □

Theorem 1.6 *The process $\{X_t\}$ is positive recurrent if and only if there exists a probability measure $\pi(x) = \pi_0 + \int_0^x g(y)\,\mathrm{d}y$ such that* (1.7) *and* (1.8) *hold for $\nu = \pi$. In that case, the solution to* (1.7) *and* (1.8) *is unique and the stationary distribution.*

Proof. The existence of a solution in the positive recurrent case follows from Theorem 1.4. If, conversely, π is a solution with $\|\pi\| = 1$, then by Lemma 1.5(iii) π is a stationary distribution for $\{Y_n\}$, the existence of which implies first that $\{Y_n\}$ cannot be transient, cf. (i), and next by (ii) that $\{Y_n\}$ is positive recurrent and that the solution to (1.7) and (1.8) is unique. It thus only remains to show that $\{X_t\}$ is indeed positive recurrent if π exists. But then by the PASTA property VII.6.7 the time–averages $\int_0^t I(X_s \le u)\,\mathrm{d}s/t$ have nonzero limits, which excludes transience and (by Proposition 1.3 and Fatou's lemma) null recurrence. □

In the case $\theta(x;0) < \infty$, it is possible to give an alternative characterization of π. For an integral kernel $K(x,y)$ defined for $0 \le y < x$, a function g and another kernel $K'(x,y)$, define

$$Kg(x) = \int_0^x K(x,z)g(z)\,\mathrm{d}z, \quad KK'(x,y) = \int_y^x K(x,z)K'(z,y)\,\mathrm{d}z$$

One easily checks the associative law $(KK')K'' = K(K'K'')$ and hence we can recursively define $K^n = K^{n-1}K = KK^{n-1}$ (here $K^1 = K$). Now take $K(x,y) = \beta\overline{B}(x-y)/r(x)$. Using $K(x,y) \le \beta/r(x)$, it follows easily by induction that

$$K^{n+1}(x,y) \le \beta^{n+1}\theta(x;y)^n/r(x)(n+1)!. \qquad (1.10)$$

Hence $K^* = \sum_1^\infty K^n$ is well defined and finite, and we have:

Corollary 1.7 *If $\theta(x;0) < \infty$, $x > 0$, then $\{X_t\}$ is positive recurrent if and only if*

$$1 + \int_0^\infty K^*(x,0)\,\mathrm{d}x \;<\; \infty, \tag{1.11}$$

in which case π_0 is the reciprocal of (1.11) *and $g(x) = \pi_0 K^*(x,0)$, $x > 0$.*

Proof. Let $g_0(x) = K(x,0)$. Then

$$g \;=\; \pi_0 g_0 + Kg \;=\; \pi_0(g_0 + Kg_0) + K^2 g \;=\; \pi_0 \sum_{n=0}^{N-1} K^n g_0 + K^N g,$$

where the first identity is just (1.8). But $K^N g \to 0$ by (1.10) and $K^n g_0(x)$ is just $K^{n+1}(x,0)$. Hence $g = \pi_0 K^*(x,0)$ which yields the desired conclusions in the positive recurrent case. If, conversely, (1.11) holds and we define π as indicated, then

$$\pi_0 g_0(x) + Kg(x) \;=\; \pi_0 g_0(x) + \pi_0 \sum_{n=2}^{\infty} K^n(x,0) \;=\; \pi_0 K^*(x,0) \;=\; g(x)$$

so that π is a probability measure satisfying (1.7). □

Also in the case $\theta(x;0) = \infty$, one can give a criterion for positive recurrence in terms of K^*:

Corollary 1.8 *The process is positive recurrent if and only if $\int_a^\infty K^*(x,a)\mathrm{d}x < \infty$ for some* (and then for all) *$a > 0$.*

Proof. Let $\{\widetilde{X}_t\}$ correspond to $\widetilde{r}(x) = r(x+a)$ and the same β, B. Then, in the obvious notation, $\widetilde{\theta}(x;0) < \infty$ because of $\inf_{x+a\geq y\geq a} r(y) > 0$ so that we may apply Corollary 1.7 to see that $\{\widetilde{X}_t\}$ is positive recurrent if and only if $\widetilde{K}^*(x;0)$ is integrable. But for $n = 1$, we have $\widetilde{K}_n(x,y) = K_n(x+a, y+a)$ (both coincide with $\beta\overline{B}(x-y)/r(x+a)$) and it follows easily by induction that this is also valid for $n > 1$. Hence $\int_0^\infty \widetilde{K}^*(x,0)\,\mathrm{d}x$ and $\int_a^\infty K^*(x,a)\,\mathrm{d}x$ are equal, in particular finite at the same time, so that we have only to show that $\{\widetilde{X}_t\}$ and $\{X_t\}$ are positive recurrent at the same time. But for $x > y > a$, $\mathbb{E}_x\tau(y) = \mathbb{E}_{x-a}\widetilde{\tau}(y-a)$, and by irreducibility properties of the processes, it is easily seen that $\mathbb{E}_x\tau(y) < \infty$ exactly when $\{X_t\}$ is positive recurrent, and that $\mathbb{E}_{x-a}\widetilde{\tau}(y-a) < \infty$ exactly when $\{\widetilde{X}_t\}$ is positive recurrent. □

The above result concludes the treatment of the general theory, and we proceed in the next section to see how the derived formulas and criteria take a more explicit form in some particular cases.

Problems

1.1 Show that $\{X_t\}$ has the strong Markov property. [*Hint:* Recall the proof of V.1.5.]

1.2 In a storage problem with $A_t = t/2$, $r(x) = 1$, $x > 0$, it seems reasonable to define $X_t = (X_0 - t/2)^+$. Show that the storage equation is not satisfied.
1.3 Show by an example that it is possible that $\lim_{x\to\infty} \theta(x;y) < \infty$ and that there then is positive recurrence.

Notes The theory was initiated by Australian authors in the 1960s; Brockwell *et al.* (1982) give a development in the present spirit (though the present proofs are in part different) and also treat the Lévy case.

The dam process (also frequently called the *storage process*) is a model example of a Markov process that is piecewise deterministic in the sense of Davis (1993); however, the general theory of such processes does not cover results of the type treated in this section.

Further aspects of the theory include tail asymptotics of the stationary distribution; see Sundt and Teugels (1995, 1997) and Asmussen and Nielsen (1995) for the light–tailed case and Asmussen (1998a) for the heavy–tailed case.

In the terminology of the theory of integral equations, (1.8) is of *Volterra type.*

2 Some Examples

The model and notation is that of the preceding section. In the following, let $\mu_B = \int_0^\infty x\, B(\mathrm{d}x) = \int_0^\infty \overline{B}(x)\,\mathrm{d}x$.

Example 2.1 *Constant release*, say $r(x) \equiv 1$. This case is already well known from the $M/G/1$ workload interpretation, but is treated here for the sake of illustration. Here $\theta(x;0) = x < \infty$ so that we may apply Corollary 1.7. Since $\int_0^\infty K(x,0)\,\mathrm{d}x = \beta\mu_B$, it is necessary for positive recurrence that $\mu_B < \infty$. If $\mu_B < \infty$, define $\rho = \beta\mu_B$, $b_0(x) = \mu_B^{-1}\overline{B}(x)$. Then $K(x,0) = \rho\, b_0(x)$ and it follows easily by induction that $K_n(x,0) = \rho^n b_0^{*n}(x)$. Hence $\pi_0^{-1} = \sum_0^\infty \rho^n$ will be finite and positive recurrence hold if and only if $\rho < 1$, in which case the expression for π in Corollary 1.7 is immediately seen to coincide with the Pollaczeck–Khintchine formula VIII.(5.5). □

Example 2.2 *Arbitrary release and exponential input*, $B(\mathrm{d}x) = \delta\mathrm{e}^{-\delta x}\,\mathrm{d}x$. Here (1.8) becomes

$$g(x) \;=\; \frac{\beta}{r(x)}\Bigl\{\pi_0\mathrm{e}^{-\delta x} + \int_0^x \mathrm{e}^{-\delta(x-y)}g(y)\,\mathrm{d}y\Bigr\}$$

which, letting $\psi(x) = \int_0^x \mathrm{e}^{\delta y}g(y)\,\mathrm{d}y$, may be rewritten as

$$\psi'(x) \;=\; \frac{\beta}{r(x)}\,\{\pi_0 + \psi(x)\}\,.$$

This is a differential equation of standard type and any solution may be written in the form

$$\psi(x) \;=\; c\mathrm{e}^{\beta\theta(x;1)} - \pi_0 \tag{2.1}$$

so that

$$g(x) \;=\; \mathrm{e}^{-\delta x}\psi'(x) \;=\; \frac{c\beta}{r(x)}\mathrm{e}^{\beta\theta(x;1)-\delta x}. \tag{2.2}$$

Thus the solution to (1.8) is unique up to a constant and we have positive recurrence if and only if

$$\alpha \;=\; \int_0^\infty \frac{\beta}{r(x)}\mathrm{e}^{\beta\theta(x;1)-\delta x}\,\mathrm{d}x \;<\; \infty.$$

It only remains to evaluate π_0 and c in the case $\alpha < \infty$. We have $\|\pi\| = 1$, which yields $\pi_0 + c\alpha = 1$. If $\theta(x;0) = \infty$, then $\pi_0 = 0$ and $c = \alpha^{-1}$. If $\theta(x;0) < \infty$, then $\psi(0) = 0$ and (2.1) with $x = 0$ yields $c = \pi_0\mathrm{e}^{\beta\theta(1;0)}$,

$$\pi_0 \;=\; \frac{1}{1+\alpha\mathrm{e}^{\beta\theta(1;0)}} \;=\; \frac{1}{1+\int_0^\infty \frac{\beta}{r(x)}\mathrm{e}^{\beta\theta(x;0)-\delta x}\,\mathrm{d}x}.$$

Note that we may rewrite (2.2) as $g(x) \;=\; \dfrac{\pi_0\beta}{r(x)}\mathrm{e}^{\beta\theta(x;0)-\delta x}$. □

Proposition 2.3 *Suppose $r(x)$ is nondecreasing. Then $\{X_t\}$ is positive recurrent if and only if*

$$\lim_{a\to\infty}\beta\int_0^\infty \frac{\overline{B}(x)}{r(x+a)}\,\mathrm{d}x \;<\; 1. \tag{2.3}$$

If $\mu_B < \infty$, (2.3) is equivalent to $\beta\mu_B < \lim_{x\to\infty} r(x)$.

Proof. The limit in (2.3) clearly exists and by the definition of $K = K_1$ is also the limit of $\int_a^\infty K(x,a)\,\mathrm{d}x$. Hence if (2.3) holds, there are a_0 and $\delta < 1$ such that $\int_a^\infty K(x,a)\,\mathrm{d}x \le \delta$ for all $a \ge a_0$. Since

$$\begin{aligned}\int_a^\infty K_{n+1}(x,a)\,\mathrm{d}x \;&=\; \int_a^\infty \mathrm{d}x\int_a^x K_n(x,y)K(y,a)\,\mathrm{d}y \\ &=\; \int_a^\infty K(y,a)\,\mathrm{d}y\int_y^\infty K_n(x,y)\,\mathrm{d}x,\end{aligned}$$

it follows by induction that $\int_a^\infty K_{n+1}(x,a)\,\mathrm{d}x \le \delta^{n+1}$ for all n. Summing yields $\int_a^\infty K^*(x,a)\,\mathrm{d}x < \infty$, $a \ge a_0$. If conversely (2.3) fails, then $\int_a^\infty K(x,a)\,\mathrm{d}x \ge 1$ for all a and thus in the same way $\int_a^\infty K^*(x,a)\,\mathrm{d}x = \infty$. Reference to Corollary 1.8 completes the proof. □

Example 2.4 *Linear release rate*, $r(x) = c + fx$. Since $r(x) \to \infty$ monotonically, positive recurrence is immediately obtained from Proposition 2.3 provided only that $\mu_B < \infty$. If $\mu_B = \infty$, we have

$$\beta\int_0^\infty \frac{\overline{B}(x)}{r(x+a)}\,\mathrm{d}x \;=\; \frac{\beta}{f}\int_0^\infty \log\Bigl(1+\frac{fx}{c+fa}\Bigr)B(\mathrm{d}x). \tag{2.4}$$

By elementary properties of the logarithm, this is finite if and only if $\int_0^\infty \log x\,B(\mathrm{d}x) < \infty$. In that case, (2.4) tends to zero as $a \to \infty$ by monotone convergence and hence (2.3) is automatic. That is, we have positive recurrence if and only if $\int_0^\infty \log x\, B(\mathrm{d}x) < \infty$. Also clearly $\theta(1;0) < \infty$, i.e. $\pi_0 > 0$, if and only if $c > 0$. □

Notes Again, Brockwell *et al.* (1982) is a basic reference. There are few explicit examples beyond the exponential case in Example 2.2. Paulsen and Gjessing (1997) give solutions in terms of special functions for Erlang(2) and H_2 distributions when $r(x) = a + bx$, but their study does not generalize even to Erlang(p) or H_p.

3 Finite Buffer Capacity Models

A simple case of finite capacity models is birth–death processes with a finite state space; see the examples in III.3. Other typical examples occur in telecommunications and data transmission systems, say that packets are stored in a finite buffer awaiting to be sent along a transmission line.

Many such models are in a natural way closely related to an infinite capacity (infinite buffer) one. The question arises what are the relations between the stationary distributions for the finite and infinite models and between the methods for computing them.

We shall consider two somewhat different set–ups. In the first, the finite capacity model $\{X_t^{(F)}\}_{t\in\mathbb{T}}$ is the restriction to a subset F of the state space E for the infinite capacity model $\{X_t\}_{t\in\mathbb{T}}$. The precise meaning of this is straightforward in discrete time ($\mathbb{T} = \mathbb{N}$); cf. the definitions preceding I.3.9. We shall not aim for the most general formulation when $\mathbb{T} = [0,\infty)$ but assume a structure that is sufficient for the examples to be considered: there are $0 = \sigma_0 < \tau_0 < \sigma_1 < \tau_1 < \sigma_2 < \cdots$ with $\sigma_n \uparrow \infty$ such that $X_t \in F$ for $\sigma_k < t < \tau_k$, $X_t \notin F$ for $\tau_k < t < \sigma_{k+1}$; we then define $X_t^{(F)} = X_t$, $0 = \sigma_0 \le t < \tau_0$, $X_{\tau_0+t}^{(F)} = X_{\sigma_1+t}$, $0 \le t < \tau_1 - \sigma_1$, and so on; see Fig. 3.1 for the case $E = \mathbb{N}$, $F = \{0, 1, \ldots, K\}$.

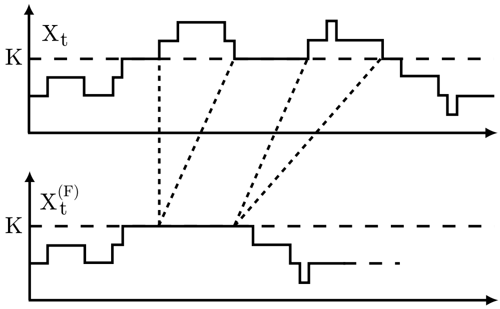

Figure 3.1

Proposition 3.1 *Assume $\{X_t\}_{t\in\mathbb{T}}$ is regenerative with state space E and generic cycle C with finite mean. Assume further that both C and $C_F = \int_0^C I(X_t \in F)\,\mathrm{d}t$ are nonlattice when $T = [0,\infty)$ and aperiodic when $T = \mathbb{N}$ (here $C_F = \sum_0^{C-1} I(X_t \in F)$). Then the stationary distributions π, $\pi^{(F)}$ of $\{X_t\}$ and $\{X_t^{(F)}\}$ exist, and $\pi^{(F)}$ is π conditioned to be in F, i.e. $\pi^{(F)}(A) = \pi(A)/\pi(F)$, $A \subseteq F$.*

Proof. Clearly, C_F is a regeneration point for $\{X_t^{(F)}\}$, and so existence follows from general regenerative process theory which also yields $\mathbb{E}C_F = \mathbb{E}C \cdot \pi(F)$. [Note that it may happen that $X_t^{(F)} \notin F$ for $t = 0, \tau_0, \tau_0 + \tau_1 - \sigma_1, \ldots$ but that these t form a null set.] We further get

$$\begin{aligned}\pi^{(F)}(A) &= \frac{1}{\mathbb{E}C_F}\,\mathbb{E}\int_0^{C_F} I\big(X_t^{(F)} \in A\big)\,\mathrm{d}t \\ &= \frac{1}{\pi(F)\mathbb{E}C}\int_0^C I(X_t \in A)\,\mathrm{d}t \;=\; \frac{\pi(A)}{\pi(F)}\end{aligned}$$

for $A \subseteq F$ and $\mathbb{T} = [0,\infty)$; the case $\mathbb{T} = \mathbb{N}$ is similar. □

Example 3.2 Let $\{X_t\}$ be an ergodic birth–death process on $E = \mathbb{N}$ with birth rates β_k and death rates δ_k and $F = \{0, 1, \ldots, K\}$. Then clearly $\{X_t^{(F)}\}$ is a birth–death process on F with the same birth–death rates as $\{X_t\}$ except possibly for the death rate, say $\widetilde{\delta}_K$ in state K. However, $\widetilde{\delta}_K$ can be identified with the exit intensity of $\{X_t^{(F)}\}$ from state K. Exit occurs if $\{X_t\}$ exits, which occurs with intensity $\beta_K + \delta_K$, and then goes to $K-1$ which occurs w.p. $\delta_K/(\beta_K + \delta_K)$. The exit intensity is the product δ_K, hence $\widetilde{\delta}_K = \delta_K$.

It follows by Proposition 3.1 that the stationary distribution for $\{X_t^{(F)}\}$ is obtained by conditioning the stationary distribution for $\{X_t\}$ to F, as was noted in III.2.6 by inspection of the explicit expressions. An example is the conditional geometric distribution in the $M/M/1$ queue with a waiting room of size $K < \infty$ and $\rho < 1$. If $\rho \geq 1$, the corresponding unrestricted $M/M/1$ queue is not ergodic and to proceed via Proposition 3.1, one needs to define the extension $\{X_t\}$ to $\mathbb{N}$ in a different way making $\{X_t\}$ ergodic (say by taking $\beta_n = 1/2$, $n \geq K$, and $\delta_n = 1$, $n > K$). □

Example 3.3 (THE FINITE DAM) In water storage problems, the capacity of a dam is apparently not infinite in practice as assumed in Section 1 but the content cannot exceed some finite level, say K. If more input at a jump occurs than the dam can contain, the excess water simply flows over instantaneously. An example of the sample paths of such a process $\{V_t^{(K)}\}$ (defined in terms of a release function $r(x)$, $0 < x \leq K$, and a compound Poisson input process with parameters β, B) is in Fig. 3.2.

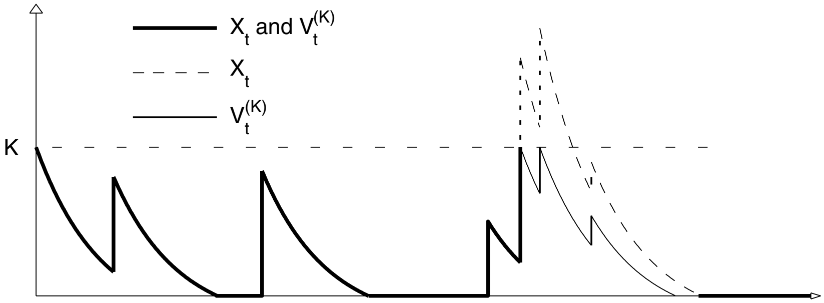

Figure 3.2

If we extend $r(\cdot)$ to (K,∞) in such a way that the infinite dam process $\{X_t\}$ is recurrent (say by letting $r(x)=2\beta\mu_B$, $x>K$), we have $\left\{V_t^{(K)}\right\} \overset{\mathscr{D}}{=} \left\{X_t^{(F)}\right\}$ where $F=[0,K]$ (but note that the sample path relation in Fig. 3.2 differs from the one in Fig. 3.1!). It follows by Proposition 3.1 and Theorem 1.4 that the stationary distribution $\pi^{(F)}$ of $\left\{V_t^{(K)}\right\}$ has an atom at 0 of size (say) $\pi_0^{(F)}>0$ if and only if $\theta(x;0)<\infty$ and that $\pi^{(F)}$ is absolutely continuous on $(0,K]$ with a density $g^{(F)}$ satisfying

$$g^{(F)}(x) \;=\; \frac{\beta}{r(x)}\left\{\pi_0^{(F)}\overline{B}(x)\;+\;\int_0^x \overline{B}(x-y)g^{(F)}(y)\,\mathrm{d}y\right\},\;\; 0<x\le K \quad (3.1)$$

(divide (1.8) by $\pi[0,K]$). Of course, these conclusions may also be obtained by copying the arguments for the infinite dam.

To exemplify important characteristics of the finite dam, consider the overflow rate

$$\beta\int_0^K \pi^{(F)}(\mathrm{d}x)\int_{K-x}^{\infty}(y-K+x)\,B(\mathrm{d}y). \qquad (3.2)$$

in the steady state. □

Example 3.4 (THE INFINITELY DEEP OR BOTTOMLESS DAM) Instead of approaching the finite dam via an infinitely high one, the suggestion has been made of using an approximation in terms of an infinitely deep or bottomless dam. This is reasonable in particular if overflow is a more predominant phenomenon than emptiness, as will be the case if the process has an upward drift (say $r(x)\equiv r$ constant and $\beta\mu_B>r$). The state space for the process, say $\{Y_t\}$, becomes $(-\infty,K]$ (equivalently, we may study the *deficit* or *depletion* process $\{K-Y_t\}$ with state space $[0,\infty)$). This model can be treated much as the infinitely high dam. For example, the derivation of (1.7) does not use the fact that $[0,x]$ is bounded to the left. Thus the steady–state density $g(x)$ is given as solution of the equation

$$g(x) \;=\; \frac{\beta}{r(x)}\int_{-\infty}^x \overline{B}(x-y)g(y)\,\mathrm{d}y,\quad -\infty<x\le K. \qquad (3.3)$$

A particularly important case is constant release $r(x) \equiv r$, in which case ergodicity is equivalent to $\beta\mu_B > r$ and it is seen by insertion that the solution of (3.3) is exponential, $g(x) = \eta e^{\eta(x-K)}$ where $\eta > 0$ is the unique solution of $\int_0^\infty e^{-\eta u} B(du) = 1 - \eta r/\beta$. □

As the second general set–up for finite capacity models, we consider a Lévy process or random walk $\{S_t\}_{t\in\mathbb{T}}$ and denote by $\{V_t^{(K)}\}_{t\in\mathbb{T}}$ the modification obtained by letting 0 and K perform as reflecting barriers. Thus, the state space is $[0,K]$ (or, in the lattice case, $\{0,1,\dots,K\}$, assuming w.l.o.g. that the lattice span is 1). Throughout, $\{V_t\}_{t\in\mathbb{T}}$ denotes $\{S_t\}$ one–sided reflected at 0 as in VII.2.

In discrete time, the process $\{V_n^{(K)}\}_{n\in\mathbb{N}}$ is given by the recursion

$$V_{n+1}^{(K)} \;=\; \bigl[V_n^{(K)} + Z_n\bigr]^+ \wedge K \tag{3.4}$$

and the initial value $V_0^{(K)} \in [0,K]$ where $\{Z_n\}$ is the sequence of random walk increments (i.e. $S_n = Z_1 + \cdots + Z_n$). The stationary distribution was found in IX.4.6 to be given by

$$\mathbb{P}_e\bigl(V_n^{(K)} \geq x\bigr) \;=\; \mathbb{P}(S_{\tau[x-K,x)} \geq x) \tag{3.5}$$

where $\tau[u,v) = \inf\{n \geq 0 : S_n \notin [u,v)\}$, $u \leq 0 \leq v$[1]. Thus, it remains to compute $\mathbb{P}(S_{\tau[u,v)} \geq v)$. The lattice case has a simple solution without restrictions on the increment distribution F. Indeed, in the lattice case we can, if necessary, truncate F to $\{0,\pm1,\dots,\pm K\}$ without changing $\mathbb{P}(S_{\tau[u,v)} \geq v)$, and get (cf. VIII.5.4 for a similar treatment of the one–boundary case):

Proposition 3.5 *Let $0 < r \leq K$, $0 < s \leq K$, and assume that F is concentrated on $\{-r,-r+1,\dots,s-1,s\}$ with $F\{-r\} > 0$, $F\{s\} > 0$. Let $0 < x \leq K$ and define $p_k = \mathbb{P}(S_{\tau[x-K,x)} = k)$, $k = x-K-1, x-K-2,\dots,x-K-r$ or $k = x,x+1,\dots,x+s-1$. Then the p_k are the solutions of the $r+s$ linear equations*

$$1 \;=\; \sum_{k=x-K-r}^{x-K-1} \alpha_j^k p_k \;+\; \sum_{k=x}^{x+s-1} \alpha_j^k p_k, \quad j = 1,\dots,r+s, \tag{3.6}$$

where $\alpha_1,\dots,\alpha_{r+s}$ are the roots of $1 = \widehat{F}[a] = \mathbb{E}a^{Z_1}$ or, equivalently the polynomial $a^r\bigl(\widehat{F}[a]-1\bigr)$. In particular, $\mathbb{P}_e\bigl(V_n^{(K)} \geq x\bigr) = p_x + \cdots + p_{x+s-1}$.

Proof. By optional stopping of the Wald martingales $\{\alpha_j^{S_n}\}_{n\in\mathbb{N}}$. □

Of course, in the lattice case one may alternatively compute $\nu_0,\dots,\nu_K$ by solving $\boldsymbol{\nu P} = \boldsymbol{\nu}$ where $\boldsymbol{P}$ is the transition matrix of $\{V_n^{(K)}\}$ (but note that K may be much larger than $r+s$).

[1] Note that $n = 0$ is included in the definition but only plays a role in the case $v = 0$ where we get $\tau[-K,0) = 0$, $S_{\tau[-K,0)} = 0$ and $\mathbb{P}_e(V_n^{(K)} \geq 0) = 1$, as should be).

Remark 3.6 Consider the lattice case with F downward skip–free (concentrated on $\{-1,0,1,2,\ldots\}$). Then the two–sided reflected process $\{V_n^{(K)}\}$ is the restriction of $\{V_n\}$ to $\{0,1,\ldots,K\}$, and we are back to Proposition 3.1. Similar remarks apply to the upward skipfree case by interchanging 0 and K. See also VIII.5b. □

In continuous time, the first problem is how to rigorously define $\{V_t^{(K)}\}_{t\geq 0} = \{V_t^{(K)}(y)\}_{t\geq 0}$ started from $y \in [0,K]$. For $y \neq K$, we do this by taking the segment up to the first hitting time $\tau(K)$ of K as the initial segment of $\{V_t(y)\}$ (the one–sided reflected process started from y) until $[K,\infty)$ is hit; we then let $V_{\tau(K)}^{(K)} = K$. For $y = K$, we similarly take the segment up to the first hitting time $\tau(0)$ of 0 by using the one–sided reflection operator (with the sign reversed and change of origin) as constructed in VII.2; at time $\tau(0)$ where this one–sided reflected (at K) process hits $(-\infty,0]$, we let $V_{\tau(0)}^{(K)} = 0$. The whole process $\{V_t^{(K)}\}$ is then constructed by glueing segments together in an obvious way. Glueing also local times together, we obtain a representation

$$V_t^{(K)} \;=\; y + S_t + L_t^{(0)} - L_t^{(K)} \tag{3.7}$$

where $\{L_t^{(0)}\}$, $\{L_t^{(K)}\}$ are the local times at 0, resp. K (both are nonnegative and nondecreasing and can increase only when $\{V_t^{(K)}\}$ is at the respective boundary).

Proposition 3.7 *The stationary distribution of the two–sided reflected Lévy process* $\{S_t\}_{t\geq 0}$ *is given by*

$$\mathbb{P}_e\bigl(V_t^{(K)} \geq x\bigr) \;=\; \mathbb{P}(S_{\tau[x-K,x)} \geq x) \tag{3.8}$$

where $\tau[u,v) = \inf\bigl\{t \geq 0 : S_t \notin [u,v)\bigr\}$, $u \leq 0 \leq v$.

Proof. Rather than attempting to extend the general duality machinery used in IX.4.6, we give a direct argument. Write $V_t^{(K)} = V_t^{(K)}(0)$, let T be fixed and let $\{R_t\} = \{R_t(x)\}$ be defined as $R_t = x - S_T + S_{T-t}$ until $(-\infty,0]$ or (K,∞) is hit; the value is then frozen at 0, resp. ∞. We shall show that

$$V_T^{(K)}(0) \geq x \iff R_T(x) = 0; \tag{3.9}$$

this yields

$$\mathbb{P}\bigl(V_T^{(K)} \geq x\bigr) \;=\; \mathbb{P}\bigl(\tau[x-K,x) \leq T,\, S_{\tau[x-K,x)} \geq x\bigr)$$

and the proposition then follows by letting $T \to \infty$. In the rest of the proof, write $V_t = V_t^{(K)}(0)$, $R_t = R_t^{(K)}(x)$.

Let $\sigma = \sup\{t \in [0,T] : V_t = 0\}$ (well–defined since $V_0 = 0$). Then $V_T = S_T - S_\sigma + L_\sigma^{(K)} - L_T^{(K)}$ by (3.7), so if $V_T \geq x$ then $S_T - S_\sigma \geq x$, and similarly, for $t \geq \sigma$

$$x \leq V_T = V_t + S_T - S_t + L_t^{(K)} - L_T^{(K)} \leq K + S_T - S_t,$$

implying $R_{T-t} \leq K$. Thus absorbtion of $\{R_t\}$ at ∞ is not possible before $T-\sigma$, and $S_T - S_\sigma \geq x$ then yields $R_{T-\sigma} = 0$ and $R_T = 0$.

Assume conversely $R_T = 0$ and write the time of absorbtion in 0 as $T-\sigma$. Then $x - S_T + S_\sigma \leq 0$, and $R_t \leq K$ for all $t \leq T-\sigma$ implies $x - S_T + S_t \leq K$ for all $t \geq \sigma$. If $V_t < K$ for all $t \in [\sigma, T]$, then $L_T^{(K)} - L_t^{(K)} = 0$ for all such t and hence

$$V_T = V_\sigma + S_T - S_\sigma + L_T^{(0)} - L_\sigma^{(0)} \geq V_\sigma + S_T - S_\sigma \geq 0 + x.$$

If $V_t = K$ for some $t \in [\sigma, T]$, denote by ω the last such t. Then $L_T^{(K)} = L_\omega^{(K)}$ and hence

$$V_T = V_\omega + S_T - S_\omega + L_T^{(0)} - L_\omega^{(0)} \geq K + S_T - S_\omega + 0 \geq x. \qquad \square$$

Example 3.8 Consider Brownian motion $\{S_t\}$ with drift μ and unit variance and the two–sided reflected version $\{V_t^{(K)}\}$ on $[0, K]$. Optional stopping of the Wald martingale $\{e^{-2\mu S_t}\}$ gives in a straightforward way

$$\mathbb{P}(S_{\tau[x-K,x)} \geq x) = \frac{e^{2\mu K} - e^{2\mu x}}{e^{2\mu K} - 1}.$$

Differentiation shows that the stationary density of $\{V_t^{(K)}\}$ is proportional to $e^{2\mu x}$. This was found earlier in IX.3.6 and when $\mu < 0$, it is compatible via Proposition 3.1 with the intuitive picture of $\{V_t^{(K)}\}$ as the restriction to $[0, K]$ of $\{V_t\}$ = Brownian motion one-sided reflected at 0 (the process obtained from $\{V_t\}$ by cutting out excursions in (K, ∞)). □

Example 3.9 Let $\{S_t\}$ be the netput process of a fluid model with background Markov process $\{J_t\}$; cf. XI.1b. An obvious extension of the proof of Proposition 3.7 shows that

$$\mathbb{P}_e\big(V_t^{(K)} \geq x\big) = \mathbb{P}_\pi\big(S^*_{\tau^*[x-K,x)} \geq x\big) \tag{3.10}$$

where π is the stationary distribution of $\{J_t\}$ and $*$ refers to the time–reversed version; cf. XI.2e. The evaluation of the r.h.s. of (3.10) is straightforward using the Wald martingales from XI.2c and has been carried out in the proof of XI.4.2. □

Example 3.10 In the case of the $M/G/1$ workload, $\{S_t\}$ is compound Poisson with unit downward drift. Denote the intensity by β and the jump size (service time) distribution by B (B is concentrated on $(0, \infty)$). The restricted workload process $\{V_t^{(K)}\}$ is just the finit dam in Example 3.3. When the service time distribution B is phase–type, $\mathbb{P}(S_{\tau[x-K,x)} \geq x)$ (and thereby $\mathbb{P}_e\big(V_t^{(K)} \geq x\big)$) may be computed applying XI.4.2 once more, now to the fluid representation of the $M/PH/1$ workload process given in XI.1c. □

Remark 3.11 When dealing with finite buffer problems, an important characteristic is the probability $\mathbb{P}_e\big(V_t^{(K)} = K\big)$ of a full buffer (in many cases, this can also be interpreted as a constant times a loss probability). Heuristically, it has sometimes been suggested to use the tail probability $\mathbb{P}_e(V_t \geq K)$ for the infinite buffer model as approximation. Whether this is justified depends very much upon the precise model. In the setting of reflected random walks or Lévy processes, care is definitely needed since $\mathbb{P}_e\big(V_t^{(K)} = K\big)$ is often 0 (say for Brownian motion or the $M/M/1$ workload) but $\mathbb{P}_e(V_t \geq K)$ not.

For a discrete–time random walk, $\mathbb{P}_e\big(V_t^{(K)} = K\big) = \mathbb{P}(S_{\tau[0,K)} \geq K)$ reduces simply to the probability that the cycle maximum exceeds K, and in the light–tailed case, XIII.5.2 and XIII.5.9 show that (in the notation there)

$$\mathbb{P}_e\big(V_t^{(K)} = K\big) \;\sim\; \mathbb{P}_L(\tau_- = \infty)\mathbb{P}_e(V_t \geq K) \;\sim\; C\mathbb{P}_L(\tau_- = \infty)\mathrm{e}^{-\gamma K} \tag{3.11}$$

as $K \to \infty$. In the heavy–tailed case, X.9.1 and X.9.4 yield

$$\mathbb{P}_e(V_t \geq K) \sim \frac{1}{|\mu_F|}\int_K^\infty \overline{F}(x)\,\mathrm{d}x\,, \quad \mathbb{P}_e\big(V_t^{(K)} = K\big) \;\sim\; \mathbb{E}\sigma\cdot\overline{F}(K)\,,$$

which in particular shows that $\mathbb{P}_e\big(V_t^{(K)} = K\big)$ is negligible compared to $\mathbb{P}_e(V_t \geq K)$. □

We will finally study an example that is slightly more complicated than reflected random walks:

Example 3.12 (MORAN'S MODEL FOR THE DAM) This is a historically important discrete Markov chain model. The input sequence $\{A_n\}$ is assumed i.i.d. and the release is constant, say b per time unit (if the content just before the release is $c < b$, only the amount c is released), and we let K denote the capacity of the dam. We will consider a slightly more general model where also the release at time n is random, say B_n rather than b (the sequence $\{B_n\}$ is assumed i.i.d. and independent of $\{A_n\}$).

We let Q_n^A denote the content just before the nth input (just after the $(n-1)$th release) and Q_n^B the content just after (just before the $(n+1)$th release). Then

$$\begin{aligned} Q_n^A &= \big[Q_{n-1}^B - B_{n-1}\big]^+, & (3.12)\\ Q_n^B &= \big(Q_n^A + A_n\big)\wedge K, & (3.13)\\ Q_n^A &= \big[(Q_{n-1}^A + A_{n-1})\wedge K - B_{n-1}\big]^+, & (3.14)\\ Q_n^B &= \big([Q_{n-1}^B - B_{n-1}]^+ + A_n\big)\wedge K. & (3.15) \end{aligned}$$

The recursions (3.14), (3.15) can be studied by much the same methods as used for (3.4). Consider e.g. (3.15) which can be written as $Q_n^B = f(Q_{n-1}^B, \boldsymbol{U}_{n-1})$ where $\boldsymbol{u} = (a,b)$, $f(x,\boldsymbol{u}) = ([x-b]^+ + a)\wedge K$ and

$\boldsymbol{U}_{n-1} = (A_n, B_{n-1})$. The inverse function g of f in the sense of IX.4.3 is then given by

$$g(x,a,b) \;=\; \begin{cases} 0 & x=0 \text{ or } x\in(0,K],\ a\geq K \\ x-(a-b) & x\in(0,K],\ a<K \\ \infty & x>K \end{cases} .$$

It follows that the dual process $\{R_n\}$ started from x evolves as the unrestricted random walk $\{(x-A_0)^+ - S_n\}$, starting from $(x-A_0)^+$ and having random walk increments $Z_n = A_n - B_{n-1}$, and that $\mathbb{P}_e(Q_n^B \geq x)$ is the probability that this process will exit $(0,K]$ to the right. □

Notes The literature on finite buffer problems is large but not always systematic. See e.g. Lindley (1959), Siegmund (1976a), Borovkov (1984), Stadje (1993), Jelenkovic (1999) and Kim and Shroff (2001).

For Skorokhod problems with two barriers, see Chen and Mandelbaum (1990). The last part of Example 3.12 is from Asmussen and Sigman (1996).

4 Some Simple Inventory Models

Example 4.1 (THE EOQ MODEL) In classical deterministic formulation of the EOQ (Economic Order Quantity) model, the inventory decreases (is sold) at a constant rate. When the inventory level reaches 0, a new batch of size Q is ordered. Thus the inventory level evolves deterministically as in Fig. 4.1(a).

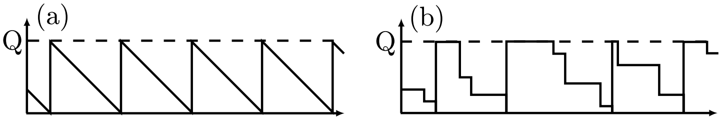

Figure 4.1

Each order has an overhead cost of a and the cost of holding inventory level x is bx per unit time. The problem is to choose Q so as to minimize the overall costs. But the rate of orders is $1/Q$ and the average inventory level $Q/2$ so that the overall cost rate is $a/Q + bQ/2$, which is a convex function with limits ∞ at $Q=0$ and $Q=\infty$. Thus there is a unique minimizer Q^*, that is easily found to be $Q^* = \sqrt{2a/b}$.

There are several stochastic versions of this model. Assume, for example, that the inventory level between orders decreases according to a compound Poisson process with Poisson rate β and downward exponential jumps with intensity δ; cf. Fig. 4.1(b). If X_t is the inventory level at time t, $\{X_t\}$ is regenerative (the cycle is the time τ between two orders) and hence $X_t \xrightarrow{\mathscr{D}} X$ so that the optimization problem becomes to minimize $a/\mathbb{E}_Q\tau + b\mathbb{E}_Q X$ w.r.t. Q. □

Example 4.2 (THE (s, S) INVENTORY MODEL) This model is given by two parameters $0 < s < S$, such that s is the inventory level at which reordering is done and S is the level to which one aims at resetting the inventory level at each order. Again, there are several possible stochastic models. For example, in discrete time, assume that the amounts of stock sold are i.i.d. r.v.'s $\{R_k\}$ with common distribution F and that it is possible to do the reordering so as to achieve exactly level S. If X_n denotes the inventory level after the nth sale, we therefore have

$$X_{n+1} \;=\; \begin{cases} X_n - R_{n+1} & \text{if } s \le X_n \le S \\ S - R_{n+1} & \text{if } -\infty < X_n < s \end{cases} .$$

The stationary distribution π is easily obtained by observing that the epochs n with $X_{n-1} < s$ are regeneration points. Thus the zero–delayed case corresponds to $X_0 = S - R_0$, the cycle length is

$$C \;=\; \inf\{n \ge 1 : R_0 + \cdots + R_{n-1} > S - s\}$$

and within the cycle the points in $[s, S)$ which are visited are of the form $S - y$ with $y \in (0, S - s)$ an epoch of a renewal process governed by F. Similarly, exactly one visit to $(-\infty, s)$ occurs at $s - z$ where z is the overshoot B_{S-s} of $S - s$. With $U = \sum_0^\infty F^{*n}$ the renewal function and $H(z) = \mathbb{P}(B_{S-s} \le z)$, we therefore have $\mathbb{E}C = U(S - s)$ and

$$\pi(\mathrm{d}y) \;=\; \begin{cases} \dfrac{U(S-\mathrm{d}y)}{U(S-s)} & \text{if } s \le y < S, \\[2ex] \dfrac{H(s-\mathrm{d}y)}{U(S-s)} & \text{if } -\infty < y < s. \end{cases} \qquad \square$$

Example 4.3 Many inventory problems involve in a crucial way the concept of *lead time* Z, defined as the delay between the placement and the actual receipt of an order. As a simple example, consider a variant of the (s, S) model, where the inventory X_t at time t is measured in discrete units $0, 1, 2, \ldots$ and orders occur according to a Poisson process with intensity δ and are handled at once provided $X_t > 0$ (orders received when $X_t = 0$ are lost). When the inventory level goes from $s + 1$ to s, an reorder of (deterministic) size $S - s$ is placed and received Z units of time later; here the lead times $Z_1, Z_2, \ldots$ are i.i.d., say with common distribution F concentrated on $(0, \infty)$. We shall compute the stationary probabilities $\pi_j = \mathbb{P}_e(X_t = j)$ and the rate (in the Palm sense) λ of reorderings. A typical application would again be to choose s, S in an optimal way; if say there is a cost of rate c_j of holding inventory level j (in particular, c_0 can be interpreted as a penalty for lost orders) and a cost d_S of ordering a batch of size S, the overall cost rate is

$$\sum_{j=0}^{S} c_j \pi_j \;+\; \lambda d_S.$$

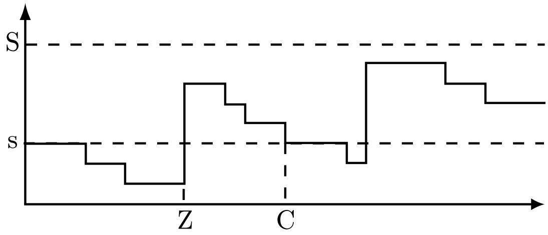

Figure 4.2

The state space for $\{X_t\}$ is $\{0, 1, \ldots, S\}$. We can view the times of downcrossing to level s as regeneration points; see Fig. 4.2 for the zero–delayed case (C is the regenerative cycle). We write p_k for the probability that the inventory level is $k = 0, \ldots, s$ just before an reorder is received. Since precisely one reorder is received per cycle, it is then immediate that (assuming $S - s > s$ for simplicity)

$$\frac{1}{\lambda} \;=\; \mathbb{E}C \;=\; \mathbb{E}Z + \mathbb{E}(C - Z) \;=\; \mu_F + \frac{1}{\delta}\sum_{k=0}^{s} p_k(S - s + k)\,.$$

Also, clearly

$$p_k \;=\; \mathbb{P}(X_{Z-} = k) \;=\; \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^{s-k}}{(s-k)!}\,F(\mathrm{d}t), \quad k = s, s-1, \ldots, 1,$$

and $p_0 = 1 - p_1 - \cdots - p_s$. Define $r_j \;=\; \int_0^C I(X_t = j)\,\mathrm{d}t$, $j = 0, 1, \ldots, S$. Then $\pi_j = r_j/\mathbb{E}C$ and

$$r_j \;=\; \begin{cases} \displaystyle\frac{1}{\delta}\sum_{k=0}^{s} p_k I(k + S - s \geq j) \;=\; \frac{1}{\delta}\sum_{k=(j-S+s)^+}^{s} p_k & s < j \leq S \\ \displaystyle\int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^{s-j}}{(s-j)!}\overline{F}(t)\,\mathrm{d}t & 0 < j \leq s \end{cases}.$$

Finally, π_0 can either be computed as $1 - \pi_1 - \cdots - \pi_S$ or from

$$r_0 \;=\; \int_0^\infty \sum_{k=s}^{\infty} \mathrm{e}^{-\delta t}\frac{(\delta t)^k}{k!}\overline{F}(t)\,\mathrm{d}t. \qquad \Box$$

Problems

4.1 Compute $\mathbb{E}_Q\tau$ and $\mathbb{E}_Q X$ in Example 4.1.
4.2 Use partial integration to obtain expressions for the r_j terms of the p_k in Example 4.3.

Notes A classical source for inventory models is Arrow *et al.* (1958). More recent treatments are in Tijms (1994), Silver *et al.* (1998), Axsäter (2000), Zipkin (2000) and Nahmias (2001).

5 Dual Insurance Risk Models

The main objects of study in insurance risk are the ruin probabilities

$$\psi(u) \;=\; \mathbb{P}(\tau(u) < \infty), \quad \psi(u,T) \;=\; \mathbb{P}(\tau(u) \le T) \tag{5.1}$$

with infinite, resp. finite horizon, where $\tau(u) = \inf\{t > 0 : \, R_t < 0 \,|\, R_0 = u\}$ is the time to ruin and $\{R_t\}_{t\ge 0}$ is a model for the reserve of the insurance company.

The traditional Cramér–Lundberg model has already been introduced in V.2.3. It is compound Poisson,

$$R_t \;=\; u + rt - \sum_{i=1}^{N_t} U_i .$$

Here r is the rate of premium inflow, $\{N_t\}$ is a Poisson process with rate β and the U_i form an independent i.i.d. sequence with common distribution B ($U_i > 0$ represents the size of the ith claim made to the company). The time change $t \to t/r$ changes r to 1 and $\psi_{r,\beta}(u,T)$ to $\psi_{1,\beta/r}(u,T/r)$ whereas $\psi_{r,\beta}(u) = \psi_{1,\beta/r}(u)$. Therefore it is no restriction to assume $r = 1$. Letting $S_t = u - R_t$ (the so–called *claim surplus process*), we then have $S_t = \sum_1^{N_t} U_i - t$. This is just the netput process generating the $M/G/1$ workload process $\{V_t\}$, so the maximum representation in III.7 (see also IX.2) yields:

Corollary 5.1 *The workload process $\{V_t\}$ of an initially empty ($V_0 = 0$) $M/G/1$ queue with $\rho < 1$ and the ruin probabilities of the Cramér–Lundberg risk model with the same β, B and $r = 1$ are connected by*

$$\mathbb{P}_0(V_T > u) \;=\; \mathbb{P}\Big(\sup_{0\le t\le T} S_t > u\Big) \;=\; \mathbb{P}(\tau(u) \le T) \;=\; \psi(u,T), \tag{5.2}$$

$$\mathbb{P}_e(V_T > u) \;=\; \mathbb{P}\Big(\sup_{0\le t<\infty} S_t > u\Big) \;=\; \mathbb{P}(\tau(u) < \infty) \;=\; \psi(u) \tag{5.3}$$

where $\tau(u) \;=\; \inf\{t > 0 : \, R_t < 0\} \;=\; \inf\{t > 0 : \, S_t > u\}$ is the time to ruin.

Results similar to Corollary 5.1 connect a number of other queueing and insurance risk models. In particular:

Corollary 5.2 *Let $\{W_n\}$, $\{V_t\}$ be the waiting time sequence, resp. workload process of a $GI/G/1$ queue and $\{R_t\}$ the process obtained by replacing the Poisson process for the Cramér–Lundberg model with the stationary renewal counting process for the $GI/G/1$ queue and the claims $U_1, U_2, \ldots$ by the service times. Then (5.2), (5.3) continue to hold. Further, if $\{S_t^{\#}\}$, $\tau^{\#}(u)$, $\psi^{\#}(u)$ refer to $\{N_t\}$ being zero–delayed with the atom at 0 removed, then*

$$\mathbb{P}_e(W_n > u) = \mathbb{P}\Big(\sup_{0\le t<\infty} S_t^{\#} > u\Big) \;=\; \mathbb{P}(\tau^{\#}(u) < \infty) \;=\; \psi^{\#}(u). \tag{5.4}$$

Proof. Use Loynes's lemma in continuous time for the first statement and the discrete time version for the second. □

Corollary 5.3 *Let* $\{J_t\}$ *be the background Markov process of a Markov additive process* (cf. XI.2), $\{S_t\}$ *the additive part,* $\{V_t\}$ *the reflected process,* $\{(J_t^*, S_t^*)\}$ *the time–reversed version and* $\tau^*(u) = \inf\{t > 0 : S_t^* > u\}$, $\psi_i^*(u) = \mathbb{P}_i(\tau^*(u) < \infty)$. *If* $\boldsymbol{\pi}$ *is the stationary distribution for* $\{J_t\}$, *then* $\mathbb{P}_e(J_t = i, V_t > u) = \pi_i \psi_i^*(u)$.

Proof. Immediate from XI.2.11. □

Corollary 5.1 can also be extended to risk models developing as the Cramér–Lundberg model except that the premium rate $r(x)$ depends on the current level x of the reserve R_t. That is,

$$R_t \;=\; u - \sum_{i=1}^{N_t} U_i + \int_0^t r(R_s)\,\mathrm{d}s \tag{5.5}$$

where $u \geq 0$. Note that for the study of the ruin probabilities the values of $r(x)$ for $x \leq 0$ are immaterial.

Proposition 5.4 *Let* $\{V_t\}$ *be an initially empty* ($V_0 = 0$) *storage process given in terms of* β, B *and* $r(x)$, $x > 0$, *as in Section 1 and let* $\{R_t\}$ *be the process* (5.5) *with* r *extended to* $(-\infty, 0)$ *in some arbitrary manner. Then*

$$\mathbb{P}_0(V_T > u) \;=\; \mathbb{P}(\tau(u) \leq T) \;=\; \psi(u,T), \tag{5.6}$$

$$\mathbb{P}_e(V_T > u) \;=\; \mathbb{P}(\tau(u) < \infty) \;=\; \psi(u). \tag{5.7}$$

Proof. Since $\{V_t\}$ is clearly stochastically monotone, we can define $\{\widetilde{R}_t\}$ as the Siegmund dual process of $\{V_t\}$, and the theory of IX.4 then shows that it suffices to verify that $\{R_t\}$ and $\{\widetilde{R}_t\}$ evolve in the same way on $(0,\infty)$, or in turn that $\mathscr{B}g = \widetilde{\mathscr{B}}g$ where $\mathscr{A}$, $\mathscr{B}$, $\widetilde{\mathscr{B}}$ are the generators of $\{V_t\}$, $\{R_t\}$, resp. $\{\widetilde{R}_t\}$, and $g \in \mathscr{K}$, the class of C^1 functions with compact support contained in $(0,\infty)$.

For $g \in \mathscr{K}$, we have up to $\mathrm{o}(h)$ terms that

$$\begin{aligned}\mathbb{E}_x g(R_h) \;&=\; g\big(x + hr(x)\big)(1-\beta h) + \beta h \int g(x-z)\,B(\mathrm{d}z)\\ &=\; g(x) \;+\; h\left\{g'(x)r(x) - \beta g(x) + \beta\int g(x-z)\,B(\mathrm{d}z)\right\}\end{aligned}$$

so that $\mathscr{B}g$ is given by $\{\cdot\}$. Similarly, for $f \in \mathscr{K}$

$$\mathscr{A}f(y) \;=\; -f'(y)r(y) - \beta f(y) + \beta\int f(y+z)\,B(\mathrm{d}z)\,.$$

Also, replacing y by $y = y - z$ and integrating by parts yields

$$\int f(y+z)g'(y)\,\mathrm{d}y \;=\; -\int f'(y)g(y-z)\,\mathrm{d}y$$

(using that f, g have compact support). It follows that (see IX.4.7 for the first identity)

$$\begin{aligned}
\int f'(x)\widetilde{\mathscr{B}}g(x)\,\mathrm{d}x &= -\int g'(y)\mathscr{A}f(y)\,\mathrm{d}y \\
&= -\int g'(y)\Big\{-f'(y)r(y) - \beta f(y) + \beta\int f(y+z)\,B(\mathrm{d}z)\Big\}\,\mathrm{d}y \\
&= \int f'(y)\Big\{g'(y)r(y) - \beta g(y) + \beta\int g(y-z)\,B(\mathrm{d}z)\Big\}\,\mathrm{d}y \\
&= \int f'(y)\mathscr{B}g(y)\,\mathrm{d}y\,.
\end{aligned}$$

The truth of this for all f under consideration implies $\widetilde{\mathscr{B}}g = \mathscr{B}g$. □

Example 5.5 A two–step premium rule

$$r(x) \;=\; \begin{cases} r_1 & x \le x_0 \\ r_2 & x > x_0 \end{cases}$$

with $r_1 > r_2$ could arise by modifying the Cramér–Lundberg model such that the company raises the premium once the reserve goes below a threshold x_0, and/or that it lowers the premium if the reserve goes above x_0, say for attracting more customers, and/or that it pays out dividend at a constant rate if the reserve is above x_0 (then r_2 has to be interpreted as the premium rate minus the dividend rate). □

Example 5.6 If in the Cramér–Lundberg model with premium rate r the company invests the reserve at interest rate δ, we obtain a model of the type (5.5) with $p(x) = r + \delta x$. □

Notes Standard textbooks on insurance risk are (in chronological order) Gerber (1979), Grandell (1990), Daykin *et al.* (1994), Rolski *et al.* (1999) and Asmussen (2000); broader expositions covering also other parts of insurance mathematics are Bühlmann (1970), Bowers *et al.* (1986), Sundt (1993) and Sundt and Teugels (2004). Besides general applied probability journals, much of the research literature is published in *Astin Bulletin*, *Insurance: Mathematics and Economics* and the *Scandinavian Actuarial Journal*.

Proposition 5.4 can alternatively be proved by a sample path argument much along the lines of Proposition 3.7, see Asmussen (2000), pp. 30–32. Yet a third approach is in Harrison and Resnick (1977).

6 The Time to Ruin

We consider the Cramér–Lundberg risk model with $\{R_t\}$, $\{S_t\}$, $\tau(u)$, $\psi(u)$, $\psi(u,T)$, etc. as in Section 5 and $r = 1$, $\rho < 1$.

A number of results on the $M/G/1$ queue translate in a straightforward way via Corollary 5.1 to results on the risk process and the ruin probabilities. For example, for a light–tailed B we have the Cramér–Lundberg approximation $\psi(u) \sim C\mathrm{e}^{-\gamma u}$, $u \to \infty$. See V.7.8 and XIII.5 for the precise conditions and values of C, γ (actually, the result originates from insurance risk and not the random walk/queueing setting considered in XIII.5!). Similarly, for a heavy–tailed B X.9.1 yields $\psi(u) \sim \rho(1-\rho)^{-1}\overline{B}_0(u)$ (subject to minor regularity conditions stated there) where $\overline{B}_0$ is the stationary excess distribution from renewal theory.

However, each of the areas of queueing and risk also has its own particular features lacking analogues in the other. For example, studies of other queue disciplines than FIFO lack interpretation in insurance risk, and in the present section we shall exemplify some of the specifics of insurance risk by undertaking a closer study of the time $\tau(u)$ to ruin which has no straightforward sample path interpretation for the $M/G/1$ queue despite the duality connection in Corollary 5.1.

We start by a detailed description of the case $u = 0$. Write $\mathbb{P}^{(0)} = \mathbb{P}(\cdot \,|\, \tau(0) < \infty)$, and recall that $\mathbb{P}(\tau(0) < \infty) = \rho$ and that $Y = S_{\tau(0)}$ has $\mathbb{P}^{(0)}$–distribution B_0 (compare e.g. Corollary 5.1 and VIII.5.7). Let K_t be the event that $\{R_t\}$ starting from $R_0 = 0$ is at a maximum at time t ($R_t > R_s$, $s < t$) and $\omega(z) = \inf\{t > 0 : R_t = z \,|\, R_0 = 0\}$; note that $\omega(z) < \infty$ a.s. since $\rho < 1$ (so that $R_t \overset{\text{a.s.}}{\to} \infty$) and $\{R_t\}$ is upward skip–free.

Lemma 6.1 *For any $T < \infty$ and any measurable $G \subseteq D[0, T)$,*

$$\mathbb{P}\left(\{S_{T-t-} - S_T\}_{t<T} \in G;\, \tau(0) \ge T\right) \;=\; \mathbb{P}_0\left(\{R_t\}_{t<T} \in G;\, K_T\right) .$$

Proof. We use the coupling $R_t = S_{T-t-} - S_T$ illustrated in Fig. 6.1. Then $\tau(0) \ge T$, i.e. $S_t < 0$ for $0 < t < T$, is equivalent to $R_{s-} < -S_T = R_T$ for all $s < T$ which in turn is the same as $R_s < R_T$ for all $s < T$, i.e. that K_T occurs. □

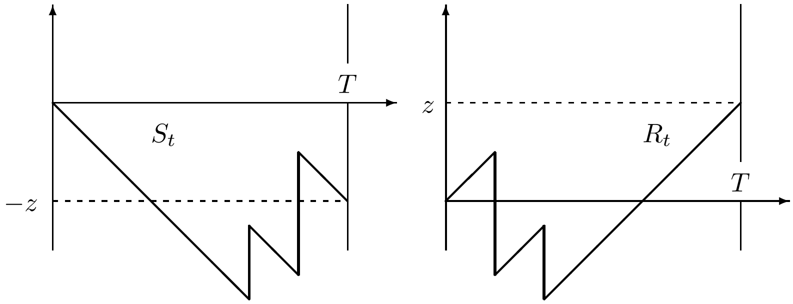

Figure 6.1

Theorem 6.2 (a) *The $\mathbb{P}^{(0)}$-distribution of $Z = -S_{\tau(0)-}$ is the stationary excess distribution B_0.*

(b) *The conditional distribution of* $\{S_{\tau(0)-t-}-z\}_{0\le t\le\tau(0)}$[2] *given* $\tau(0)<\infty$, $Z=z$ *is the same as the unconditional* $\mathbb{P}_0$*–distribution of* $\{R_t\}_{0\le t\le\omega(z)}$. *Further,* $S_{\tau(0)}$ *and* $\{S_{\tau(0)-t-}-z\}_{0\le t\le\tau(0)}$ *are conditionally independent given* $Z=z$.

Proof. Clearly,

$$\mathbb{P}(Z\in \mathrm{d}z,\tau(0)<\infty) \;=\; \int_0^\infty \mathbb{P}\bigl(\tau(0)\ge T,-S_{T-}\in\mathrm{d}z\bigr)\beta\overline{B}(z)\,\mathrm{d}T.$$

Using first Lemma 6.1 with $G=\{-S_{T-}\in\mathrm{d}z\}$ shows that the r.h.s. is $\beta\overline{B}(z)\int_0^\infty \mathbb{P}(R_{T-}\in\mathrm{d}z,K_T)\,\mathrm{d}T$. This reduces to $\beta\overline{B}(z)\,\mathrm{d}z$ since the upward movement of $\{R_t\}$ is linear at unit rate so that the expected time $\{R_t\}$ spends in a set $A\subseteq(0,\infty)$ and is at a maximum at the same time is simply the Lebesgue measure of A (take $A=\{\mathrm{d}T\}$). Since $\mathbb{P}(\tau(0)<\infty)=\rho=\beta\mu_B$, we get

$$\mathbb{P}^{(0)}(Z\in\mathrm{d}z) \;=\; (\beta\mu_B)^{-1}\mathbb{P}(Z\in\mathrm{d}z,\tau(0)<\infty) \;=\; \mu_B^{-1}\overline{B}(z)\,\mathrm{d}z \;=\; B_0(\mathrm{d}z),$$

proving (a). Part (b) is an immediate consequence of Lemma 6.1. □

As first application of Theorem 6.2, we will derive the double m.g.f. (Laplace transform) of the ruin time $\tau(u)$ and the single transform of $\tau(0)$. Let $\kappa(\alpha)=\beta(\widehat{B}[\alpha]-1)-\alpha$ be the Lévy exponent of $\{S_t\}$ and $r(a)$ the smallest of the two solutions to

$$-a \;=\; \kappa(r(a)) \;=\; \beta\bigl(\widehat{B}[r(a)]-1\bigr)-r(a)\,. \tag{6.1}$$

Theorem 6.3 (a) $\mathbb{E}\bigl[\mathrm{e}^{a\tau(0)};\,\tau(0)<\infty\bigr] \;=\; 1-\dfrac{a}{r(a)}$;

(b) $\displaystyle\int_0^\infty \mathrm{e}^{bu}\mathbb{E}\bigl[\mathrm{e}^{a\tau(u)};\,\tau(u)<\infty\bigr]\,\mathrm{d}u \;=\; \frac{-a/r(a)-\kappa(b)/b}{\kappa(b)+a}$.

Proof. Let $g_a(x)$ be the density of the measure

$$\mathbb{E}\bigl[\mathrm{e}^{a\tau(0)};\,\tau(0)<\infty,\,Y=S_{\tau(0)}\in\mathrm{d}x\bigr]$$

and write $\widehat{g}_a[b]=\int_0^\infty \mathrm{e}^{bx}g_a(x)\,\mathrm{d}x$. Optional stopping of the Wald martingale $\bigl\{\mathrm{e}^{r(a)S_t-t\kappa(r(a))}\bigr\}=\bigl\{\mathrm{e}^{r(a)S_t+at}\bigr\}$ yields $1=\mathrm{e}^{-zr(a)}\mathbb{E}\mathrm{e}^{a\omega(z)}$. Using Theorem 6.2(b) we obtain

$$\mathbb{E}\bigl[\mathrm{e}^{a\tau(0)}\mid\tau(0)<\infty,Z=z\bigr] \;=\; \mathbb{E}\mathrm{e}^{a\omega(z)}=\mathrm{e}^{zr(a)}.$$

Since clearly $\mathbb{P}^{(0)}(Y\in\mathrm{d}y\,|\,Z=z) \;=\; \overline{B}(\mathrm{d}y+z)/\overline{B}(z)$, conditioning upon $Z=z$ and using the conditional independence yields

$$\begin{aligned}\widehat{g}_a[b] \;&=\; \mathbb{E}\bigl[\mathrm{e}^{a\tau(0)+bY};\,\tau(0)<\infty\bigr]\\ &=\; \int_0^\infty \beta\overline{B}(z)\,\mathrm{d}z\,\mathrm{e}^{zr(a)}\frac{1}{\overline{B}(z)}\int_z^\infty \mathrm{e}^{b(y-z)}\,B(\mathrm{d}y)\end{aligned}$$

[2] Viewed as element of D_0, the D–functions with finite lifelength; cf. A2.

$$
\begin{aligned}
&= \beta\int_0^\infty \mathrm{e}^{by}\,B(\mathrm{d}y)\int_0^y \mathrm{e}^{(r(a)-b)z}\,\mathrm{d}z \;=\; \beta\int_0^\infty \mathrm{e}^{by}\,B(\mathrm{d}y)\frac{\mathrm{e}^{y(r(a)-b)}-1}{r(a)-b}\\
&= \frac{\beta\bigl(\widehat{B}[r(a)]-\widehat{B}[b]\bigr)}{r(a)-b} \;=\; \frac{\beta+r(a)-a-\beta\widehat{B}[b]}{r(a)-b}\\
&= \frac{r(a)-a-b-\kappa(b)}{r(a)-b}\,. \qquad (6.2)
\end{aligned}
$$

For part (a), now just take $b=0$.

For (b), define $Z_a(u)=\mathbb{E}[\mathrm{e}^{a\tau(u)};\tau(u)<\infty)$. It is then easily seen that $Z_a(u)$ is the solution of the renewal equation $Z_a(u)=z_a(u)+\int_0^u Z_a(u-x)g_a(x)\,\mathrm{d}x$ where $z_a(u)=\int_u^\infty g_a(x)\mathrm{d}x$. Hence

$$
\begin{aligned}
&\int_0^\infty \mathrm{e}^{bu}\,\mathrm{d}u\,\mathbb{E}[\mathrm{e}^{a\tau(u)};\tau(0)<\infty]\\
&\qquad=\; \widehat{Z}_a[b] \;=\; \frac{\widehat{z}_a[b]}{1-\widehat{g}_a[b]} \;=\; \frac{\bigl(\widehat{g}_a[b]-\widehat{g}_a[0]\bigr)/b}{1-\widehat{g}_a[b]}.
\end{aligned}
$$

Using (6.2), the result follows after simple algebra. □

We next turn to the question of what $\tau(u)$ looks like in the conditional distribution $\mathbb{P}^{(u)}=\mathbb{P}(\cdot\,|\,\tau(u)<\infty)$ given $\tau(u)<\infty$. The insight provided by Theorem 6.3 is quite limited, and we shall instead look into asymptotics as $u\to\infty$.

The light–tailed case is essentially settled by the analysis in XIII.5. In fact, it follows exactly as there that when the solution $\gamma>0$ exists and satisfies $\kappa'(\gamma)<\infty$ (i.e. $\widehat{B}'[\gamma]<\infty$), then $\tau(u)/u\overset{\mathbb{P}^{(u)}}{\to}1/\kappa'(\gamma)$. If in addition $\widehat{B}''[\gamma]<\infty<\infty$, then the $\mathbb{P}^{(u)}$–distribution of $\tau(u)$ is asymptotically normal with mean $u/\kappa'(\gamma)$ and variance $u\kappa''(\gamma)/\kappa'(\gamma)^3$. Hence the mean dominates the standard deviation so that in summary, given $\tau(u)<\infty$, $\tau(u)/u$ becomes more and more concentrated around $1/\kappa'(\gamma)$ as $u\to\infty$.

In the heavy–tailed case, let $U^{(u)}$, $U_0^{(u)}$ denote r.v.'s having the distributions $B^{(u)},B_0^{(u)}$of the overshoot over u corresponding to B, resp. B_0,

$$
\begin{aligned}
\overline{B}^{(u)}(x) &= \mathbb{P}(U^{(u)}>x) = \frac{\overline{B}(x+u)}{\overline{B}(u)},\\
\overline{B}_0^{(u)}(x) &= \mathbb{P}\bigl(U_0^{(u)}>x\bigr) = \frac{\int_{x+u}^\infty \overline{B}(y)\,\mathrm{d}y}{\int_u^\infty \overline{B}(y)\,\mathrm{d}y}\,.
\end{aligned}
$$

Asymptotically, the distribution of $U^{(u)},U_0^{(u)}$ typically reduces (the proof is elementary analysis and omitted):

Proposition 6.4 (a) *If B is regularly varying with tail $L(x)/x^\alpha$ for some $\alpha>1$ and some slowly varying $L(\cdot)$, then $\overline{B}_0(x)\sim L(x)/\bigl((\alpha-1)x^{\alpha-1}\mu_B\bigr)$ and $U^{(u)}/c(u)\xrightarrow{\mathscr{D}}R$, $U_0^{(u)}/c(u)\xrightarrow{\mathscr{D}}R_0$ where $c(u)=u$ and R,R_0 are Pareto* *with tails* $(1+x)^{-\alpha}$, *resp.* $(1+x)^{-(\alpha-1)}$.

(b) *If* B *is Weibull with tail* e^{-x^β} *for some* $\beta < 1$, *then* $\overline{B}_0(x) \sim \beta^{-1}x^{\beta-1}\mathrm{e}^{-x^\beta}$ *and* $U^{(u)}/c(u) \xrightarrow{\mathscr{D}} R$, $U_0^{(u)}/c(u) \xrightarrow{\mathscr{D}} R_0$ *where* $c(u) = u^{1-\beta}$ *and* R, R_0 *both are exponential with the same tail* $\mathrm{e}^{-\beta x}$.

Theorem 6.5 *In the set–up of Proposition 6.4, it holds that* $\tau(u)/c(u) \to R_0/(1-\rho)$ *in* $\mathbb{P}^{(u)}$*–distribution.*

For the proof, let R, N, N^*, $Y_1, Y_2, \ldots$ be independent r.v.'s, such that N is geometric with $\mathbb{P}(N=n) = (1-\rho)\rho^n$, $n = 0, 1, \ldots$, N^* is negative binomial with $\mathbb{P}(N^*=n) = (1-\rho)^2 n\rho^{n-1}$, $n = 1, 2, \ldots$ and $Y_1, Y_2, \ldots$ are i.i.d. with distribution B_0. Let further $Y_{(1)} < \cdots < Y_{(N)}$ be the order statistics and $M = Y_1 + \cdots + Y_N$. According to the Pollaczeck–Khinchine formula, we can think of M as $\sup_{t\ge 0} S_t$, of N as the number of ladder steps, of $Y_1, \ldots, Y_N$ as the ladder heights and of $\mathbb{P}^{(u)}$ as the conditional distribution given $M > u$.

Lemma 6.6 $\Bigl(N, Y_{(1)}, \ldots, Y_{(N-1)}, \dfrac{Y_{(N)}-u}{c(u)}\Bigr) \to \bigl(N^*, Y_1, \ldots, Y_{N^*-1}, R_0\bigr)$ *in* $\mathbb{P}^{(u)}$*–distribution.*

Proof. For $a_1, a_2, \ldots, x > 0$,

$$\begin{aligned}
&\mathbb{P}\bigl(M>u, Y_{(1)} \le a_1, \ldots, Y_{(N-1)} \le a_{N-1}, (Y_{(N)}-u)/c(u) > x\bigr)\\
&= \sum_{n=1}^{\infty}(1-\rho)\rho^n \cdot n\mathbb{P}\bigl(Y_1 \le a_1, \ldots, Y_{n-1} \le a_{n-1}, (Y_n - u)/c(u) > x\bigr)\\
&= \frac{\rho}{1-\rho}\overline{B}_0(u+c(u)x)\sum_{n=1}^{\infty}\mathbb{P}\bigl(N^* = n, Y_1 \le a_1, \ldots, Y_{n-1} \le a_{n-1}\bigr).
\end{aligned}$$

Now just note that

$$\frac{\rho}{1-\rho}\overline{B}_0(u+c(u)x) \sim \frac{\rho}{1-\rho}\overline{B}_0(u)\mathbb{P}(R_0 > x) \sim \mathbb{P}(M>u)\mathbb{P}(R_0 > x)$$

by X.9.1. □

Proof of Theorem 6.5. Recall that $Z = -S_{\tau(0)-}$, $Y = S_{\tau(0)}$. Clearly, the $\mathbb{P}^{(0)}$–distribution of Y given $Z = z$ is that of $U^{(z)}$. This shows (cf. V.3) that the joint $\mathbb{P}^{(0)}$–distribution of (Z, Y) is the same as the joint distribution of the backward and forward recurrence time distribution in a stationary renewal process with interarrival distribution B_0. Thus by symmetry, the $\mathbb{P}^{(0)}$–distribution of Z given $Y = y$ is that of $U^{(y)}$.

Let $\tau_1, \ldots, \tau_N$ be the lengths of the finite ladder segments. Thus the τ_k have the distribution of $\tau(0)$ given $\tau(0) < \infty$. Define Y_n, Z_n in terms of the nth ladder segment just as $Y = Y_1, Z = Z_1$ is defined in terms of the first, and let $\nu = \inf\{n : Y_1 + \cdots + Y_n > u\}$. Then $\tau(u) = \tau_1 + \cdots + \tau_\nu$ in $\mathbb{P}^{(u)}$–distribution. Lemma 6.6 now gives that τ_k has a limit distribution (viz. the distribution of $\omega(Z)$) for $k \ne \nu$; in particular (since N has the finite limit N^*), $\sum_{k\neq\nu}\tau_k \to 0$ in $\mathbb{P}^{(u)}$–distribution. For $k=\nu$, we write $R_0(u) = (Y_\nu - u)/c(u)$ and get

$$\begin{aligned}\mathbb{P}^{(u)}\bigl(Z_\nu > c(u)x\bigr) &= \int_0^\infty \frac{\overline{B}\bigl(y+c(u)x\bigr)}{\overline{B}(y)}\,\mathbb{P}^{(u)}(Y_\nu \in \mathrm{d}y)\\ &= \int \frac{\overline{B}\bigl(u+c(u)(r+x)\bigr)}{\overline{B}\bigl(u+c(u)r\bigr)}\,\mathbb{P}^{(u)}(R_0\in \mathrm{d}r)\,. \qquad (6.3)\end{aligned}$$

Here the integrand is bounded and has the continuous limit $\mathbb{P}(R > d(r)x)$ where

$$d(r) \;=\; \lim_{u\to\infty}\frac{c(u)}{c\bigl(u+c(u)r\bigr)} \;=\; \begin{cases}(1+r)^{-1} & \text{in the Pareto case (a),}\\ 1 & \text{in the Weibull case (b).}\end{cases}$$

Since the limit R_0 of $R_0(u)$ is continuous and concentrated on $(0,\infty)$, it follows that (6.3) asymptotically is

$$\int_0^\infty \frac{1}{[1+d(r)x]^\alpha}\,\frac{\alpha-1}{(1+r)^\alpha}\,\mathrm{d}r \;=\; \int_0^\infty \frac{\alpha-1}{(1+r+x)^\alpha}\,\mathrm{d}r \;=\; \frac{1}{x^{\alpha-1}} \;=\; \mathbb{P}(R_0 > x)$$

in case (a), whereas in case (b) we get

$$\int_0^\infty \mathrm{e}^{-\beta x}\beta\mathrm{e}^{-\beta r} \;=\; \mathrm{e}^{-\beta x} \;=\; \mathbb{P}(R_0 > x).$$

It follows that $Z_\nu/c(u) \to R_0$ in $\mathbb{P}^{(u)}$–distribution, and hence $\omega(z)/z \to (1-\rho)^{-1}$ yields

$$\tau_\nu \;=\; \omega(Z_\nu) \;\sim\; \omega\bigl(c(u)R_0\bigr) \;\sim\; \frac{c(u)R_0}{1-\rho},$$

$$\frac{\tau(u)}{c(u)} \;=\; \frac{1}{c(u)}\sum_{k<\nu}\tau_k + \frac{\tau_\nu}{c(u)} \;\to\; 0 + \frac{R_0}{1-\rho}\,. \qquad \square$$

Notes The study of the ruin time and the finite horizon ruin probabilities $\psi(u,T)$ is a classical topic in insurance risk. See Asmussen (2000), Ch. IV and IX.5, for a survey and references.

Theorem 6.5 is from Asmussen and Klüppelberg (1997) who also gave a general subexponential formulation and random walk parallels. Lemma 6.1 is also from that paper, where it was derived as an immediate application of excursion theory for Markov processes as formulated in Fitzsimmons (1987).

Appendix

A1 Polish Spaces and Weak Convergence

Polish spaces are of importance in probability theory, on the one hand, to provide a common framework comprising Euclidean space $\mathbb{R}^n$ and its nice subsets, discrete (finite or countable) sets and also some function spaces like D, and on the other hand, to possess many of the same regularity properties as $\mathbb{R}^n$ (e.g. in Polish spaces, Kolmogorov's consistency theorem holds and regular conditional distributions exist; see Neveu, 1965). Fundamental examples are:

(a) any locally compact space with a countable dense subset is Polish;
(b) if E is Polish, then $F \subseteq E$ is so if F is, say, closed or open (in fact, F is Polish if and only if F is a G_δ, i.e. of the form $\cap_0^\infty F_n$ with the F_n open);
(c) any countable product $E_0 \times E_1 \times \cdots$ of Polish spaces $E_0, E_1, \ldots$ is Polish;
(d) if E is Polish, then so is the set $\mathscr{P}(E)$ of probability distributions on E equipped with the topology for weak convergence and (in the locally compact case) the set $\mathscr{M}(E)$ of nonnegative Radon measures on E equipped with the topology for vague convergence, see below;
(e) if E is Polish, then so are function spaces like $D([0,1],E)$ and $D([0,\infty),E)$ in the standard Skorokhod topology.

Now let E be Polish and $\mathscr{C}$ the set of bounded continuous functions $E \to \mathbb{R}$. The *Borel σ–algebra* $\mathscr{E}$ on E is defined as the σ–algebra generated by the open sets (or the $f \in \mathscr{C}$) and is used throughout for measure theory. The *topology for weak convergence* of probability measures is the initial topology defined by the mappings $\mathbb{P} \to \int f \, \mathrm{d}\mathbb{P}$, $f \in \mathscr{C}$, i.e. the weakest topology making all these mappings continuous. If $\mathbb{P}_n \to \mathbb{P}$, then $\int f \, \mathrm{d}\mathbb{P}_n \to \int f \, \mathrm{d}\mathbb{P}$ not only for $f \in \mathscr{C}$ but also if f is bounded and measurable with $\mathbb{P}D_f = 0$, where D_f is the set of discontinuities of

f. Occasionally we use *Prohorov's theorem*, stating that a set $\mathscr{P}_0$ of probability measures is (weakly) relatively compact if and only if $\mathscr{P}_0$ is *tight*, i.e. if to each $\epsilon > 0$ we can find a compact set $K \subseteq E$ with $\mathbb{P}K \geq 1 - \epsilon$ for all $\mathbb{P} \in \mathscr{P}_0$. Note in particular that convergent sequences form relatively compact sets.

A closely related topology is the *topology for vague convergence* of Radon measures μ on a locally compact space E (Radon means that $\mu(K) < \infty$ when K is compact). This is defined as the initial topology defined by the set of mappings $\mu \to \int f \, d\mu$ indexed by the continuous f with compact support.

Some standard references are Billingsley (1968), Dudley (1989), Ethier and Kurtz (1986) and Pollard (1984). For D–spaces, see also Whitt (2002). For the present purposes, elementary treatments of weak convergence in $\mathbb{R}$ like that of Breiman (1968, Ch. 8) most often suffice.

For some (certainly not all) purposes the case of a general E can be reduced to the compact and/or real case by noting that E is homomorphic to a subset (necessarily a G_δ!) of $[0,1]^{\mathbb{N}}$. This follows from the following simple lemma:

Lemma A1.1 *If E is Polish, then there exists a countable class $\mathscr{K}$ of continuous functions $f : E \to [0,1]$ such that $x_n \to x$ in E if and only if $f(x_n) \to f(x)$ in $[0,1]$ for all $f \in \mathscr{K}$.*

Proof. Take $y_1, y_2, \ldots$ as a countable dense subset, let d be some metric and let $f_{k,n} : E \to [0,1]$ be continuous with $f_{k,n}(y) = 1$ for $d(y, y_k) \leq 1/n$, $f_{k,n}(y) = 0$ for $d(y, y_k) \geq 2/n$. Then $\mathscr{K} = \{f_{k,n} : k, n = 1, 2, \ldots\}$ is easily seen to have the desired property. □

A2 Right–Continuity and the Space D

The stochastic processes $\{X_t\}_{t \in \mathbb{T}}$ encountered in this book have almost exclusively a one–dimensional discrete ($\mathbb{T} = \mathbb{N}$) or continuous ($\mathbb{T} = [0, \infty)$) time parameter. Occasionally also the doubly infinite time case $\mathbb{T} = \mathbb{Z}$ or $\mathbb{T} = (-\infty, \infty)$ is encountered. The state space E is usually of an elementary type, discrete (finite or countable, e.g. $\mathbb{Z}, \mathbb{N}^p$), a well–behaved subset such as $[0, \infty)$, $(a, b]^p$ of Euclidean space $\mathbb{R}^p$ or combinations like $[0, \infty) \times \{0, 1\}$. In any case, it is more than sufficient to allow E to be a general Polish space which we then equip with the Borel σ–algebra $\mathscr{E}$. When we talk about a subset A of E, this is most often assumed to be measurable ($A \in \mathscr{E}$) without further notice.

The traditional definition of a stochastic process $\{X_t\}_{t \in \mathbb{T}}$ with state space E means just an indexed set of measurable mappings from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ into a general measurable space $(E, \mathscr{E})$. In discrete time, this is quite sufficient, but difficulties arise in continuous time. This is due to the fact that in discrete time the relevant events such as $\{X_t > 0 \text{ for all } t\}$, $\{X_t \to 0\}$, $\{\sup_{0 \leq t \leq T} X_t > u\}$, etc. are virtually always measurable since they can be obtained from elementary measurable sets of the form $X_n \in A_n$, $A_n \in \mathscr{E}$, by countable combinations of elementary set operations such as unions, intersections, differences, etc. In continuous time, this is not so, and in fact the examples above are not measurable events. A variety of suggestions (separability, joint measurability of $X_t(\omega)$ in t, ω, etc.) to overcome such difficulties have been considered, but the standard point of view in today's literature is to assume that for a.a. (or even all) $\omega \in \Omega$ the sample function $\{X_t(\omega)\}_{t\in\mathbb{T}}$ belongs to a well–behaved space of functions, of which the standard choice is D. This is quite sufficient for the present purposes, and in fact the sample paths of the processes under study exhibit most often even stronger regularity such as being piecewise continuous.

Noteworthy properties of a D–function $x = \{x_t\}_{t\in\mathbb{T}}$ are:

(i) x is given by the values on any dense countable set, say $\mathbb{Q}$ (this is elementary by right–continuity);

(ii) x is bounded on compact intervals (this is easy by a compactness argument);

(iii) x has at most countably many jumps. We shall prove this in a more general setting:

Proposition A2.1 *If $x : [0,\infty) \to E$ is right–continuous, then x is continuous except at a (at most) countable collection of points.*

Proof (adapted from Björnsson, 1988). In view of Lemma A1.1, we must show that $t \to f(x_t)$ has the desired property for each $f \in \mathscr{K}$. That is, it is sufficient to consider the case $E = [0,1]$. For $t > 0$ let

$$y_t^{(1)} = \overline{\lim_{s\uparrow t}}\, x_s - \underline{\lim_{s\uparrow t}}\, x_s, \quad y_t^{(2)} = \begin{cases} |x_t - \lim_{s\uparrow t} x_s| & \text{if } y_t^{(1)} = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then if x is discontinuous at t, we have either $y_t^{(1)} > 0$ or $y_t^{(2)} > 0$, and it is sufficient to show that for any $\epsilon > 0$ the sets $A^{(i)} = \{t : y_t^{(i)} > \epsilon\}$ are at most countable. It is clear by right–continuity that

$$0 = \lim_{s\downarrow t} y_s^{(1)} = \lim_{s\downarrow t} y_s^{(2)} \quad \text{for any } t \geq 0. \tag{A.2.1}$$

In particular, taking $t = 0$ gives $A^{(i)} \cap (0,\delta] = \emptyset$ for some $\delta > 0$ and hence

$$\tau^{(i)} = \sup\Big\{\delta > 0 : A^{(i)} \cap (0,\delta] \text{ is at most countable}\Big\}$$

is nonzero. But if $A^{(i)}$ was uncountable, we would have $\tau^{(i)} < \infty$, implying the existence of a sequence $\tau^{(i)}(n) \in A^{(i)} \cap (\tau^{(i)},\infty)$ with $\tau^{(i)}(n) \downarrow \tau^{(i)}$. Then $\underline{\lim}_{s\downarrow\tau^{(i)}} y_s^{(i)} \geq \epsilon$, contradicting (A.2.1). □

Define

$$\begin{aligned} Y_{t,f}^{(i)}(\omega) &= y_t^{(i)}\Big(f(X_t(\omega))\Big), \quad f \in \mathscr{K} \quad \text{(cf. Lemma A1.1)}, \\ C_u &= \{\omega : X_t(\omega) \text{ is continuous at } t = u\} = \bigcap_{f\in\mathscr{K}} \Big\{Y_{t,f}^{(i)} = 0,\, i = 1,2\Big\}. \end{aligned}$$

It follows by right–continuity that $\underline{\lim}_{t\uparrow u} f(X_t)$ and $\overline{\lim}_{t\uparrow u} f(X_t)$ are both measurable. Hence $Y_{t,f}^{(i)}$ and C_u are so, and it makes sense to define u to be a *fixed discontinuity* of $\{X_t\}$ if $\mathbb{P}C_u < 1$.

Corollary A2.2 *A stochastic process $\{X_t\}$ with right–continuous paths has at most countably many fixed discontinuities.*

Proof. By (A.2.1) and dominated convergence we have $\lim_{s\downarrow t} \mathbb{E}Y_{t,f}^{(i)} = 0$ for all t. Hence exactly as in Proposition A2.1 we may conclude that $\mathbb{E}Y_{t,f}^{(i)} = 0$ except for

t in a at most countable set $N_f^{(i)}$. But when $t \notin \cup_{i,f} N_f^{(i)}$, we have all $Y_{t,f}^{(i)} = 0$ a.s., implying $\mathbb{P}C_t = 1$. □

Corollary A2.3 *A stationary process $\{X_t\}$ with right–continuous paths has no fixed discontinuities.*

Proof. If $t \in [0, \infty)$ is a fixed discontinuity, then so is any $s \in [0, \infty)$ by stationarity. Hence the set of fixed discontinuities equals $[0, \infty)$, which is a contradiction since it is countable. □

We will occasionally need to consider random E–valued functions with a possibly finite (random) lifelength. We view these as measurable elements of the space $D_0(E)$, defined as the subset of $D(E \cup \{\Delta\})$ (where $\Delta \notin E$ is some extra point) of functions $\{x_t\}_{0 \le t < \infty}$ having the property $x_t = \Delta$ for all $t \ge \zeta$ where $\zeta = \inf\{t : x_t = \Delta\}$. We identify such a function with $\{x_t\}_{0 \le t < \zeta}$ and the lifelength with ζ. Since $D_0(E)$ is an open subset of $D(E \cup \{\Delta\})$, the topology and measurability structure can be defined in a straightforward way by taking restrictions.

A3 Point Processes

A point process on E is in intuitive terms just a random collection of points in E. The simplest standard example is the Poisson process on $[0, \infty)$. In fact almost exclusively the point processes encountered in this book have $E = [0, \infty)$ or $E = \mathbb{R}$ and satisfy some further regularity conditions: (a) there are no multiple points; (b) the points do not accumulate. Processes of such types are easily brought in one–to–one correspondence with sequences $\{S_n\}$ of $(0, \infty)$–valued r.v.'s (e.g. if $E = [0, \infty)$, we may just let S_n be the position of the nth point to the right of the origin) and thus no foundational difficulties arise. It may, however, frequently be revealing also to have the general abstract formulation in mind. One then requires E to be locally compact Polish and defines a point process on E to be a $\mathscr{N}$–valued random variable N, where $\mathscr{N}$ is the set of (Radon) counting measures (thus the connection to the setting above is $N(A) = \sum I(S_n \in A)$.

A particular important class of point processes is the *intensity-driven* ones, commonly referred to as *Cox processes*. Such a process is specified in terms of its intensity function $\beta(t)$, and the interpretation is at the intuitive level that an epoch occurs in $(t, t+h]$ w.p. $\beta(t)h + \mathrm{o}(h)$ independently of the past, at the formal level that $\{N_t - B(t)\}$ is a local martingale where $B(t) = \int_0^t \beta(s)\,\mathrm{d}s$. A convenient representation can be given in terms of an independent Poisson process $\{N'_t\}$ with epochs $\{T'_k\}$: the epochs $\{T_k\}$ of $\{N_t\}$ are given by $T_k = B^{-1}(T'_k)$. That this yields the desired interpretation follows at the intuitive level from

$$\begin{aligned}\mathbb{P}(N(t, t+h] = 1) &= \mathbb{P}(N'(B(t), B(t+h)] = 1) \\ &= [B(t+h) - B(t)]h + \mathrm{o}(B(t+h) - B(t))\end{aligned}$$

and $B(t+h) - B(t) = \beta(t)h + \mathrm{o}(h)$; the formal verification of the martingale property is straightforward when $\mathbb{E}B(t) < \infty$ for all t, and otherwise just use the $B^{-1}(n)$ as localization sequence.

Under mild absolute continuity requirements, any point process can be viewed as intensity-driven, but this point of view may not always be the natural one.

Of the many texts on point processes, we mention in particular Daley and Vere-Jones (1988), Kallenberg (1976) and Matthes *et al.* (1978); there is also much material in Baccelli and Brémaud (2002), Franken *et al.* (1982) and Rolski *et al.* (1999). As mentioned above, the concept occurs in an elementary manner in the present book. One piece of terminology that is used occasionally is that of the *intensity measure* which is defined as the set function $A \to \mathbb{E}N(A)$; thus e.g. for the (time–homogeneous) Poisson process with intensity β, the intensity measure is Lebesgue measure scaled by β.

A4 Stochastical Ordering

Let X, Y be real–valued r.v.'s with distributions F, G. We then say that $X \leq Y$ in the sense of stochastical ordering (written $(X \leq_{\mathrm{so}} Y)$ if (a) $\mathbb{P}(X > t) \leq \mathbb{P}(Y > t)$ for all t, i.e. if $\overline{F} \leq \overline{G}$ or, equivalently, $F \geq G$. Alternative formulations in the same situation are "X is stochastically dominated by Y (or G)," "F is stochastically smaller than G" and so on. Some of the facts that we use are the equivalence of the definition (a) to either of (b) $\mathbb{E}f(X) \leq \mathbb{E}f(Y)$ for all increasing f, or (c) there exists r.v.'s X^*, Y^* with $X \overset{\mathscr{D}}{=} X^*$, $Y \overset{\mathscr{D}}{=} Y^*$ and $X^* \leq Y^*$ a.s.

The convex ordering $\leq_{\mathrm{cx}}$ is defined by $X \leq_{\mathrm{cx}} Y$ if (b$'$) $\mathbb{E}f(X) \leq \mathbb{E}f(Y)$ for all convex f. A special case of a theorem of Strassen (see e.g. Lindvall, 1992) gives that this is equivalent to (c$'$) there exists X^*, Y^* with $X \overset{\mathscr{D}}{=} X^*$, $Y \overset{\mathscr{D}}{=} Y^*$ and $X^* = \mathbb{E}[Y^* | X^*]$ a.s. or, equivalently, such that (X^*, Y^*) is a martingale. One usually interprets $X \leq_{\mathrm{cx}} Y$ as one of the possible definitions of Y being more variable than X; for example, taking $f(x) = x$ and $f(x) = -x$ shows that $\mathbb{E}X = \mathbb{E}Y$, and taking $f(x) = x^2$ then yields $\mathbb{V}arX \leq \mathbb{V}arY$. The characterization in terms of the distribution function is (a$'$) $\mathbb{E}X = \mathbb{E}Y$ and

$$\int_x^\infty \overline{F}(y)\,\mathrm{d}y \;\leq\; \int_x^\infty \overline{G}(y)\,\mathrm{d}y \;\text{ for all } x. \tag{A.4.1}$$

In between the two orderings is the increasing convex ordering $\leq_{\mathrm{icx}}$, defined by $X \leq_{\mathrm{cx}} Y$ if (b$''$) $\mathbb{E}f(X) \leq \mathbb{E}f(Y)$ for all nondecreasing convex f. Again by Strassen, this is equivalent to (c$''$) there exists X^*, Y^* with $X \overset{\mathscr{D}}{=} X^*$, $Y \overset{\mathscr{D}}{=} Y^*$ and $X^* \leq \mathbb{E}[Y^* | X^*]$ a.s. or, equivalently, such that (X^*, Y^*) is a submartingale. The characterization in terms of the distribution function is (a$''$) $=$ (A.4.1). If X, Y are, e.g., steady–state waiting times in two different queueing systems, one often interprets $X \leq_{\mathrm{icx}} Y$ as the X–system having the better performance. For example, taking $f(x) = x$ shows that $\mathbb{E}X \leq \mathbb{E}Y$, and if one has Cramér–Lundberg asymptotics $\overline{F}(x) \sim C\mathrm{e}^{-\gamma x}$, $\overline{G}(x) \sim D\mathrm{e}^{-\eta y}$, then $\gamma \geq \eta$ (from $\mathbb{E}\mathrm{e}^{\alpha X} \leq \mathbb{E}\mathrm{e}^{\alpha Y}$ it follows that $\mathbb{E}\mathrm{e}^{\alpha X} < \infty$ for $0 < \alpha < \eta$ and therefore $\gamma \geq \eta$).

General references for orderings of r.v.'s are Shaked and Shantikumar (1994), Szekli (1995) and Müller and Stoyan (2002). For queueing applications in particular, see also Stoyan (1983), Chen and Yao (1999) and Baccelli and Brémaud (2002).

A5 Heavy Tails

The definition that a distribution B on $(0, \infty)$ is *heavy–tailed* invariably requires that $\int_0^\infty \mathrm{e}^{\epsilon x} B(\mathrm{d}x) = \infty$ for all $\epsilon > 0$. This holds in particular if

$$\lim_{x \to \infty} \frac{\overline{B}(x+y)}{\overline{B}(x)} \;=\; 1, \tag{A.5.1}$$

for all $y < \infty$, and a distribution with this property is often referred to as *long–tailed*. However, (A.5.1) is not strong enough to allow for asymptotic studies of say waiting time tails, and for such purposes one usually works within the class $\mathscr{S}$ of *subexponential distributions*, defined by the requirement

$$\lim_{x \to \infty} \frac{\overline{B^{*n}}(x)}{\overline{B}(x)} \;=\; n \tag{A.5.2}$$

for all $n = 2, 3, \ldots$ (actually, it suffices that (A.5.2) holds for $n = 2$). Some of the fundamental properties are:

Proposition A5.1 (a) *Any* $B \in \mathscr{S}$ *satisfies* (A.5.1)*;* (b) *if* $B \in \mathscr{S}$*, then for any* $z > 1$ *there is a* $D < \infty$ *such that* $\overline{B^{*n}}(u) \le \overline{B}(u) D z^n$ *for all* u*;* (c) *if* A *is a distribution on* $(0, \infty)$ *such that* $\overline{A}(x)/\overline{B}(x)$ *has a limit* $c \in [0, \infty)$*, then* $A * B \in \mathscr{S}$ *and* $\overline{A * B}(x) \sim (1+c)\overline{B}(x)$*. If* $c > 0$*, then also* $A \in \mathscr{S}$*.*

Two basic examples are regularly varying distributions, defined by $\overline{B}(x) = L(x)/x^\alpha$ where $\alpha \ge 0$ and $L(\cdot)$ is *slowly varying* ($L(tx)/L(x) \to 1$ for all t), and Weibull–like distributions with $\overline{B}(x) \sim c_1 x^\gamma \mathrm{e}^{-c_2 x^\beta}$ with $0 < \beta < 1$. The regularly varying distributions are the most heavy–tailed subexponential distributions, whereas the Weibull–like ones have light–tailed distributions (with a Gamma–like tail) as a boundary case when $\beta \uparrow 1$. In between is the third main example, the class of lognormal distributions.

Many specific investigations call for some added regularity, and there is an abundance of subclasses of $\mathscr{S}$ in the literature designed for such purposes. Usually, all the three main examples are in such subclasses, but note that the regularly varying case is sometimes the easiest to analyze. One of the main subclasses is the class $\mathscr{S}^*$ introduced by Klüppelberg and defined by

$$\int_0^{x/2} \overline{B}(x-y)\overline{B}(y)\,\mathrm{d}y \;\sim\; \overline{B}(x)\frac{\mu_B}{2}. \tag{A.5.3}$$

Surveys of heavy–tailed distributions, references and applications are in Embrechts *et al.* (1997) and Asmussen (2000). There is also much material (including motivating statistical studies) in the volume edited by Adler *et al.* (1998).

A6 Geometric Trials

Lemma A6.1 *Let* $\{\mathscr{F}_n\}_{n=1,2,\ldots}$ *be a filtration and* $\{A_n\}$ *an adapted sequence of events,* $\tau = \inf\{n \ge 1 : A_n\} = \inf\{n : I(A_n) = 1\}$*. If* $\mathbb{P}(A_{n+1} \,|\, \mathscr{F}_n) \ge \theta$ *a.s. for all* n *and some* $\theta > 0$*, then* $\tau \le_{\mathrm{so}} T$ *where* T *is a r.v. with* $\mathbb{P}(T = n) = \theta(1-\theta)^{n-1}$*. In particular,* $\tau < \infty$ *a.s.,* $\mathbb{E}\tau \le 1/\theta$ *and* $\mathbb{E}z^\tau < \infty$ *whenever* $z(1-\theta) < 1$*.*

Proof. For $n \geq 0$,

$$\begin{aligned}\mathbb{P}(\tau \geq n+1) &= \mathbb{P}(A_1^c \cdots A_n^c) = \mathbb{E}\Big[I(A_1^c) \cdots I(A_{n-1}^c)\,\mathbb{P}(A_n^c \,|\, \mathscr{F}_{n-1})\Big] \\ &\leq (1-\theta)\mathbb{P}(A_1^c \cdots A_{n-1}^c) \leq \cdots \leq (1-\theta)^n.\end{aligned}$$

□

A7 Semigroups of Positive Numbers

For $\mathscr{G} \subseteq (0,\infty)$, we say that $\mathscr{G}$ is *lattice with span* $h > 0$ if h is the greatest common divisor of the $g \in \mathscr{G}$ and *nonlattice* if no common divisor exists.

Denote by $\mathscr{G}_+$ the additive semigroup generated by $\mathscr{G}$ and by $\mathscr{G}_\pm$ the additive group. Thus, $\mathscr{G}_+, \mathscr{G}_\pm$ are the sets of all finite linear combinations $a_1g_1 + \cdots + a_ng_n$ with $g_1,\ldots,g_n \in \mathscr{G}$ and $a_1,\ldots,a_n \in \mathbb{N}\backslash\{0\}$, resp. $a_1,\ldots,a_n \in \mathbb{Z}\backslash\{0\}$.

The example we have in mind is $\mathscr{G}$ being the support of a measure on $(0,\infty)$, and the following result is a key tool for ergodic theory for Markov chains and renewal theory:

Theorem A7.1 (a) *If $\mathscr{G}$ is lattice with span h, then there is a $n_0 \in \mathbb{N}$ such that $\{n_0h, (n_0+1)h, \ldots\} \subseteq \mathscr{G}_+$;* (b) *if $\mathscr{G}$ is nonlattice, then $\mathscr{G}_+$ is dense at ∞ in the sense that* $\lim_{x\to\infty} d(x,\mathscr{G}_+) = 0$ *where* $d(x,\mathscr{G}_+) = \inf\{|x-g| : g \in \mathscr{G}_+\}$.

Lemma A7.2 *Assume $m > 0$, $m \in \mathscr{G}_\pm$. Then $nm \in \mathscr{G}_+$ for all large $n \in \mathbb{N}$.*

Proof. Assume w.l.o.g. that $m = 1$ and write $1 = m = \sum_1^j a_ig_i$ with $a_i \in \mathbb{Z}$, $g_1,\ldots,g_j \in \mathscr{G}$, let $k = \sum_1^n |a_i|g_i$ and $n_0 = k^2$. If $n \geq n_0$, we can write $n = sk + t$ with $s \geq k$, $0 \leq t < k$. Hence $n = \sum_{i=1}^j (s|a_i| + ta_i)g_i$ is in $\mathscr{G}_+$ since $s - t \geq k - k = 0$ and hence $s|a_i| + ta_i \geq 0$. □

Proof of Theorem A7.1. In (a), assume w.l.o.g. that $h = 1$ and let m be the minimal positive element of $\mathscr{G}_\pm$. If $m > 1$, there would exist $d \in \mathscr{G}$ which is not a multiple of m, i.e. $km < d < (k+1)m$ for some $k = 1, 2, \ldots$. Then $0 < d - km < m$ which contradicts the choice of m since $d - km \in \mathscr{G}_\pm$. The result then follows immediately from Lemma A7.2.

In (b), it suffices by Lemma A7.2 to show that $m = 0$ where $m = \inf\{m^* : m^* > 0, m^* \in \mathscr{G}_\pm\}$. But if $m > 0$, we can find first $d \in \mathscr{G}$ which is not a multiple of m, i.e. $0 < d - km < m$ for some k. Then also $0 < d - km^* < m$ for some $m^* \in \mathscr{G}_\pm$ which is a contradiction since $d - km^* \in \mathscr{G}_\pm$. □

A8 Total Variation Convergence

Let $(E,\mathscr{E})$ be a measurable space and ν a signed measure on $(E,\mathscr{E})$. Then the *total variation* (t.v.) of ν is defined as $\|\nu\| = \sup_{A\in\mathscr{E}}\nu(A)$. If $\nu \geq 0$, then $\|\nu\| = \nu(E)$. However, the main case for our applications is $\nu = \mathbb{P}' - \mathbb{P}''$ with $\mathbb{P}', \mathbb{P}''$ probabilities. Then $2\|\nu\| = \nu(E_+) - \nu(E_-)$, where $E = E_+ \cup E_-$ is the Jordan–Hahn decomposition of E w.r.t. ν. We say that $\nu_n \to \nu$ in t.v. if $\|\nu_n - \nu\| \to 0$, i.e. if $\nu_n(A) \to \nu(A)$ uniformly in $A \in \mathscr{E}$, which in turn is easily seen to be equivalent to $\int g \, \mathrm{d}\nu_n \to \int g \, \mathrm{d}\nu$ uniformly in the measurable g with $\|g\|_\infty \le 1$. Similarly, $X_n \to X$ in t.v. means that $\mathbb{P}(X_n \in \cdot) \to \mathbb{P}(X \in \cdot)$ in t.v.

Taking g continuous shows that t.v. convergence entails weak convergence. One important example of t.v. convergence is provided by *Scheffe's theorem* (Billingsley, 1968, p. 224) which states that if ν_n, ν are probabilities with densities f_n, f w.r.t. μ, and $f_n(x) \to f(x)$ for μ–a.a. x, then $\nu_n \to \nu$ in t.v. This means in particular that for a discrete E, the notions of weak convergence and t.v. convergence coincide. In fact, if μ is counting measure on E, then $\nu_n \to \nu$ and E being discrete implies that

$$f_n(x) \;=\; \nu_n(\{x\}) \;\to\; \nu(\{x\}) \;=\; f(x) \ \text{ for all } x.$$

Note that in much of the literature, the definition of $\|\nu\|$ differs by a factor of 2. Thus, for example, in the coupling inequality VII.(2.1) the r.h.s. is frequently encountered as $2\mathbb{P}(T > t)$ rather than $\mathbb{P}(T > t)$.

A9 Transforms

Transforms of a distribution F are denoted by $\widehat{F}[\cdot]$ which may mean either characteristic function (ch.f.) $\widehat{F}[s] = \int \mathrm{e}^{\mathrm{i}sx} F(\mathrm{d}x)$, Laplace transform $\widehat{F}[s] = \int \mathrm{e}^{-sx} F(\mathrm{d}x)$, moment generating function (m.g.f.) $\widehat{F}[s] = \int \mathrm{e}^{sx} F(\mathrm{d}x)$, or, if F is concentrated on $\mathbb{Z}$ with point probabilities $\{f_n\}$, (probability) generating function $\widehat{F}[s] = \widehat{f}[s] = \sum_{-\infty}^{\infty} s^n f_n$.

In the text, we use without further reference a number of standard facts such as that F is uniquely determined by $\widehat{F}$, that $\widehat{F * G} = \widehat{F}\widehat{G}$, that moments can be expressed in terms of derivatives of $\widehat{F}$ and so on.

The cumulant generating function (c.g.f.) is $\log \widehat{F}[s]$, where $\widehat{F}[s]$ is the m.g.f. A basic property is that its kth derivative at 0 is the kth cumulant of F (the first cumulant is the mean, the second the variance, the third the central third moment; for $k \ge 4$ the expressions quickly become less easily interpreted). For formulas connecting higher moments and cumulants, see Smith (1995).

In the book, we do no treat numerical transform inversion which of course is important in practice. Some selected references are Grübel (1991) and Abate and Whitt (1992, 1995).

A10 Stopping Times and Wald's Identity

Let $\mathbb{T} = \mathbb{N}$ or $\mathbb{T} = [0, \infty)$, and let $\{\mathscr{F}_t\}_{t \in \mathbb{T}}$ be a *filtration*, i.e. a nondecreasing family of σ–fields. A random time $\tau \le \infty$ is a *stopping time w.r.t. to* $\{\mathscr{F}_t\}_{t \in \mathbb{T}}$ if

$$\{\tau \le t\} \;\in\; \mathscr{F}_t \ \text{ for all } t \in \mathbb{T}. \tag{A.10.1}$$

The *stopping time σ–field* $\mathscr{F}_\tau$ (sometimes called the *pre–τ–field*) is then defined as the collection of all sets $A \in \mathscr{F}$, where $\mathscr{F} = \sigma(\cup_{t \in \mathbb{T}} \mathscr{F}_t)$, such that $A \cap \{\tau \le t\}$ belongs to $\mathscr{F}_t$ for all $t \in \mathbb{T}$.

In applications, it is convenient to note that measurability is an automatic result of (A.10.1) and needs not be checked separately, and also that for $\mathbb{T} = \mathbb{N}$

(A10.1) is equivalent to $\{\tau = n\} \in \mathscr{F}_n$ for all $n \in \mathbb{N}$ and $A \in \mathscr{F}_\tau$. The following result is standard and easy to prove:

Proposition A10.1 *Let τ be a stopping time. Then:* (a) *τ is $\mathscr{F}_\tau$–measurable;* (b) *if $\{X_t\}$ is a stochastic process such that X_t is $\mathscr{F}_t$–measurable for each $t \in \mathbb{T}$, and that the paths are right–continuous when $\mathbb{T} = [0, \infty)$, then $X_\tau I(\tau < \infty)$ is $\mathscr{F}_\tau$–measurable;* (c) *if σ is an additional $\{\mathscr{F}_t\}$–stopping time, then so are $\sigma \wedge \tau$, $\sigma \vee \tau$, $\sigma + \tau$. If $\sigma \le \tau$, then $\mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$.*

Part (a) of the following result is referred to as *Wald's identity* (sometimes called also *Wald's lemma*), and part (b) as *Wald's second moment identity.* For a proof and a thorough discussion, see e.g. Neveu (1972, pp. 83–85). The result states that optional stopping of the martingales $\{S_n - n\mu\}$ and $\{(S_n - n\mu)^2 - n\sigma^2\}$ is justified under very weak conditions.

Proposition A10.2 *Let τ be an a.s. finite stopping time w.r.t. $\{\mathscr{F}_n\}_{n\in\mathbb{N}}$. Further, let $X_1, X_2, \ldots$ be i.i.d. r.v.'s such that for any n X_n is $\mathscr{F}_n$–measurable and $X_{n+1}, X_{n+2}, \ldots$ are independent of $\mathscr{F}_n$, and write $S_n = X_1 + \cdots + X_n$, $\mu = \mathbb{E}X_1$. Then:*
(a) *if either $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}\tau < \infty$, or $X_1 \ge 0$, then $\mathbb{E}S_\tau = \mu\mathbb{E}\tau$;*
(a) *if $\sigma^2 = \mathbb{V}arX_1 < \infty$ and $\mathbb{E}\tau < \infty$, then $\mathbb{E}(S_\tau - \tau\mu)^2 = \sigma^2\mathbb{E}\tau$.*

The elementary case is of course $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$. In general, one might alternatively formulate the condition by $X_1, X_2, \ldots$ being i.i.d. and $\{\tau \le n\}$ being independent of $X_{n+1}, X_{n+2}, \ldots$ for any n. This means that τ is a *randomized stopping time*. Further examples of this notion occur in VII.3 and X.3.

A11 Discrete Skeletons

Limit theory for continuous–time stochastic processes ($\mathbb{T} = [0, \infty)$), being concerned with the question of existence of limits of functions like $f(t) = \mathbb{P}(X_t \in A)$ or $f(t) = \mathbb{E}g(X_t)$, can sometimes be reduced to the discrete case $\mathbb{T} = \mathbb{N}$ by means of the study of discrete skeletons $\{X_{n\delta}\}_{n\in\mathbb{N}}$. For example, elementary topology yields:

Proposition A11.1 *If $f : [0, \infty) \to \mathbb{R}$ is uniformly continuous and $\lambda(\delta) = \lim_{n\to\infty} f(n\delta)$ exists for each $\delta > 0$, then $\lambda \equiv \lambda(\delta)$ does not depend on δ, and furthermore $f(t) \to \lambda$ as $t \to \infty$ continuously.*

It is frequently much easier to show that a $f(t)$ of the type above is just continuous rather than uniformly continuous. In fact, this is sufficient:

Proposition A11.2 *The conclusion of Proposition A11.1 holds true if f is continuous and $\lambda(\delta) = \lim_{n\to\infty} f(n\delta)$ exists for each $\delta > 0$.*

This result is known in the literature as the *Croft–Kingman lemma.* The proof (Kingman, 1963) is again real topology, but much less elementary than for Proposition A11.1.

Bibliography

J. Abate and W. Whitt (1988) Transient behaviour of the M/M/1 queue via Laplace transforms. *Adv. Appl. Probab.* **20**, 145–178.

J. Abate and W. Whitt (1992) The Fourier–series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5–87.

J. Abate and W. Whitt (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Computing* **7**, 36–43.

J. Abate and W. Whitt (1999) Explicit $M/G/1$ waiting time distributions for a class of long–tail service–time distributions. *Opns. Res. Letters* **25**, 25–31.

M. Abramowitz and I. Stegun (1972) *Handbook of Mathematical Functions* (10th ed.). Dover.

R.J. Adler, R. Feldman and M.S. Taqqu (1998) *A User's Guide to Heavy Tails.* Birkhäuser.

A.S. Alfa and S. Chakravarty, eds. (1997) *Matrix–Analytic Methods for Stochastic Models.* Marcel Dekker.

A.S. Alfa and S. Chakravarty, eds. (1998) *Advances in Matrix Analytic Methods for Stochastic Models.* Notable Publications, New Jersey.

A.O. Allen (1978) *Probability, Statistics and Queueing Theory with Computer Applications.* Academic Press.

G. Alsmeyer (1991) *Erneuerungstheorie.* B.G. Teubner.

G. Alsmeyer (1994) Blackwell's renewal theorem for certain linear submartingales and coupling. *Acta Appl. Math.* **34**, 135–150.

G. Alsmeyer (1997) The Markov renewal theorem and related results. *Markov Proc. Rel. Fields* **3**, 103–127.

V. Anantharam (1988) How large delays build up in a $GI/GI/1$ queue. *Queueing Systems* **5**, 345–368.

W.J. Anderson (1991) *Continuous–Time Markov Chains. An Applications–Oriented Approach.* Springer–Verlag.

C.W. Andersson (1970) Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probab.* **7**, 99–113.

D. Anick, D. Mitra and M.M. Sondhi (1982) Stochastic theory of a data–handling system with multiple sources. *Bell System Tech. J.* **61**, 1871–1894.

K. Arrow, S. Karlin and H. Scarf (1958) *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press.

J.R. Artalejo (1999) A classified bibliography of research on retrial queues. *Top* **7**, 187–211.

S. Asmussen (1982) Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the $GI/G/1$ queue. *Adv. Appl. Probab.* **14**, 143-170.

S. Asmussen (1989) Aspects of matrix Wiener–Hopf factorization in applied probability. *The Mathematical Scientist* **14**, 101–116.

S. Asmussen (1992a) Phase–type representations in random walk and queueing problems. *Ann. Probab.* **20**, 772–789.

S. Asmussen (1992b) Light traffic equivalence in single–server queues. *Ann. Appl. Probab.* **2**, 555–574.

S. Asmussen (1995) Stationary distributions via first passage times. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 79–102. CRC Press.

S. Asmussen (1998a) Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage probabilities. *Ann. Appl. Probab.* **8**, 354–374.

S. Asmussen (1998b) Extreme value theory for queues via cycle maxima. *Extremes* **1**, 137–168.

S. Asmussen (1998c) A probabilistic look at the Wiener-Hopf equation. *SIAM Review* **40**, 189–201.

S. Asmussen (2000) *Ruin Probabilities.* World Scientific.

S. Asmussen and M. Bladt (1996) Renewal theory and queueing algorithms for matrix–exponential distributions. *Matrix–Analytic Methods in Stochastic Models* (A.S. Alfa and S. Chakravarty, eds.), 313–341. Marcel Dekker.

S. Asmussen and S. Foss (1993) Renovation, regeneration, and coupling in multiple–server queues in continuous time. *Frontiers in Pure and Applied Probability* (H. Niemi, G. Högnäs, A.N. Shiryaev and A.V. Melnikov, eds.), 1–6. VSP/TVP.

S. Asmussen, S. Foss and D. Korshunov (2003) Limit theorems for sums of random variables with a local subexponential behaviour. *J. Theor. Probab.* (to appear).

S. Asmussen, V. Kalashnikov, C. Klüppelberg, D. Konstantinides and G. Tsitsiasvili (2002) A local limit theorem for random walk maxima with heavy tails. *Statist. Probab. Letters* **56**, 399–404.

S. Asmussen and O. Kella (2000) A multidimensional martingale for Markov additive processes and its applications. *Adv. Appl. Probab.* **32**, 376–393.

S. Asmussen and O. Kella (2001) On optional stopping of some exponential martingales for Lévy processes with or without reflection. *Stoch. Proc. Appl.* **91**, 47–55.

S. Asmussen and C. Klüppelberg (1997) Large deviations for subexponential tails, with applications to insurance risk. *Stoch. Proc. Appl.* **64**, 103–125.

S. Asmussen and G. Koole (1993) Marked point processes as limits of Markovian arrival streams. *J. Appl. Probab.* **30**, 365–372.

S. Asmussen and J. Møller (2001) Calculation of the steady–state waiting time in $GI/PH/c$ and $MAP/PH/c$ queues. *Queueing Systems* **37**, 9–29.

S. Asmussen, O. Nerman and M. Olsson (1996) Fitting phase-type distributions via the EM algorithm. *Scand. J. Statist.* **23**, 419–441.

S. Asmussen and H.M. Nielsen (1995) Ruin probabilities via local adjustment coefficients. *J. Appl. Probab.* **32**, 736–755.

S. Asmussen and C.A. O'Cinneide (1999) Matrix–exponential distributions [Distributions with a rational Laplace transform]. *Encyclopedia of Statistical Sciences, Update Volume 3* (S. Kotz, C.B. Read, D.L. Banks, eds.), 435–440. Wiley.

S. Asmussen and R.Y. Rubinstein (1995) Steady–state rare events simulation in queueing models and its complexity properties. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 429–466. CRC Press.

S. Asmussen and K. Sigman (1996) Monotone stochastic recursions and their duals. *Probab. Th. Eng. Inf. Sc.* **10**, 1–20.

R. Atar and P. Dupuis (1999) Large deviations and queueing networks: methods for rate function identification. *Stoch. Proc. Appl.* **84**, 255–296.

S. Axsäter (2000) *Inventory Control.* Kluwer

F. Baccelli and P. Brémaud (2002) *Elements of Queueing Theory. Palm Martingale Calculus and Stochastic Recurrences* (2nd ed.). Springer–Verlag.

F. Baccelli and S. Foss (1994) Stability of Jackson–type queueing networks. *Queueing Systems* **17**, 5–72.

F. Baccelli and S. Foss (2003) Moments and tails in monotone–separable stochastic networks. *Ann. Appl. Probab.* **13**.

F. Baccelli and A.M. Makowski (1989) Dynamic, transient and stationary behaviour of the $M/G/1$ queue via martingales. *Ann. Probab.* **17**, 1691-1699.

F. Baccelli, S. Schlegel and V. Schmidt (1999) Asymptotics of stochastic networks with subexponential service times. *Queueing Systems* **33**, 205–232.

F. Baccelli and V. Schmidt (1996) Taylor series expansions for Poisson–driven $(\max, +)$–linear systems. *Ann. Appl. Probab.* **6**, 138–185.

F. Ball and V.T. Stefanov (2001) Further approaches to computing fundamental characteristics of birth–death processes. *J. Appl. Probab.* **38**, 995–1005.

A. Baltrunas, D.J. Daley and C. Klüppelberg (2002). Tail behaviour of the busy period of a $GI/G/1$ queue with subexponential service times. Manuscript.

O. Barndorff–Nielsen (1978) *Information and Exponential Families in Statistical Theory.* Wiley.

O. Barndorff–Nielsen, T. Mikosch and S. Resnick, eds. (2001) *Lévy Processes. Theory and Applications.* Birkhäuser.

F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios (1975) Open, closed and mixed networks of queues with different classes of customers. *J. ACM* **22**, 248–260.

H.C.P. Berbee (1979) *Random Walks with Stationary Increments and Renewal Theory.* Mathematical Centre Tracts **112**, Amsterdam.

A. Berman and R.J. Plemmons (1994) *Nonnegative Matrices in the Mathematical Sciences.* SIAM.

J. Bertoin (1996) *Lévy Processes.* Cambridge University Press.

J. Bertoin (1999) *Subordinators: Examples and Applications.* Ecole d'Ete de Probabilites de Saint–Flour XXVII. Lecture Notes in Mathematics **1727**. Springer–Verlag.

J. Bertoin and R.A. Doney (1994a) Cramér's estimate for Lévy processes. *Statist. Probab. Letters* **21**, 363–365.

J. Bertoin and R.A. Doney (1994b) Local properties of ladder height distributions. *J. Appl. Probab.* **31**, 816–821.

R.N. Bhattacharya and R.R. Rao (1976) *Normal Approximations and Asymptotic Expansions.* Wiley.

P. Billingsley (1968) *Convergence of Probability Measures.* Wiley; 2nd ed. published 1999.

N.H. Bingham (1976) Fluctuation theory in continuous time. *Adv. Appl. Probab.* **7**, 705–766.

O.J. Björnsson (1988) Remarks on functions which are continuous from one side (in Danish). *Normat* **3**, 101–103.

M. Bladt, A. Gonzales and S.L. Lauritzen (2003) The estimation of phase–type related functionals through Markov chain Monte Carlo methodology. *Scand. Act. J.* (to appear).

B. Blaszczyszyn, T. Rolski and V. Schmidt (1995) Light–traffic approximations in queues and related models. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 379–406. CRC Press.

B. Blaszczyszyn and K. Sigman (1999) Risk and duality in multidimensions. *Stoch. Proc. Appl.* **83**, 331–356.

A.N. Borodin and P. Salminen (1996) *Handbook of Brownian Motion — Facts and Formulae.* Birkhäuser.

A.A. Borovkov (1976) *Stochastic Processes in Queueing Theory.* Springer.

A.A. Borovkov (1984) *Asymptotic Methods in Queueing Theory.* Wiley.

A.A. Borovkov (1999) *Ergodicity and Stability of Stochastic Processes.* Wiley.

A.A. Borovkov and S. Foss (1992) Stochastically recursive sequences and their generalization. *Siberian Math. J.* **43**, 16–81.

S.C. Borst, O.J. Boxma, J.A. Morrison and R. Nunez Queija (2003) The equivalence between processor sharing and service in random order. *Opns. Res. Letters* (to appear).

O.J. Boxma, Q. Deng and A.P. Zwart (2002) Waiting time asymptotics for the $M/G/2$ queue with heterogeneous servers. *Queueing Systems* **40**, 5–31.

N.L. Bowers, Jr., H.U. Gerber, J.C. Hickman, D.A. Jones and C.J. Nesbitt (1986) *Actuarial Mathematics.* The Society of Actuaries, Itasca, Illinois.

M. Bramson (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30**, 89–148.

A. Brandt, P. Franken and B. Lisek (1990) *Stationary Stochastic Models.* Wiley.

L. Breiman (1968) *Probability.* Addison–Wesley. Reprinted by SIAM 1998.

P. Brémaud (1981) *Point Processes and Queues.* Springer–Verlag.

P. Brémaud (1999) *Markov Chains. Gibbs Fields, Monte Carlo Simulation and Queues.* Springer–Verlag.

P.J. Brockwell, S.I. Resnick and R.L. Tweedie (1982) Storage processes with general release rule and additive inputs. *Adv. Appl. Probab.* **14**, 392–433.

M. Brown (2004) Exploiting the waiting time paradox: applications of the renewal length transformation. *Ann. Appl. Probab.* **14** (to appear).

H. Bühlmann (1970) *Mathematical Methods in Risk Theory.* Springer–Verlag.

J.A. Buzacott and J.G. Shantikumar (1993) *Stochastic Models of Manufacturing Systems.* Prentice–Hall.

H. Carlsson and S. Wainger (1984) On the multi–dimensional renewal theorem. *Ann. Probab.* **11**, 143–157.

J.T. Chang and Y. Peres (1997) Ladder heights, Gaussian random walks and the Riemann zeta function. *Ann. Probab.* **25**, 787–802.

X. Chao, M. Miyazawa and M. Pinedo (1999) *Queueing Networks. Customers, Signals and Product Form Solutions.* Wiley.

F. Charlot, M. Ghidouche and M. Hamami (1978) Irreducibilite et recurrence au sens de Harris des "Temps d'attente" des files $GI/G/q$. *Z. Wahrscheinlichkeitsth. verw. Geb.* **43**, 187–203.

H. Chen and A. Mandelbaum (1990) Leontiff systems, RBV's and RBM's. In *Proceedings of the Imperial College Workshop on Applied Stochastic Processes* (M.H.A. Davis & R.J. Elliot, eds.). Gordon and Breach.

H. Chen and D. Yao (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization.* Springer–Verlag.

J. Choe and N. Shroff (1999) On the supremum distribution of integrated stationary Gaussian processes with negative drift. *Adv. Appl. Probab.* **31**, 135–157.

K.L. Chung (1967) *Markov Chains with Stationary Transition Probabilites* (2nd ed.). Springer–Verlag.

K.L. Chung (1974) *A Course in Probability Theory* (2nd ed.). Academic Press.

E. Çinlar (1972) Markov additive processes. II. *Z. Wahrscheinlichkeitsth. verw. Geb.* **24**, 93–121.

E. Çinlar (1975) *Introduction to Stochastic Processes.* Prentice–Hall.

J.W. Cohen (1982) *The Single Server Queue* (2nd ed.). North-Holland.

J.W. Cohen (1992) *Analysis of Random Walks.* IOS Press.

D.R. Cox (1962) *Renewal Theory.* Methuen.

D.R. Cox and W.L. Smith (1961) *Queues.* Methuen.

J.T. Cox and U. Rössler (1984) A duality relation for entrance and exit laws for Markov processes. *Stoch. Proc. Appl.* **16**, 141–156.

J.G. Dai (1995a) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limits. *Ann. Appl. Probab.* **5**, 49–77.

J.G. Dai (1995b) Stability of open queuing networks via fluid models. *Stochastic Networks* (F. Kelly and R.J. Williams, eds.), 71–90. Springer–Verlag.

D.J. Daley (1999) The Hurst parameter of a long–range dependent renewal process. *Ann. Probab.* **27**, 2035–2041.

D.J. Daley, A.Y. Kreinin and C.D. Trengrove (1992) Inequalities concerning the waiting time in single–server queues: a survey. *Queueing and Related Models* (U.N. Bhat ed.), 177–223. Clarendon Press.

D. Daley and T. Rolski (1984) A light traffic approximation for a single server queue. *Math. Opns. Res.* **9**, 624–628.

D. Daley and T. Rolski (1991) Light traffic approximations in queues. *Math. Opns. Res.* **16**, 57–71.

D. Daley and D. Vere–Jones (1988) *An Introduction to the Theory of Point Processes.* Springer–Verlag.

C.D. Daykin, T. Pentikäinen and E. Pesonen (1994) *Practical Risk Theory for Actuaries.* Chapman & Hall.

M.H.A. Davis (1993) *Markov Processes and Optimization.* Chapman & Hall.

A. de Acosta and P. Ney (1998) Large deviations lower bounds for arbitrary additive functionals of a Markov chain. *Ann. Probab.* **26**, 1660–1682.

K. Dȩbicki (2002) Ruin probabilities for integrated Gaussian processes. *Stoch. Proc. Appl.* **98**, 151–174.

K. Dȩbicki, Z. Michna and T. Rolski (1998) On the supremum from Gaussian processes over infinite horizon. *Probab. Math. Statist.* **18**, 83-100.

C. Dellacherie and P.A. Meyer (1975–93) *Probabilités et Potentiel.* Hermann.

A. Dembo and O. Zeitouni (1998) *Large Deviations Techniques and Applications.* Springer–Verlag.

J.H.A. de Smit (1995) Explicit Wiener–Hopf factorizations for the analysis of multi–dimensional queues. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 392–419. CRC Press.

R.A. Doney (1997) One–sided large deviations and renewal theorems in the case of infinite mean. *Probab. Th. Rel. Fields* **107**, 451–465.

E. van Doorn (1980) *Stochastic Monotonicity and Queueing Applications of Birth–Death Processes*. Lecture Notes in Statistics **4**. Springer.

P. Doukhan, G. Oppenheim and M.S. Taqqu, eds. (2003) *Long–Range Dependence: Theory and Applications.* Birkhäuser.

R.M. Dudley (1989) *Real Analysis and Probability.* Chapman & Hall.

N.G. Duffield and N. O'Connell (1995) Large deviations and overflow probabilities for the general single–server queue, with applications. *Math. Proc. Camb. Philos. Soc.* **118**, 363–374.

P. Dupuis and K. Ramanan (1998) A Skorokhod problem and large deviations analysis of a processor sharing model. *Queueing Systems* **28**, 109–124.

R. Durrett (1991) *Probability: Theory and Examples.* Wadsworth & Brooks/Cole.

R.J. Elliot, L. Aggoun and J.B. Moore (1995) *Hidden Markov Models. Estimation and Control.* Springer–Verlag.

M. El–Taha and S. Stidham Jr. (1999) *Sample–Path Analysis of Queueing Systems.* Kluwer.

P. Embrechts, C. Klüppelberg and T. Mikosch (1997) *Modelling Extremal Events for Finance and Insurance*. Springer–Verlag.

P. Embrechts and N. Veraverbeke (1982) Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics* **1**, 55–72

P. Erdös, W. Feller and H. Pollard (1949) A property of power series with positive coefficients. *Bull. Amer. Math. Soc.* **55**, 201–204.

S. Ethier and T. Kurtz (1986) *Markov Processes: Characterization and Convergence.* Wiley.

G. Fayolle, R. Iasnogorodski and V. Malyshev (1999) *Random Walks in the Quarter Plane. Algebraic Methods, Boundary Value Problems and Applications.* Springer–Verlag.

G. Fayolle, V. Malyshev and M. Menshikov (1995) *Topics in the Constructive Theory of Markov Chains*. Cambridge University Press.

G. Falin and J.G.C. Templeton (1997) *Retrial Queues*. Chapman & Hall.

W. Feller (1971) *An Introduction to Probability Theory and Its Applications* **II** (2nd ed.). Wiley.

P.A. Ferrari, H. Kesten, S. Martinez and P. Picco (1995) Existence of quasi–stationary distributions. A renewal dynamical approach. *Ann. Probab.* **23**, 501–521.

P.J. Fitzsimmons (1987) On the excursions of Markov processes in classical duality. *Prob. Th. Rel. Fields* **75**, 159–178.

L. Flatto (1997) The waiting time distribution for the random order service $M/M/1$ queue. *Ann. Appl. Probab.* **7**, 382–409.

R.D. Foley and D.R. McDonald (2001) Join the shortest queue; stability and exact asymptotics. *Ann. Appl. Probab.* **11**, 569–607.

S. Foss (1980) Approximation of multichannel queueing systems. *Siberian Math. J.* **21**, 132–140.

S. Foss and N. Chernova (2001) On optimality of FCFS discipline in multichannel queueing systems and networks. *Siberian Math. J.* **42**, 372–385.

S.G. Foss and V.V. Kalashnikov (1991) Regeneration and renovation in queues. *Queueing Systems* **8**, 211–223.

P. Franken, D. König, U. Arndt and V. Schmidt (1982) *Queues and Point Processes.* Wiley.

D. Freedman (1971) *Markov Chains.* Holden–Day.

B.E. Fristedt (1996) Intersections and limits of regenerative sets. *Random Discrete Structures* (D. Aldous and R. Pemantle, eds.), 121–151. Springer–Verlag.

C.D. Fuh and T. Lai (1998) Wald's equations, first passage times and moments of ladder variables in Markov random walks. *J. Appl. Probab.* **35**, 566–580.

H. Furrer, Z. Michna and A. Weron (1997) Stable Lévy motion approximation in collective risk theory. *Insurance: Mathematics and Economics* **20**, 97–114.

H.R. Gail, S.L. Hantler and B.A. Taylor (1996) Spectral analysis of $M/G/1$ and $GI/M/1$ Markov chains. *Adv. Appl. Probab.* **28**, 114–165.

D.P. Gaver and J.P. Lehoczky (1982) Channels that cooperatively service a data stream and voice messages. *IEEE Trans. Com.* **30**, 1153–1161.

E. Gelenbe and G. Pujolle (1998) *Introduction to Queueing Networks.* Wiley.

H.U. Gerber (1979). *An Introduction to Mathematical Risk Theory.* S.S. Huebner Foundation Monographs, University of Pennsylvania.

P. Glasserman and S.–G. Kou (1995a) Limits of first passage times to rare sets in regenerative processes. *Ann. Appl. Probab.* **5**, 424–445.

P. Glasserman and S.–G. Kou (1995b) Analysis of an importance sampling estimator for tandem queues. *ACM TOMACS* **5**, 22–42.

P.W. Glynn (1990) Diffusion approximations. In *Handbooks on OR & MS* **2** (D.P. Heyman & M.J. Sobel, eds.), 145–198. Elsevier.

P.W. Glynn and H. Thorisson (2001) Two–sided taboo limits for Markov processes and associated perfect simulation. *Stoch. Proc. Appl.* **91**, 1–20.

P.W. Glynn and W. Whitt (1986) A central limit version of $L = \lambda W$. *Queueing Systems* **2**, 191–215.

P.W. Glynn and W. Whitt (1989) Extensions of the queueing relations $L = \lambda W$ and $H = \lambda G$. *Opns. Res.* **37**, 634–644.

P.W. Glynn and W. Whitt (1993) Limit theorems for cumulative processes. *Stoch. Proc. Appl.* **47**, 299–314.

P.W. Glynn and W. Whitt (1994) Logarithmic asymptotics for steady–state tail probabilities in a single–server queue. In *Studies in Applied Probability* (J. Galambos and J. Gani, eds.). *J. Appl. Probab.* **31A**, 131–156.

B.W. Gnedenko and A.N. Kolmogorov (1954) *Limit Distributions for Sums of Independent Random Variables.* Addison–Wesley.

B. Gnedenko and I.N. Kovalenko (1989) *An Introduction to Queueing Theory* (2nd ed.). Birkhäuser [1st ed. published 1968 by Israel Program for Scientific Translations].

B.W. Gnedenko and D. König (1983/84) *Handbuch der Bedienungstheorie* **I–II**. Akademie–Verlag.

W.J. Gordon and G.F. Newell (1967) Closed queueing systems with exponential servers. *Opns. Res.* **15**, 254–265.

J. Grandell (1990) *Aspects of Risk Theory.* Springer–Verlag.

D. Gross and C.M. Harris (1985) *Fundamentals of Queueing Theory.* Wiley.

R. Grübel (1988) Harmonic renewal measures and the first positive sum. *J. London Math. Soc.* **38**, 179–192.

R. Grübel (1991) $G/G/1$ via FFT. Statistical Algorithm 265, *Applied Statistics* **40**, 355–365.

A. Gut (1988) *Stopped Random Walks. Theory and Applications.* Springer–Verlag.

P. Hall and C.C. Heyde (1980) *Martingale Limit Theory and Its Applications.* Academic Press.

T.E. Harris (1963) *The Theory of Branching Processes.* Springer–Verlag.

J.M. Harrison (1985) *Brownian Motion and Stochastic Flow Systems.* Wiley.

J.M. Harrison and M.I. Reiman (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* **9**, 302–308.

J.M. Harrison and S.I. Resnick (1977) The recurrence classification of risk and storage processes. *Math. Opns. Res.* **3**, 57-66.

J.M. Harrison and R. Williams (1987) Brownian models of open queueing networks with homogeneous customer populations. *Stochastics and Stochastic Reports* **22**, 77–115.

J.M. Harrison and R. Williams (1992) Brownian models of feedforward queueing networks: quasi–reversibility and product form solutions. *Ann. Appl. Probab.* **2**, 263–293.

C.R. Heathcote (1967) Complete exponential convergence and related topics. *J. Appl. Probab.* **4**, 1–40.

C.R. Heathcote and P. Winer (1969) An approximations for the moments of the waiting time. *Opns. Res.* **17**, 175–186.

D. Heath, S. Resnick & G. Samorodnitsky (1998) Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Opns. Res.* **23**, 145–165.

D. Heath, S. Resnick & G. Samorodnitsky (1999) How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Ann. Appl. Probab.* **9**, 352–375.

P. Heidelberger (1995) Fast simulation of rare events in queueing and reliability models. *ACM TOMACS* **6**, 43–85.

U. Hermann (1965) Ein Approximationssatz für Verteilungen Stationärer zufälliger Punktfolgen. *Math. Nachrichten* **30**, 377–381.

P. Heyman and M.J. Sobel (1982) *Stochastic Models in Operations Research.* McGraw–Hill.

A. Hordijk (2001) Comparisons of queues with different discrete–time arrival processes. *Probab. Th. Eng. Inf. Sc.* **15**, 1–14.

T. Huang and K. Sigman (1999) Steady–state asymptotics for tandem, split and match and other feedforward queues with heavy tailed service. *Queueing Systems* **33**, 233–259.

J. Hüssler and J. Piterbarg (1999) Extremes of a certain class of Gaussian processes. *Stoch. Proc. Appl.* **83**, 257–271.

I. Ignatiouk–Robert (2000) The large deviations of Jackson networks. *Ann. Appl. Probab.* **10**, 962–1001.

J.R. Jackson (1957) Networks of waiting lines. *Opns. Res.* **5**, 518–521.

J. Jacod and A.N. Shiryaev (1987) *Limit Theorems for Stochastic Processes.* Springer–Verlag.

J. Janssen and N. Limnios, eds. (1999) *Semi–Markov Models and Applications.* Kluwer.

P. Jelenkovic (1999) Subexponential loss rates in a $GI/G/1$ queue with applications. *Queueing Systems* **33**, 91–123.

P.R. Jelenkovic and P. Momcilovic (2002) Finite buffer queue with generalized processor sharing and heavy–tailed input. *Computer Network Journal* **40**, 433–443.

P.R. Jelenkovic and P. Momcilovic (2003) Large deviations analysis of subexponential waiting times in a processor sharing queue. *Math. Opns. Res.* (to appear).

J.L. Jensen (1995) *Saddlepoint Approximations.* Oxford University Press.

V.V. Kalashnikov (1994) *Mathematical Methods in Queueing Theory.* Kluwer.

V.V. Kalashnikov (1994) *Topics on Regenerative Processes.* Kluwer.

O. Kallenberg (1976) *Random Measures.* Akademie–Verlag.

S. Karlin and H.G. Taylor (1981) *A Second Course in Stochastic Processes.* Academic Press.

A. Karr (1975) Weak convergence of a sequence of Markov chains. *Z. Wahrscheinlichkeitsth. verw. Geb.* **33**, 41–48.

J. Keilson (1965) *Green's Function Methods in Probability Theory.* Griffin.

J. Keilson (1979) *Markov Chain Models — Rarity and Exponentiality.* Springer.

O. Kella and W. Whitt (1992) Useful martingales for stochastic storage processes with Lévy input. *J. Appl. Probab.* **29**, 396–403.

F.P. Kelly (1979) *Reversibility and Stochastic Networks.* Wiley.

F.P. Kelly (1983) Invariant measures and the Q–matrix. *Probability, Statistics and Analysis* (J.F.C. Kingman and G.E.H. Reuter, eds.), 143–160. Cambridge University Press.

F.P. Kelly (1991) Loss networks. *Ann. Appl. Probab.* **1**, 319–378.

F. Kelly and R.J. Williams, eds. (1995) *Stochastic Networks.* Springer–Verlag.

F. Kelly, S. Zachary and I. Ziedins, eds. (1996) *Stochastic Networks: Theory and Applications.* Oxford University Press.

J.G. Kemeny, J.L. Snell and A.W. Knapp (1976) *Denumerable Markov Chains* (2nd ed.). Springer–Verlag.

J.H.B. Kemperman (1961) *The Passage Problem for a Markov Chain.* University of Chicago Press.

J. Kennedy (1994) Understanding the Wiener–Hopf factorization for the simple random walk. *J. Appl. Probab.* **31**, 561–563.

H. Kesten (1995) A ratio limit theorem for (sub) Markov chains on $\{1, 2, \ldots\}$ with bounded jumps. *Adv. Appl. Probab.* **27**, 652–691.

A.Y. Khintchine (1960) *Mathematical Methods in the Theory of Queueing.* Griffin.

A.I. Khinchin (1949) *Mathematical Foundations of Statistical Mechanics.* Dover.

J. Kiefer and J. Wolfowitz (1955) On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78**, 1–18.

J. Kiefer and J. Wolfowitz (1956) On the characteristics of the general queueing process with applications to random walks. *Ann. Math. Statist.* **27**, 147–161.

H.S. Kim and N.B. Shroff (2001) On the asymptotic relationship between the overflow probability and the loss ratio. *Adv. Appl. Probab.* **33**, 836–863.

J.F.C. Kingman (1963) Ergodic properties of continuous time Markov processes and their discrete skeletons. *Proc. London Math. Soc.* **13**, 593–604.

J.F.C. Kingman (1972) *Regenerative Phenomena.* Wiley.

L. Kleinrock (1975/76) *Queueing Systems* **1–2**. Wiley.

S. Kotz, T.J. Kozubowski and K. Podgorski (2001) *The Laplace Distributions and Generalizations.* Birkhäuser.

I.N. Kovalenko (1995) Approximations of queues via small–parameter method. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 481–506. CRC Press.

U. Küchler and M. Sørensen (1997) *Exponential Families of Stochastic Processes.* Springer–Verlag.

H.J. Kushner (2001) *Heavy Traffic Analysis of Controlled Queueing and Communications Networks.* Springer–Verlag.

S.P. Lalley (1986) Renewal theorem for a class of stationary sequences. *Probab. Th. Rel. Fields* **72**, 195–213.

G. Latouche and P. Taylor, eds. (2000) *Advances in Algorithmic Methods for Stochastic Models.* Notable Publications, New Jersey.

G. Latouche and P. Taylor, eds. (2002) *Matrix–Analytic Methods. Theory and Applications.* World Scientific.

G. Latouche and V. Ramaswami (1993) A logarithmic reduction algorithm for quasi–birth–and–death processes. *J. Appl. Probab.* **30**, 650–674.

G. Latouche and V. Ramaswami (1999) *Introduction to Matrix Analytic Methods in Stochastic Modelling.* SIAM.

M.R. Leadbetter, G. Lindgren and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer–Verlag.

P. Le Gall (1962) *Les Systemes avec ou sans Attente et les Processus Stochastiques.* Dunod.

F. Le Gall and Y. Le Jan (1998) Branching processes in Lévy processes I–II. *Ann. Probab.* **26**, 213–252, 1407–1432.

D. Lindley (1952) The theory of a queue with a single server. *Proc. Cambr. Philos. Soc.* **48**, 277–289.

D. Lindley (1959) Discussion of a paper of C.B. Winsten. *J. Roy. Statist. Soc.* **B21**, 22–23.

T. Lindvall (1977) A probabilistic proof of Blackwell's renewal theorem. *Ann. Probab.* **5**, 482–485.

T. Lindvall (1992) *Lectures on the Coupling Method.* Wiley; reprinted by Dover 2002.

L. Lipsky (1992) *Queueing Theory: A Linear Algebraic Approach.* Macmillan.

V.I. Lotov (1996) On some boundary crossing problems for Gaussian random walks. *Ann. Probab.* **24**, 2154–2171.

R.M. Loynes (1962) The stability of a queue with non–independent interarrival and service times. *Proc. Camb. Philos. Soc.* **58**, 497–520.

R.B. Lund, S.P. Meyn and R.L. Tweedie (1996) Computable exponential convergence rates for stochastically ordered Markov processes. *Ann. Appl. Probab.* **6**, 218–237.

A. Martin–Löf (1979) *Statistical Mechanics and the Foundations of Thermodynamics*. Lecture Notes in Physics **101**. Springer–Verlag

L. Massoulie and A. Simonian (1999) Large buffer asymptotics for the queue with fractional Brownian input. *J. Appl. Probab.* **36**, 894–906.

K. Matthes, J. Kerstan and J. Mecke (1978) *Infinitely Divisible Point Processes.* Wiley.

I.L. MacDonald and W. Zucchini (1997) *Hidden Markov and Other Models for Discrete–Valued Time Series.* Chapman & Hall.

D. McDonald (1975) Renewal theorem and Markov chains. *Ann. Inst. Henri Poincare* **XI**, 187–197.

D. McDonald (1999) Asymptotics of first passage times for random walk in an orthant. *Ann. Appl. Probab.* **9**, 110–145.

B. Melamed and D. Yao (1995) The ASTA property. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 195–224. CRC Press.

S. Meyn and R.L. Tweedie (1993) *Markov Chains and Stochastic Stability.* Springer–Verlag.

T. Mikosch and C. Stărică (2003) Long–range dependence effects and ARCH modeling. *Long–Range Dependence: Theory and Applications* (P. Doukhan, G. Oppenheim and M.S. Taqqu, eds.), 149–168. Birkhäuser.

D.L. Minh and R.M. Sorli (1983) Simulating the $GI/G/1$ queue in heavy traffic. *Opns. Res* **31**, 966–971.

M. Miyazawa (1993) Insensitivity and product form decomposability of reallocatable GSMP. *Adv. Appl. Probab.* **25**, 415–437.

M. Miyazawa (1994) Rate conservation laws: A survey. *Queueing Systems* **15**, 1–58.

M. Miyazawa (2002) A Markov renewal approach to the asymptotic decay of the tail probabilities in risk and queueing processes. *Probab. Th. Eng. Inf. Sc.* **16**, 139–150.

M. Miyazawa (2003) Conjectures on decay rates of tail probabilities in generalized Jackson and batch movement networks. *J. Opns. Res. Soc. Japan* **41** (to appear).

M. Miyazawa (2004) Hitting probabilities in a Markov additive process with linear movements and upwards jumps: their applications to risk and queueing processes. *Ann. Appl. Probab.* **14** (to appear).

P.A.P. Moran (1959) *The Theory of Storage*. Methuen.

P.M. Morse (1958) *Queues, Inventories and Maintenance.* Wiley.

A. Müller and D. Stoyan (2002) *Comparison Methods for Stochastic Models and Risks.* Wiley.

S. Nahmias (2001) *Production and Operations Analysis* (4th ed.). Irwin.

S. Narayana and M.F. Neuts (1992) The two first moments of the counts for the Markovian arrival process. *Stochastic Models* **8**, 549–477.

M.F. Neuts (1969) The queue with Poisson input and general service times treated as a branching process. *Duke Math. J.* **36**, 215–232.

M.F. Neuts (1979) A versatile Markovian point process. *J. Appl. Probab.* **16**, 764–779.

M.F. Neuts (1981) *Matrix–Geometric Solutions in Stochastic Models.* Johns Hopkins University Press.

M.F. Neuts (1986) The caudal characteristic curve of queues. *Adv. Appl. Probab.* **16**, 221–254.

M.F. Neuts (1989) *Structured Markov Chains of the $M/G/1$ Type and Their Applications*. Marcel Dekker.

M.F. Neuts (1992) Models based on the Markovian arrival process. *IEICE Trans. Comm.* **E75–B**, 1255–1265.

J. Neveu (1961) Une generalization des processus a accroisances independantes. *Sem. Math. Abh. Hamburg.*

J. Neveu (1965) *Mathematical Foundations of the Calculus of Probability.* Holden–Day.

J. Neveu (1972) *Martingales a temps discret.* Dunod.

P. Ney and E. Nummelin (1987) Markov additive processes I–II. *Ann. Probab.* **15**, 561–592, 593–609.

I. Norros (2000) Queueing behaviour under fractional Brownian traffic. In *Self–Similar Network Traffic and Performance Evaluation* (K. Park and W. Willinger, eds.), 101–114. Wiley.

E. Nummelin (1984) *General Irreducible Markov Chains and Non–Negative Operators*. Cambridge University Press.

S. Orey (1971) *Lecture Notes on Limit Theorems for Markov Chain Transition Probabilities*. Van Nostrand Reinhold.

Z. Palmowski and T. Rolski (2002) A technique for exponential change of measure for Markov processes. *Bernoulli* **8**, 767–785.

K. Park and W. Willinger, eds. (2000) *Self–Similar Network Traffic and Performance Evaluation*. Wiley.

J. Paulsen and H.K. Gjessing (1997) Ruin theory with stochastic returns on investments. *Adv. Appl. Probab.* **29**, 965–985.

V.I. Piterbarg (2001) Large deviations of a storage process with fractional Brownian motion as input. *Extremes* **4**, 147–164.

J.W. Pitman (1974) Uniform rates of convergence for Markov chain transition probabilities. *Z. Wahrscheinlichkeitsth. verw. Geb.* **29**, 193–227.

J.W. Pitman (1986) Stationary excursions. *Seminaire de Probabilities XXI*, 289–302. Lecture Notes in Mathematics **1247**. Springer–Verlag

D. Pollard (1984) *Convergence of Stochastic Processes*. Springer–Verlag.

J.H. Pollard (1973) *Mathematical Models for the Growth of Human Populations.* Cambridge University Press.

N.U. Prabhu (1965) *Queues and Inventories.* Wiley.

N.U. Prabhu (1980) *Stochastic Storage Processes. Queues, Insurance Risk and Dams.* Springer–Verlag; 2nd ed. published 1998.

J. Preater (2002) On the severity of $M/M/\infty$ congested episodes. *J. Appl. Probab.* **29**, 228–230.

S.H. Preston, P. Heuveline and M. Guillot (2001) *Demography. Measuring and Modeling Population Processes.* Blackwell.

P. Protter (1990) *Stochastic Integration and Differential Equations.* Springer–Verlag.

M.L. Puterman (1994) *Markov Decision Processes.* Wiley.

M. Reiman (1984) Open queueing networks in heavy traffic. *Maths. Opns. Res.* **9**, 441–458.

D. Revuz (1984) *Markov Chains.* North–Holland.

D. Revuz and M. Yor (1999) *Continuous Martingales and Brownian Motion* (3rd ed.). Springer–Verlag.

P. Robert (2000) *Réseaux et files d'attente: méthodes probabilistes.* Springer–Verlag (English ed. to appear 2003).

L.C.G. Rogers (1994) Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains. *Ann. Appl. Probab.* **4**, 390–413.

L.C.G. Rogers and D. Williams (1994) *Diffusions, Markov Processes and Martingales* **1–2**. Cambridge University Press.

T. Rolski (1981) Queues with non–stationary input streams: Ross's conjecture. *Adv. Appl. Probab.* **13**, 603–618.

T. Rolski, H. Schmidli, V. Schmidt and J.L. Teugels (1999) *Stochastic Processes for Insurance and Finance.* Wiley.

J. Rosenthal (1995) Convergence rates for Markov chains. *SIAM Review* **37**, 387–405.

M. Rougham and C.E.M. Pearce (2002) Martingale methods for analyzing single–server queues. *Queueing Systems* **41**, 205–240

R.Y. Rubinstein and B. Melamed (1998) *Classical and Modern Simulation.* Wiley.

M. Rudemo (1973) Point processes generated by transitions of Markov chains. *Adv. Appl. Probab.* **5**, 262–286.

R. Ryan and K. Sigman (2000) Continuous-time monotone stochastic recursions and duality. *Adv. Appl. Probab.* **32**, 426–445.

T. Rydén (2000) Statistical estimation for Markov–modulated Poisson processes and Markovian arrival processes. *Advances in Algorithmic Methods for Stochastic Models* (G. Latouche and P. Taylor, eds.), 329–350. Notable Publications, New Jersey.

T.L. Saaty (1961) *Elements of Queueing Theory.* McGraw–Hill.

J.S. Sadowsky and W. Szpankowsky (1995) The probability of large queue lengths and waiting times in a heterogeneous multiserver queue I-II. *Adv. Appl. Probab.* **27**, 532–566, 567–583.

P. Salminen and I. Norros (2001) On busy periods of the unbounded Brownian storage. *Queueing Systems* **39**, 317–333.

L. Saloff–Coste (1996) Lectures on finite Markov chains. *Seminaire de Probabilities XXVI*, 301–413. Lecture Notes in Mathematics **1665**. Springer–Verlag.

G. Samorodnitsky, D. Heath and S. Resnick (1997) Patterns of buffer overflow in a class of queues with long memory in the input stream. *Ann. Appl. Probab.* **7**, 1021–1057.

G. Samorodnitsky and M.S. Taqqu (1994) *Stable Non–Gaussian Random Processes: Stochastic Models with Infinite Variance.* Chapman & Hall.

K. Sato (1999) *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press.

H.H. Schaefer (1970) *Topological Vector Spaces.* Springer–Verlag.

R. Schassberger (1973) *Warteschlangen.* Springer–Verlag.

A. Scheller–Wolf (2003) Necessary and sufficient for delay moment for multiserver queues: why s slow servers is better than one fast for heavy–tailed systems. *Opns. Res.* (to appear).

A. Scheller–Wolf and K. Sigman (1997) Delay moment for FIFO $GI/G/s$ queues. *Queueing Systems* **25**, 97–95.

E. Seneta (1994) *Non–Negative Matrices and Markov Chains.* Springer–Verlag.

B. Sengupta (1989) Markov processes whose steady–state distribution is matrix–exponential with an application to the $GI/PH/1$ queue. *Adv. Appl. Probab.* **21**, 159–180.

L.I. Sennot (1999) *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley.

R. Serfozo (1999) *Introduction to Stochastic Networks.* Springer–Verlag.

M. Shaked and J.G. Shantikumar (1994) *Stochastic Orders and Their Applications.* Academic Press.

M. Shalmon (1988) Analysis of the $GI/G/1$ queue and its variations via the LCFS preemptive resume discipline and its random walk interpretation. *Probab. Th. Eng. Inf. Sc.* **2**, 215–230.

A.N. Shiryaev (1996) *Probability.* Springer–Verlag.

A. Shwarz and A. Weiss (1995) *Large Deviations for Performance Analysis.* Chapman & Hall.

D. Siegmund (1976a) The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Ann. Probab.* **4**, 914–924.

D. Siegmund (1976b) Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4**, 673–684.

D. Siegmund (1979) Corrected diffusion approximations in certain random walk problems. *Adv. Appl. Probab.* **11**, 701–719.

D. Siegmund (1985) *Sequential Analysis.* Springer–Verlag.

K. Sigman (1988) Queues as Harris recurrent Markov chains. *Queueing Systems* **3**, 179–198.

K. Sigman (1992) Light traffic for workload in queues. *Queueing Systems* **11**, 429–442.

K. Sigman (1995) *Stationary Marked Point Processes: An Intuitive Approach.* Chapman & Hall.

E.A. Silver, D.F. Pyke and R. Peterson (1998) *Inventory Management and Production Planning and Scheduling* (3rd ed.). Wiley.

P.J. Smith (1995) A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician* **49**, 217–218.

W.L. Smith (1953) Distribution of queueing times. *Proc. Roy. Soc.* **49**, 449–461.

W.L. Smith (1955) Regenerative stochastic processes. *Proc. Roy. Soc.*, Ser. A, **232**, 6–31.

F. Spitzer (1976) *Principles of Random Walk* (2nd ed.). Springer–Verlag

W. Stadje (1993) A new look at the Moran dam. *J. Appl. Probab.* **30**, 489–495.

C. Stone (1966) On absolutely continuous distributions and renewal theory. *Ann. Math. Statist.* **37**, 271–275.

D. Stoyan (1983) *Comparison Methods for Queues and Other Stochastic Models* (D.J. Daley, ed.). Wiley.

B. Sundt (1993) *An Introduction to Non–Life Insurance Mathematics.* Verlag Versicherungswirtschaft e.V., Karlsruhe.

B. Sundt and J.L. Teugels (1995/97) Ruin estimates under interest force I–II. *Insurance: Mathematics and Economics* **16**, 7–22; *ibid.* **19**, 85–94.

B. Sundt and J.L. Teugels (1997) The adjustment coefficient in ruin estimates under interest force. *Insurance: Mathematics and Economics* **19**, 85–94.

B. Sundt and J.L. Teugels, eds. (2004) *Encyclopedia of Actuarial Sciences* (in production). Wiley.

R. Syski (1960) *Introduction to Congestion Theory in Telephone Systems.* Oliver and Boyd.

R. Szekli (1995) *Stohastic Orderings and Dependence in Applied Probability.* Lecture Notes in Statistics **97**. Springer–Verlag.

S. Täcklind (1942) Sur le risque dans les jeuxs inequitables. *Skand. Aktuar Tidsskr.* **1942**, 1–42.

L. Takács (1962) *Introduction to the Theory of Queues.* Oxford University Press.

H. Tanaka (1979) Stochastic equations for diffusion processes in a bounded region. *Hiroshima Math. J.* **9**, 163–177.

J.L. Teugels (1977) On the rate of convergence of a compound Poisson process. *Bull. Soc. Math. Belgique* **39**, 205–216.

J.L. Teugels (1982) Estimation of ruin probabilities. *Insurance: Mathematics and Economics* **1**, 163–175.

H. Thorisson (2000) *Coupling, Stationarity and Regeneration.* Springer–Verlag.

H. Tijms (1994) *Stochastic Models — An Algorithmic Approach.* Wiley.

N. Veraverbeke and J.L. Teugels (1975/76) The exponential rate of convergence of the distribution of the maximum of a random walk. *J. Appl. Probab.* **12**, 279–288; *ibid.* **13**, 733–740.

A. Wald (1947) *Sequential Analysis.* Wiley.

J. Walrand (1988) *An Introduction to Queueing Networks.* Prentice–Hall.

W. Whitt (1989) An interpolation approximation for the mean workload in a $GI/G/1$ queue. *Opns. Res.* **37**, 936–952.

W. Whitt (2000) The impact of a heavy–tailed service–time distribution upon the $M/GI/s$ waiting–time distribution. *Queueing Systems* **36**, 71–87.

W. Whitt (2002) *Stochastic–Process Limits.* Springer–Verlag.

P. Whittle (1986) *Systems in Stochastic Equilibrium.* Wiley.

R.J. Williams (1998) Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems* **30**, 27–88.

R.W. Wolff (1982) Poisson Arrivals See Time Averages. *Opns. Res.* **30**, 223–231.

R.W. Wolff (1987) Upper bounds on work in system for multi–channel queues. *J. Appl. Probab.* **14**, 547–551.

R.W. Wolff (1989) *Stochastic Modeling and the Theory of Queues.* Prentice–Hall.

M. Woodroofe (1982) *Nonlinear Renewal Theory in Sequential Analysis.* SIAM.

S.F. Yashkov (1987) Processor–sharing queues: some progress in analysis. *Queueing Systems* **2**, 1–17.

S.F. Yashkov (1992) Mathematical problems in the theory of shared–processor systems. *Journal of Soviet Mathematics* **58**, 101–147.

C.–H. Zhang (1988) A nonlinear renewal theory. *Ann. Probab.* **16**, 793–824.

P.H. Zipkin (2000) *Foundations of Inventory Management.* McGraw–Hill.

A.P. Zwart, S.C. Borst and M. Mandjes (2003) Exact asymptotics for fluid queues fed by multiple heavy–tailed On–Off flows. *Ann. Appl. Probab.* **13**.

A.P. Zwart and O.J. Boxma (2000) Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* **35**, 141–166.

Index

absorbtion, 6, 16, 38, 39, 83, 261, 263
arrival theorem, 133–135
asynchronous transfer mode, 307
atom, 16, 383

backward differential equation, 48–50, 54
backward recurrence time, 8, 14, 15, 38, 140–143, 150–156, 170, 172, 348, 405
balking, 72
BCMP network, 135
Bessel function, 98–103, 105–107
birth–death process, 46–48, 50, 59, 71–80, 114, 116, 127, 131, 182, 307, 390
Blumenthal's 0–1 law, 37
Boltzmann's law, 376
bottleneck behaviour, 121
branching process, 46, 113, 284
Brownian motion, 38, 96–97, 174, 244, 246, 248, 258–260, 288, 310, 311, 362, 365
  reflected, 97, 136, 174, 258–260
    between two barriers, 258, 394
busy cycle, 103, 105, 107, 168, 220, 267, 272, 291, 298–301
busy period, 102–105, 107, 110, 220, 267, 273, 284, 300, 327

cadlag, *see* Skorokhod's D–space
Campbell's formula, 219
central limit theory, 30, 55, 105–106, 161–162, 178–179, 248, 267, 301, 313, 355, 363, 367–368, 404
Cesaro average, 18, 54, 69, 177–179, 203, 278, 313–314, 385
change of measure, 177, 358–361
Chapman–Kolmogorov equations, 33, 39, 261
characteristic triplet, 246, 310
Chebycheff's covariance inequality, 272
Chernoff bound, 357
Choquet–Deny lemma, 158
CLT, *see* central limit theory
communicating states, 5–6
compensation, 246
competing risks, 45
component, 186, 194, 198
continuity of stationary distributions, 75, 87, 284–286
convergence rate, 19, 164, 191

exponential, 18, 19, 22, 24, 26, 107, 194–197, 356–357
convex conjugate, 355
convex order, 284, 336–338, 411
increasing, 336–339, 411
coupling, 11, 189–197, 205–206, 253, 342, 348, 402, 411
ϵ–, 192–194
broad sense, 189, 190
inequality, 190–191, 414
maximal, 197
narrow sense, 190
Ornstein, 227
shift, 197
Cox process, 217, 337, 410
Coxian distributions, 85
Cramér's condition (C), 189, 355
Cramér–Lundberg asymptotics, 68, 165, 365–369, 371, 373, 378–379, 395, 402, 404, 411
Croft–Kingman lemma, 415
cumulant, 106, 248, 312, 352, 414
cumulative process, 178–179, 313
current life, 151–153, 173, 213
customer class, 113, 125–134
cycle maximum, 298–301, 368, 395
cyclic class, 17, 28, 204

dam, *see* storage process
finite, 390, 394
infinitely deep, 391
Moran, 395
demography, *see* population model
departure process, 115–117, 122, 125
detailed balance, 114
diagonalization, 49, 56, 89–90, 242, 329–331, 335
diffusion, 263, 361
diffusion approximation, *see* functional CLT
corrected, 369–373
directly Riemann integrable, 154–158, 160, 163, 170, 187, 209
discrete skeleton, 51, 54, 57, 106, 415
distributional Little's law, 277
doubly exponential distribution, 92, 96, 109, 230
doubly stochastic, 15
duality, 260–264, 392–396, 399–401

EOQ model, 396
ergodic theorem, 19, 219
for Markov chains, 17, 202
Erlang distribution, 62, 81, 84, 90, 153, 349, 351, 389
excursion, 174, 241, 320, 321, 394, 406
exit vector, 83
explosion, 40–41, 43–44, 47–48, 53, 73, 117
exponential change of measure, 99, 162–165, 352–379
exponential ergodicity, *see* convergence rate, exponential
exponential family, 352–354, 364, 369
extinction, 7, 24
extreme value theory, 180, 183–185, 298, 369

FCFS, 63, 108–109, 340–343
FIFO, *see* FCFS
filtration, 37
finite capacity model, 389–396
fixed discontinuity, 409
fluid limit, 136
fluid model, 300, 301, 306–308, 328–330, 334–335, 394
forward differential equation, 48–50
forward recurrence time, 9, 14, 38, 88, 140–143, 150–156, 172, 188, 193–196, 224, 276, 405
Foster's criteria, 24, 282
Fréchet distribution, 184, 298
fractional Brownian motion, 301, 369
functional CLT, 267, 288–290
fundamental matrix, 31, 55

$\boldsymbol{G}$–matrix, 322–324, 328, 331–334
Gärtner–Ellis theorem, 357, 369
Gaussian process, 219, 301, 369
Gaussian random walk, 243
generator, 38, 263, 361, 400
geometrical ergodicity, *see* convergence rate, exponential
geometrical trial, 18, 26, 200, 346, 412
Girsanov's theorem, 361
Gumbel distribution, 184, 298, 369

$H = \lambda G$, 279
harmonic function, 20–21, 24, 25, 30

Harris chain, 198–206, 325, 345–348, 383, 384
heavy tails, 113, 165–167, 184, 295–301, 344, 354, 369, 387, 402, 404–406, 412
heavy traffic, 136, 255, 286–290, 295, 366, 369–373
heterogeneous servers, 80, 328
Hewitt–Savage 0–1 law, 158, 225, 226
hidden Markov model, 311
h–transform, 24
Hurst parameter, 301
hydrodynamical limit, 136
hyperexponential distribution, 62, 81, 153, 389

idle period, 103, 267, 271, 273, 274
imbedded Markov chain, *see* Markov chain, imbedded
importance sampling, 374, 376
increasing convex order, 336–339, 411
infinitely divisible distribution, 244–250
insensitivity, 78, 123–124, 128, 130
instantaneous state, 39, 44
insurance mathematics, 401
insurance risk, *see* ruin probability
integral equation
  Lindley, 94, 96, 109, 229, 268
  Lotka, 143, 210
  renewal, *see* renewal equation
  Volterra, 387
  Wiener–Hopf, 96, 229
integration by parts, 251, 255, 274
intensity, 40, 44–50, 304, 337, 410
intensity matrix, 38
intensity measure, 88, 152, 217, 246, 411
invariant σ–field, 19, 197, 219
inventory model, 61, 63, 396–398
inverse Gaussian distribution, 85, 288, 362–364, 367
irreducibility, 6, 13, 26, 50, 91
iterative algorithm, 241–242, 321, 323, 331–335
Ito's formula, 361

Jackson network, 117–123, 131, 134–135
  closed, 120–122, 134–135
  generalized, 136

Kelly network, 135
Kelly's lemma, 59, 124, 127, 129
Kendall classification, 61–63
Kiefer–Wolfowitz vector, 341
Kolmogorov consistency, 12, 33, 247, 252, 358, 407
Kolmogorov's loop criterion, 59
Krein–Rutman theorem, 29
Kronecker product, 324, 327

$\ell = \lambda w$, *see* Little's formula
Lévy exponent, 104, 107, 245–250
Lévy measure, 246
Lévy process, 96–97, 104, 229, 244–250, 253–260, 284, 309, 360–362, 364–366, 381, 387, 392
  characteristic triplet of, 246
  reflected, 96–97, 253–260
  spectrally positive, 250, 254, 256
Lévy–Khintchine representation, 245
ladder epoch, 220–223, 229–232
ladder height distribution, 220–222, 224–225, 227–243, 268, 272, 292, 296, 314, 322, 328, 357, 366–367, 405
ladder point, 223
Laplace distribution, *see* doubly exponential distribution
Laplace transform, 47, 83–85, 87, 105, 107, 113, 146, 149, 232, 257, 259, 403–404, 414
  inversion, 218, 239, 373, 414
large deviations, 136, 355–357, 369, 375, 376
lattice or nonlattice, 153–157, 170, 186, 208, 222, 225, 273, 354, 413
  strongly nonlattice, *see* Cramér's condition (C)
LCFS, 63, 109–110, 130, 284
lead time, 397
length–biasing, 177
LIFO, *see* LCFS
light tails, 184, 352–379, 402, 404
light traffic, 290–295

likelihood ratio, 177, 358, 359
Lindley process, 23, 92–97, 220, 237, 250, 260, 262, 267, 280, 298, 336, 341, 344
linear progression algorithm, 332
Little's formula, 71, 276, 278, 284
local balance, 58
local time, 251, 256–260, 310, 393
logarithmic asymptotics, 357, 369, 375
logarithmic reduction algorithm, 333
lognormal distribution, 412
long–range dependence, 301
long–tailed distribution, *see* heavy tails
Lorden's inequality, 160
loss system, 67, 78–80, 123–124, 307, 395
Loynes's lemma, 252–253, 255, 260, 263, 336, 400
Lundberg equation, 165, 238, 354, 365
Lundberg's inequality, 365
Lyapounov function, *see* test function

M–matrix, 56
machine repair problem, 79
Malthusian parameter, 163
manufacturing system, 71, 135
MAP, *see* Markov additive process *or* Markovian arrival process
Markov additive process, 229, 302, 309–317, 324, 328, 330, 376–379, 394, 400
Markov chain, 3–38, 57, 181, 191–192, 264, 286, 361
  ergodic, 17, 18, 25, 74, 191–192
  Harris, *see* Harris chain
  imbedded, 40, 50, 70, 93, 236, 279–284, 318, 322, 349–351, 384
  inhomogeneous, 57
  Monte Carlo, 88
  on a subset, 14, 16, 75, 201–202, 212, 319, 320, 389–391
Markov process, 7, 32–59, 139, 192, 198–206, 260–261, 309, 348, 359–360
  ergodic, 52, 58, 321
  jump, 39–59
  minimal, 41–45, 54
  piecewise deterministic, 387
  stochastically monotone, 260
Markov–modulated Poisson process, 213, 302, 306
Markov–modulation, 264, 302, 307, 322, 324, 336
Markovian arrival process, 302–306, 308, 322
martingale, 20–24, 26, 31, 38, 56, 104–105, 107, 158, 182, 235, 254–260, 285, 289, 290, 312, 329–331, 335, 358–359, 362, 411, 415
  backward, 247, 249, 254, 285, 289
  Dynkin, 38, 56
  Kella–Whitt, 255–260
  likelihood ratio, 358, 362, 376
  local, 38, 255, 361, 410
  Wald, 104, 181, 235, 255, 258, 312, 329–331, 362, 376, 392, 394, 403
matrix–geometric method, 243
mixing, 179
mixture, 81, 84
Mogulskii's theorem, 357, 376
Monte Carlo method, *see* simulation
Moran's model for the dam, 395
multiplicative functional, 359, 361

natural scale, 182
negative customers, 71
netput, 93, 136, 236, 240, 307, 308, 394, 399
number in system, *see* queue length

oblique reflection, 136
occupation measure, 12, 224, 225, 228, 299, 315, 325
on–off model, 301, 307
Orey's *C*–set theorem, 205
Ornstein–Uhlenbeck process, 206
overshoot distribution, 42, 87, 195, 221, 224, 233, 239, 294, 364–366, 370–372, 397, 404

Palm theory, 108, 213–219, 303, 327, 337

Palm–Jacobæus formula, 80
Pareto distribution, 404
PASTA, 109, 133, 178, 218, 219, 385
periodicity, 6, 9, 17, 19, 26, 28, 51, 54, 171, 202, 204, 206, 219, 268, 346
Perron–Frobenius theory, 25–29, 55, 83, 210, 305, 312
phase generator, 83
phase representation, 83, 242
  reversed, 91, 124, 130
phase–type distribution, 80–92, 123, 128, 148, 234, 239–243, 260, 303, 306–308, 324, 326, 327, 389, 394
Poisson's equation, 30–31, 55
Polish space, 33, 407–408, 410
Pollaczeck–Khinchine formula, 238, 257, 281, 296, 387, 405
population model, 7, 24, 46, 76, 80, 117, 143, 163, 210
preemptive, 65, 284
  non, 65, 109
priority class, 64
processor sharing, *see* PS
product–form, 132, 135
Prohorov's theorem, 408
PS, 63, 80, 111–113, 118, 130
  generalized, 113

QBD, *see* quasi birth–death process
quasi birth–death process, 323–324, 328, 332–334
quasi–reversibility, 125–134
quasi–stationary distribution, 16, 19
queue
  $D/G/1$, 337
  $D/M/1$, 295
  $GI/D/1$, 337
  $GI/G/1$, 92, 168, 266–301, 337, 341, 353, 356, 366, 368, 399
  $GI/G/s$, 64, 300, 340–348, 369
  $GI/M/1$, 70, 93, 236, 238, 279–281, 318, 366
  $GI/M/1$ type, 319–322, 328, 331
  $GI/M/s$, 348–351
  $GI/PH/1$, 239–243, 319, 326, 328
  $GI/PH/s$, 327, 351
  $M/D/1$, 295
  $M/E_k/1$, 97
  $M/G/1$, 70, 95, 97, 109, 236–238, 275, 281–284, 296, 301, 318, 336, 337, 367, 381, 387, 394, 399, 402
  $M/G/1$ type, 322–323, 328, 331
  $M/G/\infty$, 115
  $M/M/1$, 46, 50, 75–76, 95, 96, 98–113, 176, 181, 230, 259, 260, 354
    finite, 77, 390
    Markov–modulated, 322, 324
    multiclass, 127, 130
  $M/M/\infty$, 76–77, 182
  $M/M/m$, 77
  $M/PH/1$, 130, 242, 394
  $M/PH/\cdot$, 128–130
  $M/PH/\infty$, 130
  $MArP/G/1$, 322
  $MArP/PH/1$, 308, 324, 328
  bottleneck, 121
  doubly infinite, 98, 103, 107, 110
  fork–join, 71
  join–the–shortest, 71
  Kendall classification, 61–63
  many–server, *see* queue, $GI/\cdot/s$
  retrial, 71
  split–and–match, 71
  symmetric, 128–130
  tandem, 115–117
  with vacation, 71
queue discipline, 62–65, 71, 108–113, 340–343, 402
queue length, 65, 71, 98–107, 268, 276–279, 281, 290, 291, 349
queueing network, 114–136, 255, 290, 295, 300
  BCMP, 135
  Kelly, 135

$\boldsymbol{R}$–matrix, 319–321, 323–324, 328, 331–334
Radon measure, 383, 408, 410
random order service, *see* SIRO
random walk, xi, 15, 93, 220–243, 262, 266, 269, 272, 285, 287, 291, 296, 298, 323, 358, 360–362, 365–373, 392, 406
  Bernoulli, 15, 72, 98

maxima, 94, 109, 220–236, 242, 260, 269, 296
oscillating, 224
rare event, 179–185, 373–376
rate conservation, 150, 218, 219, 274, 276, 278, 279, 281, 283, 351, 383
rational Laplace transform, 83, 85, 87, 88, 238–239
recurrence, 5–6, 12, 13, 21, 44, 50, 51, 68, 72, 226, 316, 318, 319, 382, 384
Harris, *see* Harris chain
null, 6, 13, 16, 18, 19, 22, 68, 95, 153, 197, 203, 282, 383
of random walks, 226
positive, 6, 13–17, 21, 316, 318–321, 383–386, 388, 389
recurrent classes, 6, 14
reflection, 38, 92–97, 100, 136, 174, 250–260, 262, 307, 314
between two barriers, 258, 263, 392–396
oblique, 136, 255
reflection principle, 97, 100, 363
regeneration set, 199
regenerative phenomena, 172
regenerative process, 168–185, 188, 196, 198–206, 216, 227, 266, 268, 272, 274, 275, 277, 300, 344–348, 383, 390, 398
regenerative set, 172
regenerative simulation, 174–176
regular conditional distribution, 407
regular variation, *see* heavy tails
relaxation time, 105–107, 356–357
reneging, 72, 80
renewal density, 88, 147, 148
renewal equation, 143–150, 162–167, 208, 404
defective, 162–167
excessive, 162–167
Markov, 208–211
renewal function, 146–149, 293, 397
renewal measure, 143–150, 186–188, 224, 225, 273, 357
renewal process, 7–11, 14, 15, 62, 124, 138–143, 172–174, 193–196, 207, 212, 223, 301–303, 346, 405
alternating, 173–174, 301
imbedded, 169, 170, 178
stationary, 150–153
terminating, 10, 138, 223, 345
thinned, 293
transient, *see* renewal process, terminating
renewal sequence, 8–11, 17, 236, 267
renewal theorem, 89, 153–159, 193–196
Blackwell's, 155–159, 194, 208, 273, 297
elementary, 140, 178, 227
key, 155–156, 170, 187, 209–211, 300
defective, 11, 209–211
discrete, 9
excessive, 209–211
renewal theory, 7–11, 38, 88–92, 138–167, 193–196, 268
defective, 10, 148–149, 209–211
discrete, 7–11, 227
for dependent sequences, 143
for random walks, 143, 225–227, 229, 299
Markov, 143, 206–213, 280, 283, 303, 306, 310, 319, 326
multivariate, 143
with infinite mean, 143
renovating events, 172
residual service time, 276
Reuter's explosion condition, 47, 73
reversibility, 56–59, 114–117
Rice's formula, 219
road traffic model, 23, 95, 139, 142, 147, 207
roots, 234, 239, 242, 329–331, 392
Ross conjecture, 336–339
Rouché's theorem, 243
Round Robin, 64
ruin probability, 144–146, 149, 150, 165, 260–264, 328–330, 334, 365, 369, 399–406
two–sided, 329, 330, 364
ruin time, 402–406

(s, S) model, 397

saddlepoint, 355–357, 374–376
self–similarity, 301
semi–group, 33, 39, 413
semi–Markov, *see* renewal theory, Markov
Sengupta model, 324–328
sequential analysis, 364
shot–noise process, 147, 218
Siegmund duality, 260–262, 264, 400
simulation, 174–176, 272, 373–376
SIRO, 63, 110–111, 113
skip–free, 71, 114, 235–238, 319, 393
Skorokhod problem, 251, 255, 258, 396
Skorokhod's D–space, xii, 33, 169, 177, 245, 310, 358, 407, 409–410
  with finite lifelengths, xii, 169, 180, 403, 410
sojourn time, 65, 71, 111–113, 268
spectral method, 335
Spitzer's identity, 230, 356
Spitzer–Baxter identities, 229–233
spread–out distribution, 186–189, 194–197, 273, 347, 354, 371
squared coefficient of variation, 68, 81, 87
stability, 135
stable process, 250, 290, 364
Stam's lemma, 369, 371
state–space collapse, 136
stationarity, 11–16, 50–55, 90, 410
stationary distribution, 11, 14, 23, 25, 50, 52, 74, 89, 384
stationary measure, 11–16, 18, 51–52, 73–74, 200–203, 284, 383
  –super, 13
statistics, 86, 87, 306, 311, 352, 354, 364
steady state, xi, 68–69
stochastic integral, 255
stochastic recursion, 261–264, 392, 395
  Kiefer–Wolfowitz, 341
  Lindley, *see* Lindley process
stochastical ordering, 153, 190, 260, 342, 411, 412
Stone's decomposition, 186–189, 196, 273
stopping time, 4, 12, 34–37, 169, 199–200, 258, 315, 359, 362, 414
  randomized, 199–200, 415
storage equation, 381, 387
storage process, 380–392, 394–396, 400
Strassen's theorem, 336, 411
strong Markov property, 4, 34–35, 40, 96, 97, 141, 142, 226, 250, 253, 276, 381, 386
strong ratio property, 19
strongly nonlattice, *see* Cramér's condition (C)
subexponential distribution, *see* heavy tails
subordinator, 172, 250, 254, 364
substochastic matrix, 26, 316
symmetric queue, 128–130

taboo probability, 13
tail σ–field, 19, 197
test function, 20–24, 107, 206
throughput, 119, 132
time dependence, 48–50, 69, 98–107, 259, 367, 373, 402–406
time reversion, 56–59, 90–91, 114–117, 123, 124, 126, 127, 129, 222, 252–253, 314, 400
total variation, 19, 94, 171, 188–190, 202, 253, 274, 291–295, 345, 413
traffic equations, 119–120, 132
traffic intensity, 68, 116
transience, 5–6, 14, 16, 20, 22, 50, 51, 68, 72, 211, 226, 382, 384
  of random walks, 226
transient properties, *see* time dependence
two–barrier problem, 263, 264, 329, 330, 364, 389–396, 401
type I counter, 142
type II counter, 149

uniformization, 55, 321

vacation, 71
vague convergence, 156–158, 407, 408
virtual waiting time, *see* workload

waiting time, 65, 92, 108–113, 237, 242, 266–272, 274, 279, 281, 284–301, 326, 328, 336, 351, 356, 369, 399
waiting time paradox, 152, 213
Wald martingale, *see* martingale, Wald
Wald's identity, 95, 157, 160, 174, 180, 225, 258, 273, 313, 364, 414, 415
  second moment, 162, 415
Weibull distribution, 184, 298, 405, 412
Wiener–Hopf factorization, 145, 227–243, 272, 311, 314–316, 335, 371, 373
workload, 65, 97, 108–109, 236, 237, 260, 268, 272–276, 279, 281, 291, 308, 337, 341, 381, 387, 394, 399

Yaglom limit, 16

Z–matrix, 56

Applications of Mathematics

(continued from page ii)