A First Course in Stochastic Models

Henk C. Tijms

WILEY

A First Course in Stochastic Models

Henk C. Tijms
Vrije Universiteit, Amsterdam, The Netherlands

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Tijms, H. C.
A first course in stochastic models / Henk C. Tijms.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-49880-7 (acid-free paper)—ISBN 0-471-49881-5 (pbk. : acid-free paper)
1. Stochastic processes. I. Title.

QA274.T46 2003
519.2′3—dc21

2002193371

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-471-49880-7 (Cloth)
ISBN 0-471-49881-5 (Paper)

Typeset in 10/12pt Times from LaTeX files supplied by the author, by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by T J International Ltd, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface

The teaching of applied probability needs a fresh approach. The field of applied probability has changed profoundly in the past twenty years and yet the textbooks in use today do not fully reflect the changes. The development of computational methods has greatly contributed to a better understanding of the theory. It is my conviction that theory is better understood when the algorithms that solve the problems the theory addresses are presented at the same time. This textbook tries to recognize what the computer can do without letting the theory be dominated by the computational tools. In some ways, the book is a successor of my earlier book *Stochastic Modeling and Analysis*. However, the set-up of the present text is completely different. The theory has a more central place and provides a framework in which the applications fit. Without a solid basis in theory, no applications can be solved. The book is intended as a first introduction to stochastic models for senior undergraduate students in computer science, engineering, statistics and operations research, among others. Readers of this book are assumed to be familiar with the elementary theory of probability.

I am grateful to my academic colleagues Richard Boucherie, Avi Mandelbaum, Rein Nobel and Rien van Veldhuizen for their helpful comments, and to my students Gaya Branderhorst, Ton Dieker, Borus Jungbacker and Sanne Zwart for their detailed checking of substantial sections of the manuscript. Julian Rampelmann and Gloria Wirz-Wagenaar were helpful in transcribing my handwritten notes into a nice Latex manuscript.

Finally, users of the book can find supporting educational software for Markov chains and queues on my website http://staff.feweb.vu.nl/tijms.

CHAPTER 1

The Poisson Process and Related Processes

1.0 INTRODUCTION

The Poisson process is a counting process that counts the number of occurrences of some specific event through time. Examples include the arrivals of customers at a counter, the occurrences of earthquakes in a certain region, the occurrences of breakdowns in an electricity generator, etc. The Poisson process is a natural modelling tool in numerous applied probability problems. It not only models many real-world phenomena, but the process allows for tractable mathematical analysis as well.

The Poisson process is discussed in detail in Section 1.1. Basic properties are derived including the characteristic memoryless property. Illustrative examples are given to show the usefulness of the model. The compound Poisson process is dealt with in Section 1.2. In a Poisson arrival process customers arrive singly, while in a compound Poisson arrival process customers arrive in batches. Another generalization of the Poisson process is the non-stationary Poisson process that is discussed in Section 1.3. The Poisson process assumes that the intensity at which events occur is time-independent. This assumption is dropped in the non-stationary Poisson process. The final Section 1.4 discusses the Markov modulated arrival process in which the intensity at which Poisson arrivals occur is subject to a random environment.

1.1 THE POISSON PROCESS

There are several equivalent definitions of the Poisson process. Our starting point is a sequence $X_1, X_2, \ldots$ of positive, independent random variables with a common probability distribution. Think of X_n as the time elapsed between the $(n-1)$th and nth occurrence of some specific event in a probabilistic situation. Let

$$S_0 = 0 \quad \text{and} \quad S_n = \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots .$$

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

Then S_n is the epoch at which the nth event occurs. For each $t \geq 0$, define the random variable $N(t)$ by

$$N(t) = \text{the largest integer } n \geq 0 \text{ for which } S_n \leq t.$$

The random variable $N(t)$ represents the number of events up to time t.

Definition 1.1.1 *The counting process* $\{N(t),\ t \geq 0\}$ *is called a Poisson process with rate* λ *if the interoccurrence times* $X_1, X_2, \ldots$ *have a common exponential distribution function*

$$P\{X_n \leq x\} = 1 - e^{-\lambda x}, \quad x \geq 0.$$

The assumption of exponentially distributed interoccurrence times seems to be restrictive, but it appears that the Poisson process is an excellent model for many real-world phenomena. The explanation lies in the following deep result that is only roughly stated; see Khintchine (1969) for the precise rationale for the Poisson assumption in a variety of circumstances (the Palm–Khintchine theorem). Suppose that at microlevel there are a *very large* number of independent stochastic processes, where each *separate* microprocess generates only *rarely* an event. Then at macrolevel the superposition of all these microprocesses behaves approximately as a Poisson process. This insightful result is analogous to the well-known result that the number of successes in a very large number of independent Bernoulli trials with a very small success probability is approximately Poisson distributed. The superposition result provides an explanation of the occurrence of Poisson processes in a wide variety of circumstances. For example, the number of calls received at a large telephone exchange is the superposition of the individual calls of many subscribers each calling infrequently. Thus the process describing the overall number of calls can be expected to be close to a Poisson process. Similarly, a Poisson demand process for a given product can be expected if the demands are the superposition of the individual requests of many customers each asking infrequently for that product. Below it will be seen that the reason of the mathematical tractability of the Poisson process is its memoryless property. Information about the time elapsed since the last event is not relevant in predicting the time until the next event.

1.1.1 The Memoryless Property

In the remainder of this section we use for the Poisson process the terminology of 'arrivals' instead of 'events'. We first characterize the distribution of the counting variable $N(t)$. To do so, we use the well-known fact that the sum of k independent random variables with a common exponential distribution has an Erlang distribution. That is,

$$P\{S_k \le t\} = 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \quad t \ge 0. \tag{1.1.1}$$

The Erlang (k, λ) distribution has the probability density $\lambda^k t^{k-1} e^{-\lambda t}/(k-1)!$.

Theorem 1.1.1 *For any $t > 0$,*

$$P\{N(t) = k\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, \ldots . \tag{1.1.2}$$

That is, $N(t)$ is Poisson distributed with mean λt.

Proof The proof is based on the simple but useful observation that the number of arrivals up to time t is k or more if and only if the kth arrival occurs before or at time t. Hence

$$\begin{aligned} P\{N(t) \ge k\} &= P\{S_k \le t\} \\ &= 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}. \end{aligned}$$

The result next follows from $P\{N(t) = k\} = P\{N(t) \ge k\} - P\{N(t) \ge k+1\}$.

The following remark is made. To memorize the expression (1.1.1) for the distribution function of the Erlang (k, λ) distribution it is easiest to reason in reverse order: since the number of arrivals in $(0, t)$ is Poisson distributed with mean λt and the kth arrival time S_k is at or before t only if k or more arrivals occur in $(0, t)$, it follows that $P\{S_k \le t\} = \sum_{j=k}^{\infty} e^{-\lambda t}(\lambda t)^j/j!$.

The memoryless property of the Poisson process

Next we discuss the memoryless property that is characteristic for the Poisson process. For any $t \ge 0$, define the random variable γ_t as

$$\gamma_t = \text{the waiting time from epoch } t \text{ until the next arrival.}$$

The following theorem is of utmost importance.

Theorem 1.1.2 *For any $t \ge 0$, the random variable γ_t has the same exponential distribution with mean $1/\lambda$. That is,*

$$P\{\gamma_t \le x\} = 1 - e^{-\lambda x}, \quad x \ge 0, \tag{1.1.3}$$

independently of t.

Proof Fix $t \geq 0$. The event $\{\gamma_t > x\}$ occurs only if one of the mutually exclusive events $\{X_1 > t + x\}$, $\{X_1 \leq t,\ X_1 + X_2 > t + x\}$, $\{X_1 + X_2 \leq t,\ X_1 + X_2 + X_3 > t + x\}, \ldots$ occurs. This gives

$$P\{\gamma_t > x\} = P\{X_1 > t + x\} + \sum_{n=1}^{\infty} P\{S_n \leq t,\ S_{n+1} > t + x\}.$$

By conditioning on S_n, we find

$$\begin{aligned} P\{S_n \leq t,\ S_{n+1} > t + x\} &= \int_0^t P\{S_{n+1} > t + x \mid S_n = y\} \lambda^n \frac{y^{n-1}}{(n-1)!} e^{-\lambda y}\, dy \\ &= \int_0^t P\{X_{n+1} > t + x - y\} \lambda^n \frac{y^{n-1}}{(n-1)!} e^{-\lambda y}\, dy. \end{aligned}$$

This gives

$$\begin{aligned} P\{\gamma_t > x\} &= e^{-\lambda(t+x)} + \sum_{n=1}^{\infty} \int_0^t e^{-\lambda(t+x-y)} \lambda^n \frac{y^{n-1}}{(n-1)!} e^{-\lambda y}\, dy \\ &= e^{-\lambda(t+x)} + \int_0^t e^{-\lambda(t+x-y)} \lambda\, dy \\ &= e^{-\lambda(t+x)} + e^{-\lambda(t+x)}(e^{\lambda t} - 1) = e^{-\lambda x}, \end{aligned}$$

proving the desired result. The interchange of the sum and the integral in the second equality is justified by the non-negativity of the terms involved.

The theorem states that at each point in time the waiting time until the next arrival has the *same* exponential distribution as the *original* interarrival time, regardless of how long ago the last arrival occurred. The Poisson process is the only renewal process having this memoryless property. How much time is elapsed since the last arrival gives no information about how long to wait until the next arrival. This remarkable property does not hold for general arrival processes (e.g. consider the case of constant interarrival times). The *lack of memory* of the Poisson process explains the mathematical tractability of the process. In specific applications the analysis does not require a state variable keeping track of the time elapsed since the last arrival. The memoryless property of the Poisson process is of course closely related to the lack of memory of the exponential distribution.

Theorem 1.1.1 states that the number of arrivals in the time interval $(0, s)$ is Poisson distributed with mean λs. More generally, the number of arrivals in any time interval of length s has a Poisson distribution with mean λs. That is,

$$P\{N(u + s) - N(u) = k\} = e^{-\lambda s} \frac{(\lambda s)^k}{k!}, \quad k = 0, 1, \ldots, \tag{1.1.4}$$

independently of u. To prove this result, note that by Theorem 1.1.2 the time elapsed between a given epoch u and the epoch of the first arrival after u has the

same exponential distribution as the time elapsed between epoch 0 and the epoch of the first arrival after epoch 0. Next mimic the proof of Theorem 1.1.1.

To illustrate the foregoing, we give the following example.

Example 1.1.1 A taxi problem

Group taxis are waiting for passengers at the central railway station. Passengers for those taxis arrive according to a Poisson process with an average of 20 passengers per hour. A taxi departs as soon as four passengers have been collected or ten minutes have expired since the first passenger got in the taxi.

(a) Suppose you get in the taxi as first passenger. What is the probability that you have to wait ten minutes until the departure of the taxi?

(b) Suppose you got in the taxi as first passenger and you have already been waiting for five minutes. In the meantime two other passengers got in the taxi. What is the probability that you will have to wait another five minutes until the taxi departs?

To answer these questions, we take the minute as time unit so that the arrival rate $\lambda = 1/3$. By Theorem 1.1.1 the answer to question (a) is given by

$$P\{\text{less than 3 passengers arrive in } (0, 10)\} = \sum_{k=0}^{2} e^{-10/3} \frac{(10/3)^k}{k!} = 0.3528.$$

The answer to question (b) follows from the memoryless property stated in Theorem 1.1.2 and is given by

$$P\{\gamma_5 > 5\} = e^{-5/3} = 0.1889.$$

In view of the lack of memory of the Poisson process, it will be intuitively clear that the Poisson process has the following properties:

(A) *Independent increments: the numbers of arrivals occurring in disjoint intervals of time are independent.*

(B) *Stationary increments: the number of arrivals occurring in a given time interval depends only on the length of the interval.*

A formal proof of these properties will not be given here; see Exercise 1.8. To give the infinitesimal-transition rate representation of the Poisson process, we use

$$1 - e^{-h} = h - \frac{h^2}{2!} + \frac{h^3}{3!} - \cdots = h + o(h) \quad \text{as } h \to 0.$$

The mathematical symbol $o(h)$ is the generic notation for any function $f(h)$ with the property that $\lim_{h\to 0} f(h)/h = 0$, that is, $o(h)$ is some unspecified term that is negligibly small compared to h itself as $h \to 0$. For example, $f(h) = h^2$ is an $o(h)$-function. Using the expansion of e^{-h}, it readily follows from (1.1.4) that

(C) *The probability of one arrival occurring in a time interval of length* Δt is $\lambda \Delta t + o(\Delta t)$ *for* $\Delta t \to 0$.

(D) *The probability of two or more arrivals occurring in a time interval of length* Δt *is* $o(\Delta t)$ *for* $\Delta t \to 0$.

The property (D) states that the probability of two or more arrivals in a very small time interval of length Δt is negligibly small compared to Δt itself as $\Delta t \to 0$.

The Poisson process could alternatively be defined by taking (A), (B), (C) and (D) as postulates. This alternative definition proves to be useful in the analysis of continuous-time Markov chains in Chapter 4. Also, the alternative definition of the Poisson process has the advantage that it can be generalized to an arrival process with time-dependent arrival rate.

1.1.2 Merging and Splitting of Poisson Processes

Many applications involve the merging of independent Poisson processes or the splitting of events of a Poisson process in different categories. The next theorem shows that these situations again lead to Poisson processes.

Theorem 1.1.3 *(a) Suppose that* $\{N_1(t),\ t \geq 0\}$ *and* $\{N_2(t),\ t \geq 0\}$ *are independent Poisson processes with respective rates* λ_1 *and* λ_2*, where the process* $\{N_i(t)\}$ *corresponds to type* i *arrivals. Let* $N(t) = N_1(t) + N_2(t)$, $t \geq 0$*. Then the merged process* $\{N(t),\ t \geq 0\}$ *is a Poisson process with rate* $\lambda = \lambda_1 + \lambda_2$*. Denoting by* Z_k *the interarrival time between the* $(k-1)$*th and* k*th arrival in the merged process and letting* $I_k = i$ *if the* k*th arrival in the merged process is a type* i *arrival, then for any* $k = 1, 2, \ldots,$

$$P\{I_k = i \mid Z_k = t\} = \frac{\lambda_i}{\lambda_1 + \lambda_2}, \quad i = 1, 2, \tag{1.1.5}$$

independently of t.

(b) Let $\{N(t),\ t \geq 0\}$ *be a Poisson process with rate* λ*. Suppose that each arrival of the process is classified as being a type* 1 *arrival or type* 2 *arrival with respective probabilities* p_1 *and* p_2*, independently of all other arrivals. Let* $N_i(t)$ *be the number of type* i *arrivals up to time* t*. Then* $\{N_1(t)\}$ *and* $\{N_2(t)\}$ *are two independent Poisson processes having respective rates* λp_1 *and* λp_2.

Proof We give only a sketch of the proof using the properties (A), (B), (C) and (D).

(a) It will be obvious that the process $\{N(t)\}$ satisfies the properties (A) and (B). To verify property (C) note that

$$\begin{aligned} &P\{\text{one arrival in } (t, t+\Delta t]\} \\ &\quad = \sum_{i=1}^{2} P\left\{\begin{array}{c}\text{one arrival of type } i \text{ and no arrival} \\ \text{of the other type in } (t, t+\Delta t]\end{array}\right\} \\ &\quad = [\lambda_1 \Delta t + o(\Delta t)][1 - \lambda_2 \Delta t + o(\Delta t)] \\ &\qquad + [\lambda_2 \Delta t + o(\Delta t)][1 - \lambda_1 \Delta t + o(\Delta t)] \\ &\quad = (\lambda_1 + \lambda_2)\Delta t + o(\Delta t) \quad \text{as } \Delta t \to 0. \end{aligned}$$

Property (D) follows by noting that

$$\begin{aligned} P\{\text{no arrival in } (t, t+\Delta t]\} &= [1 - \lambda_1 \Delta t + o(\Delta t)][1 - \lambda_2 \Delta t + o(\Delta t)] \\ &= 1 - (\lambda_1 + \lambda_2)\Delta t + o(\Delta t) \quad \text{as } \Delta t \to 0. \end{aligned}$$

This completes the proof that $\{N(t)\}$ is a Poisson process with rate $\lambda_1 + \lambda_2$. To prove the other assertion in part (a), denote by the random variable Y_i the interarrival time in the process $\{N_i(t)\}$. Then

$$\begin{aligned} P\{Z_k > t, I_k = 1\} &= P\{Y_2 > Y_1 > t\} \\ &= \int_t^\infty P\{Y_2 > Y_1 > t \mid Y_1 = x\} \lambda_1 e^{-\lambda_1 x}\, dx \\ &= \int_t^\infty e^{-\lambda_2 x} \lambda_1 e^{-\lambda_1 x}\, dx = \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}. \end{aligned}$$

By taking $t = 0$, we find $P\{I_k = 1\} = \lambda_1/(\lambda_1 + \lambda_2)$. Since $\{N(t)\}$ is a Poisson process with rate $\lambda_1 + \lambda_2$, we have $P\{Z_k > t\} = \exp[-(\lambda_1 + \lambda_2)t]$. Hence

$$P\{I_k = 1, Z_k > t\} = P\{I_k = 1\} P\{Z_k > t\},$$

showing that $P\{I_k = 1 \mid Z_k = t\} = \lambda_1/(\lambda_1 + \lambda_2)$ independently of t.

(b) Obviously, the process $\{N_i(t)\}$ satisfies the properties (A), (B) and (D). To verify property (C), note that

$$\begin{aligned} P\{\text{one arrival of type } i \text{ in } (t, t+\Delta t]\} &= (\lambda \Delta t) p_i + o(\Delta t) \\ &= (\lambda p_i)\Delta t + o(\Delta t). \end{aligned}$$

It remains to prove that the processes $\{N_1(t)\}$ and $\{N_2(t)\}$ are independent. Fix $t > 0$. Then, by conditioning,

$$
\begin{aligned}
&P\{N_1(t) = k,\ N_2(t) = m\} \\
&\quad = \sum_{n=0}^{\infty} P\{N_1(t) = k,\ N_2(t) = m \mid N(t) = n\} P\{N(t) = n\} \\
&\quad = P\{N_1(t) = k,\ N_2(t) = m \mid N(t) = k + m\} P\{N(t) = k + m\} \\
&\quad = \binom{k+m}{k} p_1^k p_2^m e^{-\lambda t} \frac{(\lambda t)^{k+m}}{(k+m)!} \\
&\quad = e^{-\lambda p_1 t} \frac{(\lambda p_1 t)^k}{k!} e^{-\lambda p_2 t} \frac{(\lambda p_2 t)^m}{m!},
\end{aligned}
$$

showing that $P\{N_1(t) = k, N_2(t) = m\} = P\{N_1(t) = k\} P\{N_2(t) = m\}$.

The remarkable result (1.1.5) states that the next arrival is of type i with probability $\lambda_i/(\lambda_1 + \lambda_2)$ regardless of how long it takes until the next arrival. This result is characteristic for competing Poisson processes which are independent of each other. As an illustration, suppose that long-term parkers and short-term parkers arrive at a parking lot according to independent Poisson processes with respective rates λ_1 and λ_2. Then the merged arrival process of parkers is a Poisson process with rate $\lambda_1 + \lambda_2$ and the probability that a newly arriving parker is a long-term parker equals $\lambda_1/(\lambda_1 + \lambda_2)$.

Example 1.1.2 A stock problem with substitutable products

A store has a leftover stock of Q_1 units of product 1 and Q_2 units of product 2. Both products are taken out of production. Customers asking for product 1 arrive according to a Poisson process with rate λ_1. Independently of this process, customers asking for product 2 arrive according to a Poisson process with rate λ_2. Each customer asks for one unit of the concerning product. The two products serve as substitute for each other, that is, a customer asking for a product that is sold out is satisfied with the other product when still in stock. What is the probability distribution of the time until both products are sold out? What is the probability that product 1 is sold out before product 2?

To answer the first question, observe that both products are sold out as soon as $Q_1 + Q_2$ demands have occurred. The aggregated demand process is a Poisson process with rate $\lambda_1 + \lambda_2$. Hence the time until both products are sold out has an Erlang $(Q_1 + Q_2, \lambda_1 + \lambda_2)$ distribution. To answer the second question, observe that product 1 is sold out before product 2 only if the first $Q_1 + Q_2 - 1$ aggregated demands have no more than $Q_2 - 1$ demands for product 2. Hence, by (1.1.5), the desired probability is given by

$$
\sum_{k=0}^{Q_2-1} \binom{Q_1 + Q_2 - 1}{k} \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{Q_1+Q_2-1-k}.
$$

1.1.3 The $M/G/\infty$ Queue*

Suppose that customers arrive at a service facility according to a Poisson process with rate λ. The service facility has an ample number of servers. In other words, it is assumed that each customer gets immediately assigned a new server upon arrival. The service times of the customers are independent random variables having a common probability distribution with finite mean μ. The service times are independent of the arrival process. This versatile model is very useful in applications. An interesting question is: what is the limiting distribution of the number of busy servers? The surprisingly simple answer to this question is that the limiting distribution is a Poisson distribution with mean $\lambda\mu$:

$$\lim_{t\to\infty} P(k \text{ servers are busy at time } t) = e^{-\lambda\mu}\frac{(\lambda\mu)^k}{k!} \tag{1.1.6}$$

for $k = 0, 1, \ldots$. This limiting distribution does not require the shape of the service-time distribution, but uses the service-time distribution only through its mean μ. This famous *insensitivity* result is extremely useful for applications. The $M/G/\infty$ model has applications in various fields. A nice application is the $(S-1, S)$ inventory system with back ordering. In this model customers asking for a certain product arrive according to a Poisson process with rate λ. Each customer asks for one unit of the product. The initial on-hand inventory is S. Each time a customer demand occurs, a replenishment order is placed for exactly one unit of the product. A customer demand that occurs when the on-hand inventory is zero also triggers a replenishment order and the demand is back ordered until a unit becomes available to satisfy the demand. The lead times of the replenishment orders are independent random variables each having the same probability distribution with mean τ. Some reflections show that this $(S-1, S)$ inventory system can be translated into the $M/G/\infty$ queueing model: identify the outstanding replenishment orders with customers in service and identify the lead times of the replenishment orders with the service times. Thus the limiting distribution of the number of outstanding replenishment orders is a Poisson distribution with mean $\lambda\tau$. In particular,

$$\text{the long-run average on-hand inventory} = \sum_{k=0}^{S}(S-k)\,e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}.$$

Returning to the $M/G/\infty$ model, we first give a heuristic argument for (1.1.6) and next a rigorous proof.

Heuristic derivation

Suppose first that the service times are deterministic and are equal to the constant $D = \mu$. Fix t with $t > D$. If each service time is precisely equal to the constant

*This section can be skipped at first reading.

D, then the only customers present at time t are those customers who have arrived in $(t - D, t]$. Hence the number of customers present at time t is Poisson distributed with mean λD proving (1.1.6) for the special case of deterministic service times. Next consider the case that the service time takes on finitely many values $D_1, \ldots, D_s$ with respective probabilities $p_1, \ldots, p_s$. Mark the customers with the same fixed service time D_k as type k customers. Then, by Theorem 1.1.3, type k customers arrive according to a Poisson process with rate λp_k. Moreover the various Poisson arrival processes of the marked customers are independent of each other. Fix now t with $t > \max_k D_k$. By the above argument, the number of type k customers present at time t is Poisson distributed with mean $(\lambda p_k)D_k$. Thus, by the independence property of the split Poisson process, the total number of customers present at time t has a Poisson distribution with mean

$$\sum_{k=1}^{s} \lambda p_k D_k = \lambda \mu.$$

This proves (1.1.6) for the case that the service time has a discrete distribution with finite support. Any service-time distribution can be arbitrarily closely approximated by a discrete distribution with finite support. This makes plausible that the insensitivity result (1.1.6) holds for any service-time distribution.

Rigorous derivation

The differential equation approach can be used to give a rigorous proof of (1.1.6). Assuming that there are no customers present at epoch 0, define for any $t > 0$

$$p_j(t) = P\{\text{there are } j \text{ busy servers at time } t\}, \quad j = 0, 1, \ldots .$$

Consider now $p_j(t + \Delta t)$ for Δt small. The event that there are j servers busy at time $t + \Delta t$ can occur in the following mutually exclusive ways:

(a) no arrival occurs in $(0,\Delta t)$ and there are j busy servers at time $t + \Delta t$ due to arrivals in $(\Delta t, t + \Delta t)$,

(b) one arrival occurs in $(0, \Delta t)$, the service of the first arrival is completed before time $t + \Delta t$ and there are j busy servers at time $t + \Delta t$ due to arrivals in $(\Delta t, t + \Delta t)$,

(c) one arrival occurs in $(0, \Delta t)$, the service of the first arrival is not completed before time $t + \Delta t$ and there are $j - 1$ other busy servers at time $t + \Delta t$ due to arrivals in $(\Delta t, t + \Delta t)$,

(d) two or more arrivals occur in $(0, \Delta t)$ and j servers are busy at time $t + \Delta t$.

Let $B(t)$ denote the probability distribution of the service time of a customer. Then, since a probability distribution function has at most a countable number of

discontinuity points, we find for almost all $t > 0$ that

$$p_j(t+\Delta t) = (1-\lambda\Delta t)p_j(t) + \lambda\Delta t B(t+\Delta t)p_j(t) + \lambda\Delta t\{1-B(t+\Delta t)\}p_{j-1}(t) + o(\Delta t).$$

Subtracting $p_j(t)$ from $p_j(t+\Delta t)$, dividing by Δt and letting $\Delta t \to 0$, we find

$$p_0'(t) = -\lambda(1-B(t))p_0(t)$$
$$p_j'(t) = -\lambda(1-B(t))p_j(t) + \lambda(1-B(t))p_{j-1}(t), \quad j = 1, 2, \ldots .$$

Next, by induction on j, it is readily verified that

$$p_j(t) = e^{-\lambda\int_0^t (1-B(x))\,dx}\frac{\left[\lambda\int_0^t (1-B(x))\,dx\right]^j}{j!}, \quad j = 0, 1, \ldots .$$

By a continuity argument this relation holds for all $t \geq 0$. Since $\int_0^\infty [1-B(x)]\,dx = \mu$, the result (1.1.6) follows. Another proof of (1.1.6) is indicated in Exercise 1.14.

Example 1.1.3 A stochastic allocation problem

A nationwide courier service has purchased a large number of transport vehicles for a new service the company is providing. The management has to allocate these vehicles to a number of regional centres. In total C vehicles have been purchased and these vehicles must be allocated to F regional centres. The regional centres operate independently of each other and each regional centre services its own group of customers. In region i customer orders arrive at the base station according to a Poisson process with rate λ_i for $i = 1, \ldots, F$. Each customer order requires a separate transport vehicle. A customer order that finds all vehicles occupied upon arrival is delayed until a vehicle becomes available. The processing time of a customer order in region i has a lognormal distribution with mean $E(S_i)$ and standard deviation $\sigma(S_i)$. The processing time includes the time the vehicle needs to return to its base station. The management of the company wishes to allocate the vehicles to the regions in such a way that all regions provide, as nearly as possible, a uniform level of service to the customers. The service level in a region is measured as the long-run fraction of time that all vehicles are occupied (it will be seen in Section 2.4 that the long-run fraction of delayed customer orders is also given by this service measure).

Let us assume that the parameters are such that each region gets a *large* number of vehicles and most of the time is able to directly provide a vehicle for an arriving customer order. Then the $M/G/\infty$ model can be used as an approximate model to obtain a satisfactory solution. Let the dimensionless quantity R_i denote

$$R_i = \lambda_i E(S_i), \quad i = 1, \ldots, F,$$

that is, R_i is the average amount of work that is offered per time unit in region i. Denoting by c_i the number of vehicles to be assigned to region i, we take c_i of the form

$$c_i \approx R_i + k\sqrt{R_i}, \quad i = 1, \ldots, F,$$

for an appropriate constant k. By using this *square-root rule*, each region will provide nearly the same service level to its customers. To explain this, we use for each region the $M/G/\infty$ model to approximate the probability that all vehicles in the region are occupied at an arbitrary point of time. It follows from (1.1.6) that for region i this probability is approximated by

$$\sum_{k=c_i}^{\infty} e^{-R_i} \frac{R_i^k}{k!}$$

when c_i vehicles are assigned to region i. The Poisson distribution with mean R can be approximated by a normal distribution with mean R and standard deviation $\sqrt{R}$ when R is large enough. Thus we use the approximation

$$\sum_{k=c_i}^{\infty} e^{-R_i} \frac{R_i^k}{k!} \approx 1 - \Phi\left(\frac{c_i - R_i}{\sqrt{R_i}}\right), \quad i = 1, \ldots, F,$$

where $\Phi(x)$ is the standard normal distribution function. By requiring that

$$\Phi\left(\frac{c_1 - R_1}{\sqrt{R_1}}\right) \approx \cdots \approx \Phi\left(\frac{c_F - R_F}{\sqrt{R_F}}\right),$$

we find the square-root formula for c_i. The constant k in this formula must be chosen such that

$$\sum_{i=1}^{F} c_i = C.$$

Together this requirement and the square-root formula give

$$k \approx \frac{C - \sum_{i=1}^{F} R_i}{\sum_{i=1}^{F} \sqrt{R_i}}.$$

This value of k is the guideline for determining the allocation $(c_1, \ldots, c_F)$ so that each region, as nearly as possible, provides a uniform service level. To illustrate this, consider the numerical data:

$$c = 250,\ F = 5,\ \lambda_1 = 5,\ \lambda_2 = 10,\ \lambda_3 = 10,\ \lambda_4 = 50,\ \lambda_5 = 37.5,$$

$$E(S_1) = 2,\ E(S_2) = 2.5,\ E(S_3) = 3.5,\ E(S_4) = 1,\ E(S_5) = 2,$$

$$\sigma(S_1) = 1.5,\ \sigma(S_2) = 2,\ \sigma(S_3) = 3,\ \sigma(S_4) = 1,\ \sigma(S_5) = 2.7.$$

Then the estimate for k is 1.8450. Substituting this value into the square-root formula for c_i, we find $c_1 \approx 15.83$, $c_2 \approx 34.23$, $c_3 \approx 45.92$, $c_4 \approx 63.05$ and $c_5 \approx 90.98$. This suggests the allocation

$$(c_1^*,\ c_2^*,\ c_3^*,\ c_4^*,\ c_5^*) = (16,\ 34,\ 46,\ 63,\ 91).$$

Note that in determining this allocation we have used the distributions of the processing times only through their first moments. The actual value of the long-run fraction of time during which all vehicles are occupied in region i depends (to a slight degree) on the probability distribution of the processing time S_i. Using simulation, we find the values 0.056, 0.058, 0.050, 0.051 and 0.050 for the service level in the respective regions 1, 2, 3, 4 and 5.

The $M/G/\infty$ queue also has applications in the analysis of inventory systems.

Example 1.1.4 A two-echelon inventory system with repairable items

Consider a two-echelon inventory system consisting of a central depot and a number N of regional bases that operate independently of each other. Failed items arrive at the base level and are either repaired at the base or at the central depot, depending on the complexity of the repair. More specifically, failed items arrive at the bases $1, \ldots, N$ according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_N$. A failed item at base j can be repaired at the base with probability r_j; otherwise the item must be repaired at the depot. The average repair time of an item is μ_j at base j and μ_0 at the depot. It takes an average time of τ_j to ship an item from base j to the depot and back. The base immediately replaces a failed item from base stock if available; otherwise the replacement of the failed item is back ordered until an item becomes available at the base. If a failed item from base j arrives at the depot for repair, the depot immediately sends a replacement item to the base j from depot stock if available; otherwise the replacement is back ordered until a repaired item becomes available at the depot. In the two-echelon system a total of J spare parts are available. The goal is to spread these parts over the bases and the depot in order to minimize the total average number of back orders outstanding at the bases. This repairable-item inventory model has applications in the military, among others.

An approximate analysis of this inventory system can be given by using the $M/G/\infty$ queueing model. Let $(S_0, S_1, \ldots, S_N)$ be a given design for which S_0 spare parts have been assigned to the depot and S_j spare parts to base j for $j = 1, \ldots, N$ such that $S_0 + S_1 + \cdots + S_N = J$. At the depot, failed items arrive according to a Poisson process with rate

$$\lambda_0 = \sum_{j=1}^{N} \lambda_j (1 - r_j).$$

Each failed item arriving at the depot immediately goes to repair. The failed items arriving at the depot can be thought of as customers arriving at a queueing system

with infinitely many servers. Hence the limiting distribution of the number of items in repair at the depot at an arbitrary point of time is a Poisson distribution with mean $\lambda_0\mu_0$. The available stock at the depot is positive only if less than S_0 items are in repair at the depot. Why? Hence a delay occurs for the replacement of a failed item arriving at the depot only if S_0 or more items are in repair upon arrival of the item. Define now

$W_0 =$ the long-run average amount of time a failed item at the depot waits before a replacement is shipped,

$L_0 =$ the long-run average number of failed items at the depot waiting for the shipment of a replacement.

A simple relation exists between L_0 and W_0. On average λ_0 failed items arrive at the depot per time unit and on average a failed item at the depot waits W_0 time units before a replacement is shipped. Thus the average number of failed items at the depot waiting for the shipment of a replacement equals $\lambda_0 W_0$. This heuristic argument shows that

$$L_0 = \lambda_0 W_0.$$

This relation is a special case of Little's formula to be discussed in Section 2.3. The relation $W_0 = L_0/\lambda_0$ leads to an explicit formula for W_0, since L_0 is given by

$$L_0 = \sum_{k=S_0}^{\infty} (k - S_0) e^{-\lambda_0\mu_0} \frac{(\lambda_0\mu_0)^k}{k!}.$$

Armed with an explicit expression for W_0, we are able to give a formula for the long-run average number of back orders outstanding at the bases. For each base j the failed items arriving at base j can be thought of as customers entering service in a queueing system with infinitely many servers. Here the service time should be defined as the repair time in case of repair at the base and otherwise as the time until receipt of a replacement from the depot. Thus the average service time of a customer at base j is given by

$$\beta_j = r_j\mu_j + (1 - r_j)(\tau_j + W_0), \quad j = 1, \ldots, N.$$

The situation at base j can only be modelled approximately as an $M/G/\infty$ queue. The reason is that the arrival process of failed items interferes with the replacement times at the depot so that there is some dependency between the service times at base j. Assuming that this dependency is not substantial, we nevertheless use the $M/G/\infty$ queue as an approximating model and approximate the limiting distribution of the number of items in service at base j by a Poisson distribution with

mean $\lambda_j \beta_j$ for $j = 1, \ldots, N$. In particular,

$$\text{the long-run average number of back orders outstanding at base } j \approx \sum_{k=S_j}^{\infty} (k - S_j) e^{-\lambda_j \beta_j} \frac{(\lambda_j \beta_j)^k}{k!}, \qquad j = 1, \ldots, N.$$

This expression and the expression for W_0 enables us to calculate the total average number of outstanding back orders at the bases for a given assignment $(S_0, S_1, \ldots, S_N)$. Next, by some search procedure, the optimal values of $S_0, S_1, \ldots, S_N$ can be calculated.

1.1.4 The Poisson Process and the Uniform Distribution

In any small time interval of the same length the occurrence of a Poisson arrival is equally likely. In other words, Poisson arrivals occur completely randomly in time. To make this statement more precise, we relate the Poisson process to the uniform distribution.

Lemma 1.1.4 *For any $t > 0$ and $n = 1, 2, \ldots,$*

$$P\{S_k \le x \mid N(t) = n\} = \sum_{j=k}^{n} \binom{n}{j} \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{n-j} \tag{1.1.7}$$

for $0 \le x \le t$ and $1 \le k \le n$. In particular, for any $1 \le k \le n$,

$$E(S_k \mid N(t) = n) = \frac{kt}{n+1} \quad \text{and} \quad E(S_k - S_{k-1} \mid N(t) = n) = \frac{t}{n+1}. \tag{1.1.8}$$

Proof Since the Poisson process has independent and stationary increments,

$$\begin{aligned} P\{S_k \le x \mid N(t) = n\} &= \frac{P\{S_k \le x, N(t) = n\}}{P\{N(t) = n\}} \\ &= \frac{P\{N(x) \ge k, N(t) = n\}}{P\{N(t) = n\}} \\ &= \frac{1}{P\{N(t) = n\}} \sum_{j=k}^{n} P\{N(x) = j, N(t) - N(x) = n - j\} \\ &= \frac{1}{e^{-\lambda t}(\lambda t)^n / n!} \sum_{j=k}^{n} e^{-\lambda x} \frac{(\lambda x)^j}{j!} e^{-\lambda(t-x)} \frac{[\lambda(t-x)]^{n-j}}{(n-j)!} \\ &= \sum_{j=k}^{n} \binom{n}{j} \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{n-j}, \end{aligned}$$

proving the first assertion. Since $E(U) = \int_0^\infty P\{U > u\}\,du$ for any non-negative random variable U, the second assertion follows from (1.1.7) and the identity

$$\frac{(p+q+1)!}{p!q!}\int_0^1 y^p(1-y)^q\,dy = 1, \quad p, q = 0, 1, \ldots .$$

The right-hand side of (1.1.7) can be given the following interpretation. Let $U_1, \ldots, U_n$ be n independent random variables that are uniformly distributed on the interval $(0, t)$. Then the right-hand side of (1.1.7) also represents the probability that the smallest kth among $U_1, \ldots, U_n$ is less than or equal to x. This is expressed more generally in Theorem 1.1.5.

Theorem 1.1.5 *For any $t > 0$ and $n = 1, 2, \ldots$,*

$$P\{S_1 \le x_1, \ldots, S_n \le x_n \mid N(t) = n\} = P\{U_{(1)} \le x_1, \ldots, U_{(n)} \le x_n\},$$

where $U_{(k)}$ denotes the smallest kth among n independent random variables $U_1, \ldots, U_n$ that are uniformly distributed over the interval $(0, t)$.

The proof of this theorem proceeds along the same lines as that of Lemma 1.1.4. In other words, given the occurrence of n arrivals in $(0, t)$, the n arrival epochs are statistically indistinguishable from n independent observations taken from the uniform distribution on $(0, t)$. Thus Poisson arrivals occur completely randomly in time.

Example 1.1.5 A waiting-time problem

In the harbour of Amsterdam a ferry leaves every T minutes to cross the North Sea canal, where T is fixed. Passengers arrive according to a Poisson process with rate λ. The ferry has ample capacity. What is the expected total waiting time of all passengers joining a given crossing? The answer is

$$E(\text{total waiting time}) = \frac{1}{2}\lambda T^2. \tag{1.1.9}$$

To prove this, consider the first crossing of the ferry. The random variable $N(T)$ denotes the number of passengers joining this crossing and the random variable S_k represents the arrival epoch of the kth passenger. By conditioning, we find

$$E(\text{total waiting time}) = \sum_{n=0}^{\infty} E(\text{total waiting time} \mid N(T) = n)P\{N(T) = n\}$$

$$= \sum_{n=1}^{\infty} E(T - S_1 + T - S_2 + \cdots + T - S_n \mid N(T) = n) e^{-\lambda T} \frac{(\lambda T)^n}{n!}$$

$$= \sum_{n=1}^{\infty} E(T - U_{(1)} + T - U_{(2)} + \cdots + T - U_{(n)}) e^{-\lambda T} \frac{(\lambda T)^n}{n!}.$$

This gives

$$E(\text{total waiting time up to time } T) = \sum_{n=1}^{\infty} E(nT - (U_1 + \cdots + U_n)) e^{-\lambda T} \frac{(\lambda T)^n}{n!}$$

$$= \sum_{n=1}^{\infty} \left(nT - n\frac{T}{2} \right) e^{-\lambda T} \frac{(\lambda T)^n}{n!} = \frac{T}{2} \lambda T,$$

which proves the desired result.

The result (1.1.9) is simple but very useful. It is sometimes used in a somewhat different form that can be described as follows. Messages arrive at a communication channel according to a Poisson process with rate λ. The messages are stored in a buffer with ample capacity. A holding cost at rate $h > 0$ per unit of time is incurred for each message in the buffer. Then, by (1.1.9),

$$E(\text{holding costs incurred up to time } T) = \frac{h}{2} \lambda T^2. \tag{1.1.10}$$

Clustering of Poisson arrival epochs

Theorem 1.1.5 expresses that Poisson arrival epochs occur completely randomly in time. This is in agreement with the lack of memory of the exponential density $\lambda e^{-\lambda x}$ of the interarrival times. This density is largest at $x = 0$ and decreases as x increases. Thus short interarrival times are relatively frequent. This suggests that the Poisson arrival epochs show a tendency to cluster. Indeed this is confirmed by simulation experiments. Clustering of points in Poisson processes is of interest in many applications, including risk analysis and telecommunication. It is therefore important to have a formula for the probability that a given time interval of length T contains *some* time window of length w in which n or more Poisson events occur. An exact expression for this probability is difficult to give, but a simple and excellent approximation is provided by

$$1 - P(n-1, \lambda w) \exp\left[- \left(1 - \frac{\lambda w}{n} \right) \lambda (T - w) p(n-1, \lambda w) \right],$$

where $p(k, \lambda w) = e^{-\lambda w} (\lambda w)^k / k!$ and $P(n, \lambda w) = \sum_{k=0}^{n} p(k, \lambda w)$. The approximation is called Alm's approximation; see Glaz and Balakrishnan (1999). To illustrate the clustering phenomenon, consider the following example. In the first five months of the year 2000, trams hit and killed seven people in Amsterdam, each

case caused by the pedestrian's carelessness. In the preceding years such accidents occurred on average 3.7 times per year. Is the clustering of accidents in the year 2000 exceptional? It is exceptional if seven or more fatal accidents occur during the *coming* five months, but it is not exceptional when over a period of ten years (say) seven or more accidents happen in *some* time window having a length of five months. The above approximation gives the value 0.104 for the probability that over a period of ten years there is some time window having a length of five months in which seven or more fatal accidents occur. The exact value of the probability is 0.106.

1.2 COMPOUND POISSON PROCESSES

A compound Poisson process generalizes the Poisson process by allowing jumps that are not necessarily of unit magnitude.

Definition 1.2.1 *A stochastic process* $\{X(t),\ t \geq 0\}$ *is said to be a compound Poisson process if it can be represented by*

$$X(t) = \sum_{i=1}^{N(t)} D_i, \quad t \geq 0,$$

where $\{N(t),\ t \geq 0\}$ *is a Poisson process with rate* λ*, and* $D_1, D_2, \ldots$ *are independent and identically distributed non-negative random variables that are also independent of the process* $\{N(t)\}$.

Compound Poisson processes arise in a variety of contexts. As an example, consider an insurance company at which claims arrive according to a Poisson process and the claim sizes are independent and identically distributed random variables, which are also independent of the arrival process. Then the cumulative amount claimed up to time t is a compound Poisson variable. Also, the compound Poisson process has applications in inventory theory. Suppose customers asking for a given product arrive according to a Poisson process. The demands of the customers are independent and identically distributed random variables, which are also independent of the arrival process. Then the cumulative demand up to time t is a compound Poisson variable.

The mean and variance of the compound Poisson variable $X(t)$ are given by

$$E[X(t)] = \lambda t E(D_1) \quad \text{and} \quad \sigma^2[X(t)] = \lambda t E(D_1^2), \quad t \geq 0. \tag{1.2.1}$$

This result follows from (A.9) and (A.10) in Appendix A and the fact that both the mean and variance of the Poisson variable $N(t)$ are equal to λt.

Discrete compound Poisson distribution

Consider first the case of discrete random variables $D_1, D_2, \ldots$:

$$a_j = P\{D_1 = j\}, \quad j = 0, 1, \ldots .$$

Then a simple algorithm can be given to compute the probability distribution of the compound Poisson variable $X(t)$. For any $t \geq 0$, let

$$r_j(t) = P\{X(t) = j\}, \quad j = 0, 1, \ldots .$$

Define the generating function $A(z)$ by

$$A(z) = \sum_{j=0}^{\infty} a_j z^j, \quad |z| \leq 1.$$

Also, for any fixed $t > 0$, define the generating function $R(z, t)$ as

$$R(z, t) = \sum_{j=0}^{\infty} r_j(t) z^j, \quad |z| \leq 1.$$

Theorem 1.2.1 *For any fixed $t > 0$ it holds that:*

(a) the generating function $R(z, t)$ is given by

$$R(z, t) = e^{-\lambda t\{1 - A(z)\}}, \quad |z| \leq 1 \tag{1.2.2}$$

(b) the probabilities $\{r_j(t),\ j = 0, 1, \ldots\}$ satisfy the recursion

$$r_j(t) = \frac{\lambda t}{j} \sum_{k=0}^{j-1} (j - k) a_{j-k} r_k(t), \quad j = 1, 2, \ldots , \tag{1.2.3}$$

starting with $r_0(t) = e^{-\lambda t(1 - a_0)}$.

Proof Fix $t \geq 0$. By conditioning on the number of arrivals up to time t,

$$\begin{aligned} r_j(t) &= \sum_{n=0}^{\infty} P\{X(t) = j \mid N(t) = n\} P\{N(t) = n\} \\ &= \sum_{n=0}^{\infty} P\{D_0 + \cdots + D_n = j\} e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad j = 0, 1, \ldots \end{aligned}$$

with $D_0 = 0$. This gives, after an interchange of the order of summation,

$$\sum_{j=0}^{\infty} r_j(t) z^j = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \sum_{j=0}^{\infty} P\{D_0 + \cdots + D_n = j\} z^j.$$

Since the D_i are independent of each other, it follows that

$$\sum_{j=0}^{\infty} P\{D_0 + \cdots + D_n = j\} z^j = E(z^{D_0 + \cdots + D_n})$$
$$= E(z^{D_0}) \cdots E(z^{D_n}) = [A(z)]^n.$$

Thus

$$R(z, t) = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} [A(z)]^n = e^{-\lambda t[1 - A(z)]}$$

which proves (1.2.2). To prove part (b) for fixed t, we write $R(z) = R(z, t)$ for ease of notation. It follows immediately from the definition of the generating function that the probability $r_j(t)$ is given by

$$r_j(t) = \frac{1}{j!} \left. \frac{d^j R(z)}{dz^j} \right|_{z=0}.$$

It is not possible to obtain (1.2.3) directly from this relation and (1.2.2). The following intermediate step is needed. By differentiation of (1.2.2), we find

$$R'(z) = \lambda t A'(z) R(z), \quad |z| \leq 1.$$

This gives

$$\sum_{j=1}^{\infty} j r_j(t) z^{j-1} = \lambda t \left[\sum_{k=1}^{\infty} k a_k z^{k-1} \right] \left[\sum_{\ell=0}^{\infty} r_\ell(t) z^\ell \right]$$
$$= \sum_{k=1}^{\infty} \sum_{\ell=0}^{\infty} \lambda t k a_k r_\ell(t) z^{k+\ell-1}.$$

Replacing $k + l$ by j and interchanging the order of summation yields

$$\sum_{j=1}^{\infty} j r_j(t) z^{j-1} = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} \lambda t k a_k r_{j-k}(t) z^{j-1}$$
$$= \sum_{j=1}^{\infty} \left[\sum_{k=1}^{j} \lambda t k a_k r_{j-k}(t) \right] z^{j-1}.$$

Next equating coefficients gives the recurrence relation (1.2.3).

The recursion scheme for the $r_j(t)$ is easy to program and is numerically stable. It is often called Adelson's recursion scheme after Adelson (1966). In the insurance literature the recursive scheme is known as Panjer's algorithm. Note that for the special case of $a_1 = 1$ the recursion (1.2.3) reduces to the familiar recursion

scheme for computing Poisson probabilities. An alternative method to compute the compound Poisson probabilities $r_j(t)$, $j = 0, 1, \ldots$ is to apply the discrete FFT method to the explicit expression (1.2.2) for the generating function of the $r_j(t)$; see Appendix D.

Continuous compound Poisson distribution

Suppose now that the non-negative random variables D_i are continuously distributed with probability distribution function $A(x) = P\{D_1 \leq x\}$ having the probability density $a(x)$. Then the compound Poisson variable $X(t)$ has the positive mass $e^{-\lambda t}$ at point zero and a density on the positive real line. Let

$$a^*(s) = \int_0^\infty e^{-sx} a(x)\, dx$$

be the Laplace transform of $a(x)$. In the same way that (1.2.2) was derived,

$$E[e^{-sX(t)}] = e^{-\lambda t\{1-a^*(s)\}}.$$

Fix $t > 0$. How do we compute $P\{X(t) > x\}$ as function of x? Several computational methods can be used. The probability distribution function $P\{X(t) > x\}$ for $x \geq 0$ can be computed by using a numerical method for Laplace inversion; see Appendix F. By relation (E.7) in Appendix E, the Laplace transform of $P\{X(t) > x\}$ is given by

$$\int_0^\infty e^{-sx} P\{X(t) > x\}\, dx = \frac{1 - e^{-\lambda t\{1-a^*(s)\}}}{s}.$$

If no explicit expression is available for $a^*(s)$ (as is the case when the D_i are lognormally distributed), an alternative is to use the integral equation

$$P\{X(t) > x\} = \int_0^t \left[1 - A(x) + \int_0^x P\{X(t-u) > x - y\} a(y)\, dy\right] \lambda e^{-\lambda u}\, du.$$

This integral equation is easily obtained by conditioning on the epoch of the first Poisson event and by conditioning on D_1. The corresponding integral equation for the density of $X(t)$ can be numerically solved by applying the discretization algorithm given in Den Iseger *et al.* (1997). This discretization method uses spline functions and is very useful when one is content with an approximation error of about 10^{-8}. Finally, for the special case of the D_i having a gamma distribution, the probability $P\{X(t) > x\}$ can simply be computed from

$$P\{X(t) > x\} = \sum_{n=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \{1 - B^{n*}(x)\}, \quad x > 0,$$

where the n-fold convolution function $B^{n*}(x)$ is the probability distribution function of $D_1 + \cdots + D_n$. If the D_i have a gamma distribution with shape parameter

α and scale parameter β, the sum $D_1 + \cdots + D_n$ has a gamma distribution with shape parameter $n\alpha$ and scale parameter β. The computation of the gamma distribution offers no numerical difficulties; see Appendix B. The assumption of a gamma distribution is appropriate in many inventory applications with $X(t)$ representing the cumulative demand up to time t.

1.3 NON-STATIONARY POISSON PROCESSES

The non-stationary Poisson process is another useful stochastic process for counting events that occur over time. It generalizes the Poisson process by allowing for an arrival rate that need not be constant in time. Non-stationary Poisson processes are used to model arrival processes where the arrival rate fluctuates significantly over time. In the discussion below, the arrival rate function $\lambda(t)$ is assumed to be piecewise continuous.

Definition 1.3.1 *A counting process $\{N(t),\ t \geq 0\}$ is said to be a non-stationary Poisson process with intensity function $\lambda(t),\ t \geq 0$, if it satisfies the following properties:*

(a) $N(0) = 0$

(b) the process $\{N(t)\}$ has independent increments

(c) $P\{N(t + \Delta t) - N(t) = 1\} = \lambda(t)\Delta t + o(\Delta t)$ *as* $\Delta t \to 0$

(d) $P\{N(t + \Delta t) - N(t) \geq 2\} = o(\Delta t)$ *as* $\Delta t \to 0$.

The next theorem proves that the total number of arrivals in a given time interval is Poisson distributed.

Theorem 1.3.1 *For any $t, s \geq 0$,*

$$P\{N(t + s) - N(t) = k\} = e^{-[M(t+s)-M(t)]} \frac{[M(t + s) - M(t)]^k}{k!}, \qquad (1.3.1)$$

for $k = 0, 1, \ldots$, where $M(x) = \int_0^x \lambda(y)\, dy$, $x \geq 0$.

Proof The proof is instructive. Fix $t \geq 0$. Put for abbreviation

$$p_k(s) = P\{N(t + s) - N(t) = k\}, \quad k = 0, 1, \ldots .$$

Consider now $p_k(s + \Delta s)$ for Δs small. Since the probability of two or more arrivals in a small time interval of length Δs is negligibly small compared with Δs as $\Delta s \to 0$, it follows that the only possibility for the process to be in state k at time $t + s + \Delta s$ is that the process is either in state $k - 1$ or in state k at time $t + s$. Hence, by conditioning on the state of the process at time $t + s$ and given that the process has independent increments,

$$p_k(s + \Delta s) = p_{k-1}(s)[\lambda(t + s)\Delta s + o(\Delta s)] + p_k(s)[1 - \lambda(t + s)\Delta s + o(\Delta s)]$$

as $\Delta s \to 0$. Subtracting $p_k(s)$ from both sides of this equation and dividing by Δs, we obtain

$$p_k'(s) = -\lambda(t+s)[p_k(s) - p_{k-1}(s)], \quad k = 1, 2, \ldots .$$

For $k = 0$, we have $p_0'(s) = -\lambda(t+s)p_0(s)$. The boundary conditions $p_0(0) = 1$ and $p_k(0) = 0$ for $k \geq 1$ apply. It is well known from the theory of differential equations that the solution of the first-order differential equation

$$y'(s) + a(s)y(s) = b(s), \quad s \geq 0$$

is given by

$$y(s) = e^{-A(s)} \int_0^s b(x)e^{A(x)}\,dx + ce^{-A(s)}$$

for some constant c, where $A(s) = \int_0^s a(x)\,dx$. The constant c is determined by a boundary condition on $y(0)$. This gives after some algebra

$$p_0(s) = e^{-[M(s+t)-M(t)]}, \quad s \geq 0.$$

By induction the expression for $p_k(s)$ next follows from $p_k'(s) + \lambda(t+s)p_k(s) = \lambda(t+s)p_{k-1}(s)$. We omit the details.

Note that $M(t)$ represents the expected number of arrivals up to time t.

Example 1.3.1 A canal touring problem

A canal touring boat departs for a tour through the canals of Amsterdam every T minutes with T fixed. Potential customers pass the point of departure according to a Poisson process with rate λ. A potential customer who sees that the boat leaves t minutes from now joins the boat with probability $e^{-\mu t}$ for $0 \leq t \leq T$. Which stochastic process describes the arrival of customers who actually join the boat (assume that the boat has ample capacity)? The answer is that this process is a non-stationary Poisson process with arrival rate function $\lambda(t)$, where

$$\lambda(t) = \lambda e^{-\mu(T-t)} \quad \text{for } 0 \leq t < T \quad \text{and} \quad \lambda(t) = \lambda(t-T) \quad \text{for } t \geq T.$$

This follows directly from the observation that for Δt small

$$P\{\text{a customer joins the boat in } (t, t+\Delta t)\}$$
$$= (\lambda \Delta t) \times e^{-\mu(T-t)} + o(\Delta t), \quad 0 \leq t < T.$$

Thus, by Theorem 1.3.1, the number of passengers joining a given tour is Poisson distributed with mean $\int_0^T \lambda(t)\,dt = (\lambda/\mu)(1 - e^{-\mu T})$.

Another illustration of the usefulness of the non-stationary Poisson process is provided by the following example.

Example 1.3.2 Replacement with minimal repair

A machine has a stochastic lifetime with a continuous distribution. The machine is replaced by a new one at fixed times $T, 2T, \ldots$, whereas a minimal repair is done at each failure occurring between two planned replacements. A minimal repair returns the machine into the condition it was in just before the failure. It is assumed that each minimal repair takes a negligible time. What is the probability distribution of the total number of minimal repairs between two planned replacements?

Let $F(x)$ and $f(x)$ denote the probability distribution function and the probability density of the lifetime of the machine. Also, let $r(t) = f(t)/[1 - F(t)]$ denote the failure rate function of the machine. It is assumed that $f(x)$ is continuous. Then the answer to the above question is

$$P\{\text{there are } k \text{ minimal repairs between two planned replacements}\}$$
$$= e^{-M(T)} \frac{[M(T)]^k}{k!}, \quad k = 0, 1, \ldots,$$

where $M(T) = \int_0^T r(t)\, dt$. This result follows directly from Theorem 1.3.1 by noting that the process counting the number of minimal repairs between two planned replacements satisfies the properties (a), (b), (c) and (d) of Definition 1.3.1. Use the fact that the probability of a failure of the machine in a small time interval $(t, t + \Delta t]$ is equal to $r(t)\Delta t + o(\Delta t)$, as shown in Appendix B.

1.4 MARKOV MODULATED BATCH POISSON PROCESSES*

The Markov modulated batch Poisson process generalizes the compound Poisson process by allowing for correlated interarrival times. This process is used extensively in the analysis of teletraffic models (a special case is the composite model of independent on-off sources multiplexed together). A so-called phase process underlies the arrival process, where the evolution of the phase process occurs isolated from the arrivals. The phase process can only assume a finite number of states $i = 1, \ldots, m$. The sojourn time of the phase process in state i is exponentially distributed with mean $1/\omega_i$. If the phase process leaves state i, it goes to state j with probability p_{ij}, independently of the duration of the stay in state i. It is assumed that $p_{ii} = 0$ for all i. The arrival process of customers is a compound Poisson process whose parameters depend on the state of the phase process. If the phase process is in state i, then batches of customers arrive according to a Poisson process with rate λ_i where the batch size has the

*This section contains specialized material that is not used in the sequel.

discrete probability distribution $\{a_k^{(i)}, k = 1, 2, \dots\}$. It is no restriction to assume that $a_0^{(i)} = 0$; otherwise replace λ_i by $\lambda_i(1 - a_0^{(i)})$ and $a_k^{(i)}$ by $a_k^{(i)}/(1 - a_0^{(i)})$ for $k \geq 1$.

For any $t \geq 0$ and $i, j = 1, \dots, m$, define

$$P_{ij}(k, t) = P\{\text{the total number of customers arriving in } (0, t) \text{ equals } k \text{ and the phase process is in state } j \text{ at time } t \mid \text{the phase process is in state } i \text{ at the present time } 0\}, \quad k = 0, 1, \dots .$$

Also, for any $t > 0$ and $i, j = 1, \dots, m$, let us define the generating function $P_{ij}^*(z, t)$ by

$$P_{ij}^*(z, t) = \sum_{k=0}^{\infty} P_{ij}(k, t) z^k, \quad |z| \leq 1.$$

To derive an expression for $P_{ij}^*(z, t)$, it is convenient to use matrix notation. Let $Q = (q_{ij})$ be the $m \times m$ matrix whose (i, j)th element is given by

$$q_{ii} = -\omega_i \quad \text{and} \quad q_{ij} = \omega_i p_{ij} \quad \text{for } j \neq i.$$

Define the $m \times m$ diagonal matrices Λ and A_k by

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m) \quad \text{and} \quad A_k = \operatorname{diag}(a_k^{(1)}, \dots, a_k^{(m)}), \quad k = 1, 2, \dots . \tag{1.4.1}$$

Let the $m \times m$ matrix D_k for $k = 0, 1, \dots$ be defined by

$$D_0 = Q - \Lambda \text{ and } D_k = A_k \Lambda, \quad k = 1, 2, \dots . \tag{1.4.2}$$

Using $(D_k)_{ij}$ to denote the (i, j)th element of the matrix D_k, define the generating function $D_{ij}(z)$ by

$$D_{ij}(z) = \sum_{k=0}^{\infty} (D_k)_{ij} z^k, \quad |z| \leq 1.$$

Theorem 1.4.1 *Let $P^*(z, t)$ and $D(z)$ denote the $m \times m$ matrices whose (i, j)th elements are given by the generating functions $P_{ij}^*(z, t)$ and $D_{ij}(z)$. Then, for any $t > 0$,*

$$P^*(z, t) = e^{D(z)t}, \quad |z| \leq 1, \tag{1.4.3}$$

where e^{At} is defined by $e^{At} = \Sigma_{n=0}^{\infty} A^n t^n / n!$.

Proof The proof is based on deriving a system of differential equations for the $P_{ij}(k, t)$. Fix i, j, k and t. Consider $P_{ij}(k, t + \Delta t)$ for Δt small. By conditioning

on what may happen in $(t, t + \Delta t)$, it follows that

$$P_{ij}(k, t + \Delta t) = P_{ij}(k, t)(1 - \lambda_j \Delta t)(1 - \omega_j \Delta t) + \sum_{s \neq j} P_{is}(k, t)[(\omega_s \Delta t) \times p_{sj}]$$
$$+ \sum_{\ell=0}^{k-1} P_{ij}(\ell, t)\left[(\lambda_j \Delta t) \times a_{k-\ell}^{(j)}\right] + o(\Delta t).$$

Using the definition of the q_{ij}, we rewrite this relation as

$$P_{ij}(k, t + \Delta t) = P_{ij}(k, t)(1 - \lambda_j \Delta t) + \sum_{s=1}^{m} P_{is}(k, t) q_{sj} \Delta t$$
$$+ \sum_{\ell=0}^{k-1} P_{ij}(\ell, t) \lambda_j a_{k-\ell}^{(j)} \Delta t + o(\Delta t),$$

which implies that

$$\frac{d}{dt} P_{ij}(k, t) = -\lambda_j P_{ij}(k, t) + \sum_{s=1}^{m} P_{is}(k, t) q_{sj} + \lambda_j \sum_{\ell=0}^{k-1} P_{ij}(\ell, t) a_{k-\ell}^{(j)}.$$

Letting $P(k, t)$ be the $m \times m$ matrix whose (i, j)th element is $P_{ij}(k, t)$, we have in matrix notation that

$$\frac{d}{dt} P(k, t) = P(k, t)(Q - \Lambda) + \sum_{\ell=0}^{k-1} P(\ell, t) A_{k-\ell} \Lambda.$$

Using the definition of the matrices D_k, we find next that

$$\frac{d}{dt} P(k, t) = P(k, t) D_0 + \sum_{\ell=0}^{k-1} P(\ell, t) D_{k-\ell}$$
$$= \sum_{\ell=0}^{k} P(\ell, t) D_{k-\ell}.$$

Multiply componentwise both sides of this matrix equation by z^k and sum over k. Since the generating function of the convolution of two sequences is the product of the generating functions of the two sequences, it follows that

$$\frac{d}{dt} P^*(z, t) = P^*(z, t) D(z).$$

For each fixed i this equation gives a system of linear differential equations in $P_{ij}^*(z, t)$ for $j = 1, \ldots, m$. Thus, by a standard result from the theory of linear

differential equations, we obtain

$$P_i^*(z,t) = e^{D(z)t} P_i^*(z,0) \tag{1.4.4}$$

where $P_i^*(z,t)$ is the ith row of the matrix $P^*(z,t)$. Since $P_i^*(z,0)$ equals the ith unit vector $e_i = (0,\ldots,1,\ldots,0)$, it next follows that $P^*(z,t) = e^{D(z)t}$, as was to be proved.

In general it is a formidable task to obtain the numerical values of the probabilities $P_{ij}(k,t)$ from the expression (1.4.4), particularly when m is large.* The numerical approach of the discrete FFT method is only practically feasible when the computation of the matrix $e^{D(z)t}$ is not too burdensome. Numerous algorithms for the computation of the matrix exponential e^{At} have been proposed, but they do not always provide high accuracy. The computational work is simplified when the $m \times m$ matrix A has m different eigenvalues $\mu_1,\ldots,\mu_m$ (say), as is often the case in applications. It is well known from linear algebra that the matrix A can then be diagonalized as

$$A = S\chi S^{-1},$$

where the diagonal matrix χ is given by $\chi = \text{diag}(\mu_1,\ldots,\mu_m)$ and the column vectors of the matrix S are the linearly independent eigenvectors associated with the eigenvalues $\mu_1,\ldots,\mu_m$. Moreover, by $A^n = S\chi^n S^{-1}$, it holds that

$$e^{At} = S\ \text{diag}(e^{\mu_1 t},\ldots,e^{\mu_m t})S^{-1}.$$

Fast codes for the computation of eigenvalues and eigenvectors of a (complex) matrix are widely available.

To conclude this section, it is remarked that the matrix $D(z)$ in the matrix exponential $e^{D(z)t}$ has a very simple form for the important case of single arrivals (i.e. $a_i^{(1)} = 1$ for $i = 1,\ldots,m$). It then follows from (1.4.1) and (1.4.2) that

$$D(z) = Q - \Lambda + \Lambda z, \quad |z| \le 1.$$

The arrival process with single arrivals is called the *Markov modulated Poisson process*. A special case of this process is the *switched Poisson process* which has only two arrival rates ($m = 2$). This model is frequently used in applications. In the special case of the switched Poisson process, the following explicit expressions can be given for the generating functions $P_{ij}^*(z,t)$:

$$P_{ii}^*(z,t) = \frac{1}{r_2(z) - r_1(z)}\Big[\{r_2(z) - (\lambda_i(1-z) + \omega_i)\}e^{-r_1(z)t} - \{r_1(z) - (\lambda_i(1-z) + \omega_i)\}e^{-r_2(z)t}\Big], \quad i = 1,2,$$

*It is also possible to formulate a direct probabilistic algorithm for the computation of the probabilities $P_{ij}(k,t)$. This algorithm is based on the uniformization method for continuous-time Markov chains; see Section 4.5.

$$P_{12}^*(z,t) = \omega_1 \frac{e^{-r_1(z)t} - e^{-r_2(z)t}}{r_2(z) - r_1(z)} \quad \text{and} \quad P_{21}^*(z,t) = \omega_2 \frac{e^{-r_1(z)t} - e^{-r_2(z)t}}{r_2(z) - r_1(z)},$$

where

$$\begin{aligned} r_{1,2}(z) = {} & \frac{1}{2}\left(\lambda_1(1-z) + \omega_1 + \lambda_2(1-z) + \omega_2\right) \\ & \pm \frac{1}{2}\Big[\{\lambda_1(1-z) + \omega_1 + \lambda_2(1-z) + \omega_2\}^2 \\ & \quad - 4\{(\lambda_1(1-z) + \omega_1)(\lambda_2(1-z) + \omega_2) - \omega_1\omega_2\}\Big]^{1/2}. \end{aligned}$$

It is a matter of straightforward but tedious algebra to derive these expressions. The probabilities $P_{ij}(k,t)$ can be readily computed from these expressions by applying the discrete FFT method.

EXERCISES

1.1 A businessman parks his car illegally in the streets of Amsterdam twice a day for a period of exactly one hour. Parking surveillances occur according to a Poisson process with an average of λ passes per hour. What is the probability of the businessman getting a fine on a given day?

1.2 At a shuttle station, passengers arrive according to a Poisson process with rate λ. A shuttle departs as soon as seven passengers have arrived. There is an ample number of shuttles at the station.

(a) What is the conditional distribution of the time a customer has to wait until departure when upon arrival the customer finds j other customers waiting for $j = 0, 1, \ldots, 6$?

(b) What is the probability that the nth customer will not have to wait? (*Hint*: distinguish between the case that n is a multiple of 7 and the case that n is not a multiple of 7.)

(c) What is the long-run fraction of customers who, upon arrival, find j other customers waiting for $j = 0, 1, \ldots 6$?

(d) What is the long-run fraction of customers who wait more than x time units until departure?

1.3 Answer (a), (b) and (c) in Exercise 1.2 assuming that the interarrival times of the customers have an Erlang $(2, \lambda)$ distribution.

1.4 You leave work at random times between 5 pm and 6 pm to take the bus home. Bus numbers 1 and 3 bring you home. You take the first bus that arrives. Bus number 1 arrives exactly every 10 minutes, whereas bus number 3 arrives according to a Poisson process with the same average frequency as bus number 1. What is the probability that you take bus number 1 home on a given day? Can you explain why this probability is larger than $1/2$?

1.5 You wish to cross a one-way traffic road on which cars drive at a constant speed and pass according to a Poisson process with rate λ. You can only cross the road when no car has come round the corner for c time units. What is the probability of the number of passing cars before you can cross the road when you arrive at a random moment? What property of the Poisson process do you use?

1.6 Consider a Poisson arrival process with rate λ. For each fixed $t > 0$, define the random variable δ_t as the time elapsed since the last arrival before or at time t (assume that an arrival occurs at epoch 0).

(a) Show that the random variable δ_t has a truncated exponential distribution: $P\{\delta_t = t\} = e^{-\lambda t}$ and $P\{\delta_t > x\} = e^{-\lambda x}$ for $0 \leq x < t$.

(b) Prove that the random variables γ_t (= waiting time from time t until the next arrival) and δ_t are independent of each other by verifying $P\{\gamma_t > u, \delta_t > v\} = P\{\gamma_t > u\}P\{\delta_t > v\}$ for all $u \geq 0$ and $0 \leq v < t$.

1.7 Suppose that fast and slow cars enter a one-way highway according to independent Poisson processes with respective rates λ_1 and λ_2. The length of the highway is L. A fast car travels at a constant speed of s_1 and a slow car at a constant speed of s_2 with $s_2 < s_1$. When a fast car encounters a slower one, it cannot pass it and the car has to reduce its speed to s_2. Show that the long-run average travel time per fast car equals $L/s_2 - (1/\lambda_2)[1 - \exp(-\lambda_2(L/s_2 - L/s_1))]$. (*Hint*: tag a fast car and express its travel time in terms of the time elapsed since the last slow car entered the highway.)

1.8 Let $\{N(t)\}$ be a Poisson process with interarrival times $X_1, X_2, \ldots$. Prove for any $t, s > 0$ that for all $n, k = 0, 1, \ldots$

$$P\{N(t+s) - N(t) \leq k, N(t) = n\} = P\{N(s) \leq k\}P\{N(t) = n\}.$$

In other words, the process has stationary and independent increments. (*Hint:* evaluate the probability $P\{X_1 + \cdots + X_n \leq t < X_1 + \cdots + X_{n+1}, X_1 + \cdots + X_{n+k+1} > t + s\}$.)

1.9 An information centre provides services in a bilingual environment. Requests for service arrive by telephone. Major language service requests and minor language service requests arrive according to independent Poisson processes with respective rates of λ_1 and λ_2 requests per hour. The service time of each request is exponentially distributed with a mean of $1/\mu_1$ minutes for a major language request and a mean of $1/\mu_2$ minutes for a minor language request.

(a) What is the probability that in the next hour a total of n service requests will arrive?

(b) What is the probability density of the service time of an arbitrarily chosen service request?

1.10 Short-term parkers and long-term parkers arrive at a parking lot according to independent Poisson processes with respective rates λ_1 and λ_2. The parking times of the customers are independent of each other. The parking time of a short-term parker has a uniform distribution on $[a_1, b_1]$ and that of a long-term parker has a uniform distribution on $[a_2, b_2]$. The parking lot has ample capacity.

(a) What is the mean parking time of an arriving car?

(b) What is the probability distribution of the number of occupied parking spots at any time $t > b_2$?

1.11 Oil tankers with world's largest harbour Rotterdam as destination leave from harbours in the Middle East according to a Poisson process with an average of two tankers per day. The sailing time to Rotterdam has a gamma distribution with an expected value of 10 days and a standard deviation of 4 days. What is the probability distribution of the number of oil tankers that are under way from the Middle East to Rotterdam at an arbitrary point in time?

1.12 Customers with items to repair arrive at a repair facility according to a Poisson process with rate λ. The repair time of an item has a uniform distribution on $[a, b]$. There are ample repair facilities so that each defective item immediately enters repair. The exact repair time can be determined upon arrival of the item. If the repair time of an item takes longer than τ time units with τ a given number between a and b, then the customer gets a loaner for the defective item until the item returns from repair. A sufficiently large supply of loaners is available. What is the average number of loaners which are out?

1.13 On a summer day, buses with tourists arrive in the picturesque village of Edam according to a Poisson process with an average of five buses per hour. The village of Edam is world famous for its cheese. Each bus stays either one hour or two hours in Edam with equal probabilities.

(a) What is the probability distribution of the number of tourist buses in Edam at 4 o'clock in the afternoon?

(b) Each bus brings 50, 75 or 100 tourists with respective probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$. Calculate a normal approximation to the probability that more than 1000 bus tourists are in Edam at 4 o'clock in the afternoon. (*Hint:* the number of bus tourists is distributed as the convolution of two compound Poisson distributions.)

1.14 Batches of containers arrive at a stockyard according to a Poisson process with rate λ. The batch sizes are independent random variables having a common discrete probability distribution $\{\beta_j,\ j = 1, 2, \dots\}$ with finite second moment. The stockyard has ample space to store any number of containers. The containers are temporarily stored at the stockyard. The holding times of the containers at the stockyard are independent random variables having a general probability distribution function $B(x)$ with finite mean μ. Also, the holding times of containers from the same batch are independent of each other. This model is called the batch-arrival $M^X/G/\infty$ queue with *individual service*. Let $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ be the generating function of the batch size and let $\{p_j\}$ denote the limiting distribution of the number of the containers present at the stockyard.

(a) Use Theorem 1.1.5 to prove that $P(z) = \sum_{j=0}^{\infty} p_j z^j$ is given by

$$P(z) = \exp\left(-\lambda \int_0^{\infty} [1 - \beta((1-z)B(x) + z)]\, dx\right).$$

(b) Verify that the mean m and the variance ν of the limiting distribution of the number of containers at the stockyard are given by

$$m = \lambda E(X)\mu \text{ and } \nu = \lambda E(X)\mu + \lambda E[X(X-1)] \int_0^{\infty} \{1 - B(x)\}^2\, dx,$$

where the random variable X has the batch-size distribution $\{\beta_j\}$.

(c) Investigate how good the approximation to $\{p_j\}$ performs when a negative binomial distribution is fitted to the mean m and the variance ν. Verify that this approximation is exact when the service times are exponentially distributed and the batch size is geometrically distributed with mean $\beta > 1$.

1.15 Consider Exercise 1.14 assuming this time that containers from the same batch are kept at the stockyard over the same holding time and are thus simultaneously removed. The holding times for the various batches have a general distribution function $B(x)$. This model is called the batch-arrival $M^X/G/\infty$ queue with *group service.*

(a) Argue that the limiting distribution $\{p_j\}$ of the number of containers present at the stockyard is *insensitive* to the form of the holding-time distribution and requires only its mean μ.

(b) Argue that the limiting distribution $\{p_j\}$ is a compound Poisson distribution with generating function $\exp(-\lambda D\{1 - \beta(z)\})$ with $D = \mu$.

1.16 In a certain region, traffic accidents occur according to a Poisson process. Calculate the probability that exactly one accident has occurred on each day of some week when it is given that seven accidents have occurred in that week. Can you explain why this probability is so small?

1.17 Suppose calls arrive at a computer-controlled exchange according to a Poisson process at a rate of 25 calls per second. Compute an approximate value for the probability that during the busy hour there is some period of 3 seconds in which 125 or more calls arrive.

1.18 In any given year claims arrive at an insurance company according to a Poisson process with an unknown parameter λ, where λ is the outcome of a gamma distribution with shape parameter α and scale parameter β. Prove that the total number of claims during a given year has a negative binomial distribution with parameters α and $\beta/(\beta + 1)$.

1.19 Claims arrive at an insurance company according to a Poisson process with rate λ. The claim sizes are independent random variables and have the common discrete distribution $a_k = -\alpha^k[k \ln(1-\alpha)]^{-1}$ for $k = 1, 2, \dots$, where α is a constant between 0 and 1. Verify

that the total amount claimed during a given year has a negative binomial distribution with parameters $-\lambda/\ln(1-\alpha)$ and $1-\alpha$.

1.20 An insurance company has two policies with fixed remittances. Claims from the policies 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. Each claim from policy i is for a fixed amount of c_i, where c_1 and c_2 are positive integers. Explain how to compute the probability distribution of the total amount claimed during a given time period.

1.21 It is only possible to place orders for a certain product during a random time T which has an exponential distribution with mean $1/\mu$. Customers who wish to place an order for the product arrive according to a Poisson process with rate λ. The amounts ordered by the customers are independent random variables $D_1, D_2, \ldots$ having a common discrete distribution $\{a_j,\ j = 1, 2, \ldots\}$.

(a) Verify that the mean m and the variance σ^2 of the total amount ordered during the random time T are given by

$$m = \frac{\lambda}{\mu} E(D_1) \quad \text{and} \quad \sigma^2 = \frac{\lambda}{\mu} E(D_1^2) + \frac{\lambda^2}{\mu^2} E^2(D_1).$$

(b) Let $\{p_k\}$ be the probability distribution of the total amount ordered during the random time T. Argue that the p_k can be recursively computed from

$$p_k = \frac{\lambda}{\lambda + \mu} \sum_{j=1}^{k} p_{k-j} a_j, \quad k = 1, 2, \ldots,$$

starting with $p_0 = \mu/(\lambda + \mu)$.

1.22 Consider a non-stationary Poisson arrival process with arrival rate function $\lambda(t)$. It is assumed that $\lambda(t)$ is continuous and bounded in t. Let $\lambda > 0$ be any upper bound on the function $\lambda(t)$. Prove that the arrival epochs of the non-stationary Poisson arrival process can be generated by the following procedure:

(a) Generate arrival epochs of a Poisson process with rate λ.

(b) Thin out the arrival epochs by accepting an arrival occurring at epoch s with probability $\lambda(s)/\lambda$ and rejecting it otherwise.

1.23 Customers arrive at an automatic teller machine in accordance with a non-stationary Poisson process. From 8 am until 10 am customers arrive at a rate of 5 an hour. Between 10 am and 2 pm the arrival rate steadily increases from 5 per hour at 10 am to 25 per hour at 2 pm. From 2 pm to 8 pm the arrival rate steadily decreases from 25 per hour at 2 pm to 4 per hour at 8 pm. Between 8 pm and midnight the arrival rate is 3 an hour and from midnight to 8 am the arrival rate is 1 per hour. The amounts of money withdrawn by the customers are independent and identically distributed random variables with a mean of \$100 and a standard deviation of \$125.

(a) What is the probability distribution of the number of customers withdrawing money during a 24-hour period?

(b) Calculate an approximation to the probability that the total withdrawal during 24 hours is more than \$25 000.

1.24 Parking-fee dodgers enter the parking lot of the University of Amsterdam according to a Poisson process with rate λ. The parking lot has ample capacity. Each fee dodger parks his/her car during an Erlang $(2, \mu)$ distributed time. It is university policy to inspect the parking lot every T time units, with T fixed. Each newly arrived fee dodger is fined. What is the probability distribution of the number of fee dodgers who are fined at an inspection?

1.25 Suppose customers arrive according to a non-stationary Poisson process with arrival rate function $\lambda(t)$. Any newly arriving customer is marked as a type k customer with probability p_k for $k = 1, \ldots, L$, independently of the other customers. Prove that the customers of

the types $1, \ldots, L$ arrive according to independent non-stationary Poisson processes with respective arrival rate functions $p_1\lambda(t), \ldots, p_L\lambda(t)$.

1.26 Consider the infinite-server queueing model from Section 1.1.3, but assume now that customers arrive according to a non-stationary Poisson process with arrival rate function $\lambda(t)$. Let $B(x)$ be the probability distribution function of the service time of a customer. Assuming that the system is empty at epoch 0, prove that the number of busy servers at time t has a Poisson distribution with mean $\int_0^t \lambda(x)\{1 - B(t-x)\}dx$.

1.27 Consider the $M/G/\infty$ queue from Section 1.1.3 again. Let the random variable L be the length of a busy period. A busy period begins when an arrival finds the system empty and finishes when there are no longer any customers in the system. Argue that $P\{L > t\}$ can be obtained from the integral equation

$$P\{L > t\} = 1 - B(t) + \int_0^t \{B(t) - B(x)\}P\{L > t - x\}\lambda e^{-\lambda x}dx, \quad t \geq 0,$$

where $B(t)$ is the probability distribution function of the service time of a customer. *Remark*: it was shown in Shanbhag (1966) that the Laplace transform of $P\{L > t\}$ is given by

$$\frac{1}{s}\left(1 - \frac{\lambda + s}{\lambda} + \frac{1}{\lambda}\left\{\int_0^\infty \exp\left(-sx - \lambda\int_0^x (1 - B(y))dy\right)dx\right\}^{-1}\right).$$

BIBLIOGRAPHIC NOTES

A treatment of the Poisson process can be found in numerous texts. A good treatment is given in the books of Ross (1996) and Wolff (1989). The Poisson process is fundamental to all areas of applied probability. The infinite-server queue with Poisson input has many applications. The applications in Examples 1.1.3 and 1.1.4 are taken from papers of Parikh (1977) and Sherbrooke (1968).

REFERENCES

Adelson, R.M. (1966) Compound Poisson distributions. *Operat. Res. Quart.* **17**, 73–75.

Den Iseger, P.W., Smith, M.A.J. and Dekker, R. (1997) Computing compound Poisson distributions faster. *Insurance Mathematics and Economics*, **20**, 23–34.

Glaz, J. and Balakrishnan, N. (1999) *Scan Statistics and Applications*. Birkhäuser, Boston.

Khintchine, A.Y. (1969) *Mathematical Methods in the Theory of Queueing*. Hafter, New York.

Parikh, S.C. (1977) On a fleet sizing and allocation problem. *Management Sci.*, **23**, 972–977.

Ross, S.M. (1996) *Stochastic Processes*, 2nd edn. John Wiley & Sons, Inc., New York.

Shanbhag, D.N. (1966) On infinite server queues with batch arrivals. *J. Appl. Prob.*, **3**, 274–279.

Sherbrooke, C.C. (1968) Metric: a multi-echelon technique for recoverable item control, *Operat. Res.*, **16**, 122–141.

Wolff, R.W. (1989) *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.

CHAPTER 2

Renewal-Reward Processes

2.0 INTRODUCTION

The renewal-reward model is an extremely useful tool in the analysis of applied probability models for inventory, queueing and reliability applications, among others. Many stochastic processes are regenerative; that is, they regenerate themselves from time to time so that the behaviour of the process after the regeneration epoch is a probabilistic replica of the behaviour of the process starting at time zero. The time interval between two regeneration epochs is called a cycle. The sequence of regeneration cycles constitutes a so-called renewal process. The long-run behaviour of a regenerative stochastic process on which a reward structure is imposed can be studied in terms of the behaviour of the process during a single regeneration cycle. The simple and intuitively appealing renewal-reward model has numerous applications.

In Section 2.1 we first discuss some elementary results from renewal theory. A more detailed treatment of renewal theory will be given in Chapter 8. Section 2.2 deals with the renewal-reward model. It shows how to calculate long-run averages such as the long-run average reward per time unit and the long-run fraction of time the system spends in a given set of states. Illustrative examples will be given. Section 2.3 discusses the formula of Little. This formula is a kind of law of nature and relates among others the average queue size to the average waiting time in queueing systems. Another fundamental result that is frequently used in queueing and inventory applications is the property that Poisson arrivals see time averages (PASTA). This result is discussed in some detail in Section 2.4. The PASTA property is used in Section 2.5 to obtain the famous Pollaczek–Khintchine formula from queueing theory. The renewal-reward model is used in Section 2.6 to obtain a generalization of the Pollaczek–Khintchine formula in the framework of a controlled queue. Section 2.7 shows how renewal theory and an up- and down-crossing argument can be combined to derive a relation between time-average and customer-average probabilities in queues.

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

2.1 RENEWAL THEORY

As a generalization of the Poisson process, renewal theory concerns the study of stochastic processes counting the number of events that take place as a function of time. Here the interoccurrence times between successive events are independent and identically distributed random variables. For instance, the events could be the arrival of customers to a waiting line or the successive replacements of light bulbs. Although renewal theory originated from the analysis of replacement problems for components such as light bulbs, the theory has many applications to quite a wide range of practical probability problems. In inventory, queueing and reliability problems, the analysis is often based on an appropriate identification of embedded renewal processes for the specific problem considered. For example, in a queueing process the embedded events could be the arrival of customers who find the system empty, or in an inventory process the embedded events could be the replenishment of stock when the inventory position drops to the reorder point or below it.

Formally, let $X_1, X_2, \ldots$ be a sequence of non-negative, independent random variables having a common probability distribution function

$$F(x) = P\{X_k \leq x\}, \quad x \geq 0$$

for $k = 1, 2, \ldots$. Letting $\mu_1 = E(X_k)$, it is assumed that

$$0 < \mu_1 < \infty.$$

The random variable X_n denotes the interoccurrence time between the $(n-1)$th and nth event in some specific probability problem. Define

$$S_0 = 0 \quad \text{and} \quad S_n = \sum_{i=1}^{n} X_i, \quad n = 1, 2, \ldots .$$

Then S_n is the epoch at which the nth event occurs. For each $t \geq 0$, let

$$N(t) = \text{the largest integer } n \geq 0 \text{ for which } S_n \leq t.$$

Then the random variable $N(t)$ represents the number of events up to time t.

Definition 2.1.1 *The counting process $\{N(t), t \geq 0\}$ is called the renewal process generated by the interoccurrence times $X_1, X_2, \ldots$.*

It is said that a renewal occurs at time t if $S_n = t$ for some n. For each $t \geq 0$, the number of renewals up to time t is finite with probability 1. This is an immediate consequence of the strong law of large numbers stating that $S_n/n \to E(X_1)$ with probability 1 as $n \to \infty$ and thus $S_n \leq t$ only for finitely many n. The Poisson process is a special case of a renewal process. Here we give some other examples of a renewal process.

Example 2.1.1 A replacement problem

Suppose we have an infinite supply of electric bulbs, where the burning times of the bulbs are independent and identically distributed random variables. If the bulb in use fails, it is immediately replaced by a new bulb. Let X_i be the burning time of the ith bulb, $i = 1, 2, \ldots$. Then $N(t)$ is the total number of bulbs to be replaced up to time t.

Example 2.1.2 An inventory problem

Consider a periodic-review inventory system for which the demands for a single product in the successive weeks $t = 1, 2, \ldots$ are independent random variables having a common *continuous* distribution. Let X_i be the demand in the ith week, $i = 1, 2, \ldots$. Then $1 + N(u)$ is the number of weeks until depletion of the current stock u.

2.1.1 The Renewal Function

An important role in renewal theory is played by the *renewal function* $M(t)$ which is defined by

$$M(t) = E[N(t)], \quad t \geq 0. \tag{2.1.1}$$

For $n = 1, 2, \ldots$, define the probability distribution function

$$F_n(t) = P\{S_n \leq t\}, \quad t \geq 0.$$

Note that $F_1(t) = F(t)$. A basic relation is

$$N(t) \geq n \quad \text{if and only if} \quad S_n \leq t. \tag{2.1.2}$$

This relation implies that

$$P\{N(t) \geq n\} = F_n(t), \quad n = 1, 2, \ldots . \tag{2.1.3}$$

Lemma 2.1.1 *For any $t \geq 0$,*

$$M(t) = \sum_{n=1}^{\infty} F_n(t). \tag{2.1.4}$$

Proof Since for any non-negative integer-valued random variable N,

$$E(N) = \sum_{k=0}^{\infty} P\{N > k\} = \sum_{n=1}^{\infty} P\{N \geq n\},$$

the relation (2.1.4) is an immediate consequence of (2.1.3).

In Exercise 2.4 the reader is asked to prove that $M(t) < \infty$ for all $t \geq 0$. In Chapter 8 we will discuss how to compute the renewal function $M(t)$ in general. The infinite series (2.1.4) is in general not useful for computational purposes. An exception is the case in which the interoccurrence times $X_1, X_2, \ldots$ have a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. Then the sum $X_1 + \cdots + X_n$ has a gamma distribution with shape parameter $n\alpha$ and scale parameter λ. In this case $F_n(t)$ is the so-called incomplete gamma integral for which efficient numerical procedures are available; see Appendix B. Let us explain this in more detail for the case that α is a positive integer r so that the interoccurrence times $X_1, X_2, \ldots$ have an Erlang (r, λ) distribution with scale parameter λ. Then $F_n(t)$ becomes the Erlang (nr, λ) distribution function

$$F_n(t) = 1 - \sum_{k=0}^{nr-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad t \geq 0$$

and thus

$$M(t) = \sum_{n=1}^{\infty} \left[1 - \sum_{k=0}^{nr-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \right], \quad t \geq 0. \tag{2.1.5}$$

In this particular case $M(t)$ can be efficiently computed from a rapidly converging series. For the special case that the interoccurrence times are exponentially distributed ($r = 1$), the expression (2.1.5) reduces to the explicit formula

$$M(t) = \lambda t, \quad t \geq 0.$$

This finding is in agreement with earlier results for the Poisson process.

Remark 2.1.1 The phase method

A very useful interpretation of the renewal process $\{N(t)\}$ can be given when the interoccurrence times $X_1, X_2, \ldots$ have an Erlang distribution. Imagine that tokens arrive according to a Poisson process with rate λ and that the arrival of each rth token triggers the occurrence of an event. Then the events occur according to a renewal process in which the interoccurrence times have an Erlang (r, λ) distribution with scale parameter λ. The explanation is that the sum of r independent, exponentially distributed random variables with the same scale parameter λ has an Erlang (r, λ) distribution. The phase method enables us to give a tractable expression of the probability distribution of $N(t)$ when the interoccurrence times have an Erlang (r, λ) distribution. In this case $P\{N(t) \geq n\}$ is equal to the probability that nr or more arrivals occur in a Poisson arrival process with rate λ. You are asked to work out the equivalence in Exercise 2.5.

Asymptotic expansion

A very useful asymptotic expansion for the renewal function $M(t)$ can be given under a weak regularity condition on the interoccurrence times. This condition

will be formulated in Section 8.2. For the moment it is sufficient to assume that the interoccurrence times have a positive density on some interval. Further it is assumed that $\mu_2 = E(X_1^2)$ is finite. Then it will be shown in Theorem 8.2.3 that

$$\lim_{t\to\infty}\left[M(t) - \frac{t}{\mu_1}\right] = \frac{\mu_2}{2\mu_1^2} - 1. \tag{2.1.6}$$

The approximation

$$M(t) \approx \frac{t}{\mu_1} + \frac{\mu_2}{2\mu_1^2} - 1 \quad \text{for } t \text{ large}$$

is practically useful for already moderate values of t provided that the squared coefficient of variation of the interoccurrence times is not too large and not too close to zero.

2.1.2 The Excess Variable

In many practical probability problems an important quantity is the random variable γ_t defined as the time elapsed from epoch t until the next renewal after epoch t. More precisely, γ_t is defined as

$$\gamma_t = S_{N(t)+1} - t;$$

see also Figure 2.1.1 in which a renewal epoch is denoted by $\times$. Note that $S_{N(t)+1}$ is the epoch of the first renewal that occurs after time t. The random variable γ_t is called the *excess* or *residual life* at time t. For the replacement problem of Example 2.1.1 the random variable γ_t denotes the residual lifetime of the light bulb in use at time t.

Lemma 2.1.2 *For any $t \geq 0$,*

$$E(\gamma_t) = \mu_1[1 + M(t)] - t. \tag{2.1.7}$$

Proof Fix $t \geq 0$. To prove (2.1.7), we apply Wald's equation from Appendix A. To do so, note that $N(t) \leq n - 1$ if and only if $X_1 + \cdots + X_n > t$. Hence the event $\{N(t) + 1 = n\}$ depends only on $X_1, \ldots, X_n$ and is thus independent of $X_{n+1}, X_{n+2}, \ldots$. Hence

$$E\left[\sum_{k=1}^{N(t)+1} X_k\right] = E(X_1)E[N(t) + 1],$$

which gives (2.1.7).

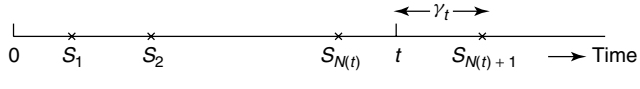


Figure 2.1.1 The excess life

In Corollary 8.2.4 it will be shown that

$$\lim_{t\to\infty} E(\gamma_t) = \frac{\mu_2}{2\mu_1} \quad \text{and} \quad \lim_{t\to\infty} E(\gamma_t^2) = \frac{\mu_3}{3\mu_1} \tag{2.1.8}$$

with $\mu_k = E(X_1^k)$ for $k = 1, 2, 3$, provided that the interoccurrence times have a positive density on some interval. An illustration of the usefulness of the concept of excess variable is provided by the next example.

Example 2.1.3 The average order size in an (s, S) ***inventory system***

Suppose a periodic-review inventory system for which the demands $X_1, X_2, \ldots$ for a single product in the successive weeks $1, 2, \ldots$ are independent random variables having a common probability density $f(x)$ with finite mean α and finite standard deviation σ. Any demand exceeding the current inventory is backlogged until inventory becomes available by the arrival of a replenishment order. The inventory position is reviewed at the beginning of each week and is controlled by an (s, S) rule with $0 \leq s < S$. Under this control rule, a replenishment order of size $S - x$ is placed when the review reveals that the inventory level x is below the reorder point s; otherwise, no ordering is done. We assume instantaneous delivery of every replenishment order.

We are interested in the average order size. Since the inventory process starts from scratch each time the inventory position is ordered up to level S, the operating characteristics can be calculated by using a renewal model in which the weekly demand sizes $X_1, X_2, \ldots$ represent the interoccurrence times of renewals. The number of weeks between two consecutive orderings equals the number of weeks needed for a cumulative demand larger than $S - s$. The order size is the sum of $S - s$ and the undershoot of the reorder point s at the epoch of ordering (see Figure 2.1.2 in which a renewal occurrence is denoted by an $\times$). Denote by $\{N(t)\}$ the renewal process associated with the weekly demands $X_1, X_2, \ldots$. Then the number of weeks needed for a cumulative demand exceeding $S - s$ is given by $1 + N(S - s)$. The undershoot of the reorder point s is just the excess life γ_{S-s} of the renewal process. Hence

$$E[\text{order size}] = S - s + E(\gamma_{S-s}).$$

From (2.1.8) it follows that the average order size can be approximated by

$$E[\text{order size}] \approx S - s + \frac{\sigma^2 + \alpha^2}{2\alpha}$$

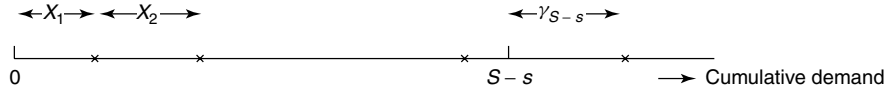


Figure 2.1.2 The inventory process modelled as a renewal process

provided that $S - s$ is sufficiently large compared with E(weekly demand). In practice this is a useful approximation for $S-s > \alpha$ when the weekly demand is not highly variable and has a squared coefficient of variation between 0.2 and 1 (say).

Another illustration of the importance of the excess variable is given by the famous waiting-time paradox.

Example 2.1.4 The waiting-time paradox

We have all experienced long waits at a bus stop when buses depart irregularly and we arrive at the bus stop at random. A theoretical explanation of this phenomenon is provided by the expression for $\lim_{t\to\infty} E(\gamma_t)$. Therefore it is convenient to rewrite (2.1.8) as

$$\lim_{t\to\infty} E(\gamma_t) = \frac{1}{2}(1 + c_X^2)\mu_1, \tag{2.1.9}$$

where

$$c_X^2 = \frac{\sigma^2(X_1)}{E^2(X_1)}$$

is the squared coefficient of variation of the interdeparture times $X_1, X_2, \ldots$. The equivalent expression (2.1.9) follows from (2.1.8) by noting that

$$1 + c_X^2 = 1 + \frac{\mu_2 - \mu_1^2}{\mu_1^2} = \frac{\mu_2}{\mu_1^2}. \tag{2.1.10}$$

The representation (2.1.9) makes clear that

$$\lim_{t\to\infty} E(\gamma_t) = \begin{cases} < \mu_1 & \text{if } c_X^2 < 1, \\ > \mu_1 & \text{if } c_X^2 > 1. \end{cases}$$

Thus the mean waiting time for the next bus depends on the regularity of the bus service and increases with the coefficient of variation of the interdeparture times. If we arrive at the bus stop at random, then for highly irregular service ($c_X^2 > 1$) the mean waiting time for the next bus is even larger than the mean interdeparture time. This surprising result is sometimes called the *waiting-time paradox*. A heuristic explanation is that it is more likely to hit a long interdeparture time than a short one when arriving at the bus stop at random. To illustrate this, consider the extreme situation in which the interdeparture time is 0 minutes with probability 9/10 and is 10 minutes with probability 1/10. Then the mean interdeparture time is 1 minute, but your mean waiting time for the next bus is 5 minutes when you arrive at the bus stop at random.

2.2 RENEWAL-REWARD PROCESSES

A powerful tool in the analysis of numerous applied probability models is the renewal-reward model. This model is also very useful for theoretical purposes. In

Chapters 3 and 4, ergodic theorems for Markov chains will be proved by using the renewal-reward theorem. The renewal-reward model is a simple and intuitively appealing model that deals with a so-called regenerative process on which a cost or reward structure is imposed. Many stochastic processes have the property of regenerating themselves at certain points in time so that the behaviour of the process after the regeneration epoch is a probabilistic replica of the behaviour starting at time zero and is independent of the behaviour before the regeneration epoch.

A formal definition of a regenerative process is as follows.

Definition 2.2.1 *A stochastic process* $\{X(t), t \in T\}$ *with time-index set* T *is said to be regenerative if there exists a (random) epoch* S_1 *such that:*

(a) $\{X(t + S_1), t \in T\}$ *is independent of* $\{X(t), 0 \leq t < S_1\}$,

(b) $\{X(t + S_1), t \in T\}$ *has the same distribution as* $\{X(t), t \in T\}$.

It is assumed that the index set T is either the interval $T = [0, \infty)$ or the countable set $T = \{0, 1, \dots\}$. In the former case we have a continuous-time regenerative process and in the other case a discrete-time regenerative process. The state space of the process $\{X(t)\}$ is assumed to be a subset of some Euclidean space.

The existence of the regeneration epoch S_1 implies the existence of further regeneration epochs $S_2, S_3, \dots$ having the same property as S_1. Intuitively speaking, a regenerative process can be split into independent and identically distributed renewal cycles. A *cycle* is defined as the time interval between two consecutive regeneration epochs. Examples of regenerative processes are:

(i) The continuous-time process $\{X(t), t \geq 0\}$ with $X(t)$ denoting the number of customers present at time t in a single-server queue in which the customers arrive according to a renewal process and the service times are independent and identically distributed random variables. It is assumed that at epoch 0 a customer arrives at an empty system. The regeneration epochs $S_1, S_2, \dots$ are the epochs at which an arriving customer finds the system empty.

(ii) The discrete-time process $\{I_n, n = 0, 1, \dots\}$ with I_n denoting the inventory level at the beginning of the nth week in the (s, S) inventory model dealt with in Example 2.1.3. Assume that the inventory level equals S at epoch 0. The regeneration epochs are the beginnings of the weeks in which the inventory level is ordered up to the level S.

Let us define the random variables $C_n = S_n - S_{n-1}$, $n = 1, 2, \dots$, where $S_0 = 0$ by convention. The random variables $C_1, C_2, \dots$ are independent and identically distributed. In fact the sequence $\{C_1, C_2, \dots\}$ underlies a renewal process in which the events are the occurrences of the regeneration epochs. Hence we can interpret C_n as

$$C_n = \text{the length of the } n\text{th renewal cycle}, \quad n = 1, 2, \dots .$$

Note that the cycle length C_n assumes values from the index set T. In the following it is assumed that

$$0 < E(C_1) < \infty.$$

In many practical situations a reward structure is imposed on the regenerative process $\{X(t), t \in T\}$. The reward structure usually consists of reward rates that are earned continuously over time and lump rewards that are only earned at certain state transitions. Let

$$R_n = \text{the total reward earned in the } n\text{th renewal cycle}, \quad n = 1, 2, \ldots .$$

It is assumed that $R_1, R_2, \ldots$ are independent and identically distributed random variables. In applications R_n typically depends on C_n. In case R_n can take on both positive and negative values, it is assumed that $E(|R_1|) < \infty$. Let

$$R(t) = \text{the cumulative reward earned up to time } t.$$

The process $\{R(t), t \geq 0\}$ is called a *renewal-reward process.* We are now ready to prove a theorem of utmost importance.

Theorem 2.2.1 (renewal-reward theorem)

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{E(R_1)}{E(C_1)} \quad \textit{with probability } 1.$$

In other words, for almost any realization of the process, the long-run average reward per time unit is equal to the expected reward earned during one cycle divided by the expected length of one cycle.

To prove this theorem we first establish the following lemma.

Lemma 2.2.2 *For any $t \geq 0$, let $N(t)$ be the number of cycles completed up to time t. Then*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{E(C_1)} \quad \textit{with probability } 1.$$

Proof By the definition of $N(t)$, we have

$$C_1 + \cdots + C_{N(t)} \leq t < C_1 + \cdots + C_{N(t)+1}.$$

Since $P\{C_1 + \cdots + C_n < \infty\} = 1$ for all $n \geq 1$, it is not difficult to verify that

$$\lim_{t\to\infty} N(t) = \infty \quad \text{with probability } 1.$$

The above inequality gives

$$\frac{C_1 + \cdots + C_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{C_1 + \cdots + C_{N(t)+1}}{N(t)+1} \frac{N(t)+1}{N(t)}.$$

By the strong law of large numbers for a sequence of independent and identically distributed random variables, we have

$$\lim_{n\to\infty} \frac{C_1 + \cdots + C_n}{n} = E(C_1) \quad \text{with probability 1.}$$

Hence, by letting $t \to \infty$ in the above inequality, the desired result follows.

Lemma 2.2.2 is also valid when $E(C_1) = \infty$ provided that $P\{C_1 < \infty\} = 1$. The reason is that the strong law of large numbers for a sequence $\{C_n\}$ of *non-negative* random variables does not require that $E(C_1) < \infty$. Next we prove Theorem 2.2.1.

Proof of Theorem 2.2.1 For ease, let us first assume that the rewards are non-negative. Then, for any $t > 0$,

$$\sum_{i=1}^{N(t)} R_i \le R(t) \le \sum_{i=1}^{N(t)+1} R_i.$$

This gives

$$\frac{\sum_{i=1}^{N(t)} R_i}{N(t)} \times \frac{N(t)}{t} \le \frac{R(t)}{t} \le \frac{\sum_{i=1}^{N(t)+1} R_i}{N(t)+1} \times \frac{N(t)+1}{t}.$$

By the strong law of large numbers for the sequence $\{R_n\}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_i = E(R_1) \quad \text{with probability 1.}$$

As pointed out in the proof of Lemma 2.2.2, $N(t) \to \infty$ with probability 1 as $t \to \infty$. Letting $t \to \infty$ in the above inequality and using Lemma 2.2.2, the desired result next follows for the case that the rewards are non-negative. If the rewards can assume both positive and negative values, then the theorem is proved by treating the positive and negative parts of the rewards separately. We omit the details.

In a natural way Theorem 2.2.1 relates the behaviour of the renewal-reward process over time to the behaviour of the process over a single renewal cycle. It is noteworthy that the outcome of the long-run average actual reward per time unit can be predicted with probability 1. If we are going to run the process over an infinitely long period of time, then we can say beforehand that in the long run the average *actual* reward per time unit will be equal to the constant $E(R_1)/E(C_1)$ with probability 1. This is a much stronger and more useful statement than the statement that the long-run *expected* average reward per time unit equals $E(R_1)/E(C_1)$ (it indeed holds that $\lim_{t\to\infty} E[R(t)]/t = E(R_1)/E(C_1)$; this *expected-value version* of the renewal-reward theorem is a direct consequence of Theorem 2.2.1 when $R(t)/t$ is bounded in t but otherwise requires a hard proof). Also it is noted that for the case of non-negative rewards R_n the renewal-reward theorem is also valid when $E(R_1) = \infty$ (the assumption $E(C_1) < \infty$ cannot be dropped for Theorem 2.2.1).

Example 2.2.1 Alternating up- and downtimes

Suppose a machine is alternately up and down. Denote by $U_1, U_2, \ldots$ the lengths of the successive up-periods and by $D_1, D_2, \ldots$ the lengths of the successive down-periods. It is assumed that both $\{U_n\}$ and $\{D_n\}$ are sequences of independent and identically distributed random variables with finite positive expectations. The sequences $\{U_n\}$ and $\{D_n\}$ are not required to be independent of each other. Assume that an up-period starts at epoch 0. What is the long-run fraction of time the machine is down? The answer is

$$\text{the long-run fraction of time the machine is down} = \frac{E(D_1)}{E(U_1) + E(D_1)} \quad \text{with probability 1.} \tag{2.2.1}$$

To verify this, define the continuous-time stochastic process $\{X(t), t \geq 0\}$ by

$$X(t) = \begin{cases} 1 & \text{if the machine is up at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The process $\{X(t)\}$ is a regenerative process. The epochs at which an up-period starts can be taken as regeneration epochs. The long-run fraction of time the machine is down can be interpreted as a long-run average cost per time unit by assuming that a cost at rate 1 is incurred while the machine is down and a cost at rate 0 otherwise. A regeneration cycle consists of an up-period and a down-period. Hence

$$E(\text{length of one cycle}) = E(U_1 + D_1)$$

and

$$E(\text{cost incurred during one cycle}) = E(D_1).$$

By applying the renewal-reward theorem, it follows that the long-run average cost per time unit equals $E(D_1)/[E(U_1) + E(D_1)]$, proving the result (2.2.1).

The intermediate step of interpreting the long-run fraction of time that the process is in a certain state as a long-run average cost (reward) per time unit is very helpful in many situations.

Limit theorems for regenerative processes

An important application of the renewal-reward theorem is the characterization of the long-run fraction of time a regenerative process $\{X(t), t \in T\}$ spends in some given set B of states. For the set B of states, define for any $t \in T$ the indicator variable

$$I_B(t) = \begin{cases} 1 & \text{if } X(t) \in B, \\ 0 & \text{if } X(t) \notin B. \end{cases}$$

Also, define the random variable

T_B = the amount of time the process spends in the set B of states during one cycle.

Note that $T_B = \int_0^{S_1} I_B(u)\,du$ for a continuous-time process $\{X(t)\}$; otherwise, T_B equals the number of indices $0 \le k < S_1$ with $X(k) \in B$. The following theorem is an immediate consequence of the renewal-reward theorem.

Theorem 2.2.3 *For the regenerative process $\{X(t)\}$ it holds that the long-run fraction of time the process spends in the set B of states is $E(T_B)/E(C_1)$ with probability 1.*

That is,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t I_B(u)\,du = \frac{E(T_B)}{E(C_1)} \quad \text{with probability 1}$$

for a continuous-time process $\{X(t)\}$ and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} I_B(k) = \frac{E(T_B)}{E(C_1)} \quad \text{with probability 1}$$

for a discrete-time process $\{X(n)\}$.

Proof The long-run fraction of time the process $\{X(t)\}$ spends in the set B of states can be interpreted as a long-run average reward per time unit by assuming that a reward at rate 1 is earned while the process is in the set B and a reward at rate 0 is earned otherwise. Then

$$E(\text{reward earned during one cycle}) = E(T_B).$$

The desired result next follows by applying the renewal-reward theorem.

Since $E(I_B(t)) = P\{X(t) \in B\}$, we have as consequence of Theorem 2.2.3 and the bounded convergence theorem that, for a continuous-time process,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t P\{X(u) \in B\}\,du = \frac{E(T_B)}{E(C_1)}.$$

Note that $(1/t)\int_0^t P\{X(u) \in B\}\,du$ can be interpreted as the probability that an outside observer arriving at a randomly chosen point in $(0, t)$ finds the process in the set B.

In many situations the ratio $E(T_B)/E(C_1)$ could be interpreted both as the long-run fraction of time the process $\{X(t)\}$ spends in the set B of states and as the probability of finding the process in the set B when the process has reached statistical equilibrium. This raises the question whether $\lim_{t\to\infty} P\{X(t) \in B\}$ always exists. This ordinary limit need not always exist. A counterexample is provided by

periodic discrete-time Markov chains; see Chapter 3. For completeness we state the following theorem.

Theorem 2.2.4 *For the regenerative process* $\{X(t), t \in T\}$,

$$\lim_{t\to\infty} P\{X(t) \in B\} = \frac{E(T_B)}{E(C_1)}$$

provided that the probability distribution of the cycle length has a continuous part in the continuous-time case and is aperiodic in the discrete-time case.

A distribution function is said to have a continuous part if it has a positive density on some interval. A discrete distribution $\{a_j, j = 0, 1, \ldots\}$ is said to be aperiodic if the greatest common divisor of the indices $j \geq 1$ for which $a_j > 0$ is equal to 1. The proof of Theorem 2.2.4 requires deep mathematics and is beyond the scope of this book. The interested reader is referred to Miller (1972). It is remarkable that the proof of Theorem 2.2.3 for the time-average limit $\lim_{t\to\infty} (1/t) \int_0^t I_B(u)\, du$ is much simpler than the proof of Theorem 2.2.4 for the ordinary limit $\lim_{t\to\infty} P\{X(t) \in B\}$. This is all the more striking when we take into account that the time-average limit is in general much more useful for practical purposes than the ordinary limit. Another advantage of the time-average limit is that it is easier to understand than the ordinary limit. In interpreting the ordinary limit one should be quite careful. The ordinary limit represents the probability that an outside person will find the process in some state of the set B when inspecting the process at an arbitrary point in time after the process has been in operation for a very long time. It is essential for this interpretation that the outside person has *no information* about the past of the process when inspecting the process. How much more concrete is the interpretation of the time-average limit as the long-run fraction of time the process will spend in the set B of states!

To illustrate Theorem 2.2.4, consider again Example 2.2.1. In this example we analysed the long-run average behaviour of the regenerative process $\{X(t)\}$, where $X(t) = 1$ if the machine is up at time t and $X(t) = 0$ otherwise. It was shown that the long-run fraction of time the machine is down equals $E(D)/[E(U) + E(D)]$, where the random variables U and D denote the lengths of an up-period and a down-period. This result does not require any assumption about the shapes of the probability distributions of U and D. However, some assumption is needed in order to conclude that

$$\lim_{t\to\infty} P\{\text{the system is down at time } t\} = \frac{E(D)}{E(U) + E(D)}. \tag{2.2.2}$$

It is sufficient to assume that the distribution function of the length of an up-period has a positive density on some interval.

We state without proof a central limit theorem for the renewal-reward process.

Theorem 2.2.5 *Assume that* $R(t) \geq 0$ *with* $E(C_1^2) < \infty$ *and* $E(R_1^2) < \infty$. *Then*

$$\lim_{t\to\infty} P\left\{\frac{R(t) - gt}{v\sqrt{t/\mu_1}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy, \quad x \geq 0,$$

where $\mu_1 = E(C_1)$, $\mu_2 = E(C_1^2)$, $g = E(R_1)/E(C_1)$ *and* $v^2 = E(R_1 - gC_1)^2$.

A proof of this theorem can be found in Wolff (1989). In applying this theorem, the difficulty is usually to find the constant v. In specific applications one might use simulation to find v. As a special case, Theorem 2.2.5 includes a central limit theorem for the renewal process $\{N(t)\}$ studied in Section 2.1. Taking the rewards R_n equal to 1 it follows that the renewal process $\{N(t)\}$ is asymptotically $N(t/\mu_1, \sigma^2 t/\mu_1^3)$ distributed with $\sigma^2 = \mu_2 - \mu_1^2$.

Next we give two illustrative examples of the renewal-reward model.

Example 2.2.2 A stochastic clearing system

In a communication system messages requiring transmission arrive according to a Poisson process with rate λ. The messages are temporarily stored in a buffer having ample capacity. Every T time units, the buffer is cleared from all messages present. The buffer is empty at time $t = 0$. A fixed cost of $K > 0$ is incurred for each clearing of the buffer. Also, for each message there is a holding cost of $h > 0$ for each time unit the message has to wait in the buffer. What is the value of T for which the long-run average cost per time unit is minimal?

We first derive an expression for the average cost per time unit for a given value of the control parameter T. To do so, observe that the stochastic process describing the number of messages in the system regenerates itself each time the buffer is cleared from all messages present. This fact uses the lack of memory of the Poisson arrival process so that at any clearing epoch it is not relevant how long ago the last message arrived. Taking a cycle as the time interval between two successive clearings of the buffer, we have

$$\text{the expected length of one cycle } = T.$$

To specify the expected cost incurred during one cycle, we need an expression for the total waiting time of all messages arriving during one cycle. It was shown in Example 1.1.4 that

$$E[\text{total waiting time in } (0, T)] = \frac{1}{2}\lambda T^2.$$

This gives

$$E[\text{cost incurred during one cycle}] = K + \frac{1}{2}h\lambda T^2.$$

Hence, by the renewal-reward theorem,

$$\text{the long-run average cost per time unit } = \frac{1}{T}\left(K + \frac{1}{2}h\lambda T^2\right)$$

with probability 1. When $K = 0$ and $h = 1$, the system incurs a cost at rate j whenever there are j messages in the buffer, in which case the average cost per time unit gives the average number of messages in the buffer. Hence

$$\text{the long-run average number of messages in the buffer} = \frac{1}{2}\lambda T.$$

Putting the derivative of the cost function equal to 0, it follows that the long-run average cost is minimal for

$$T^* = \sqrt{\frac{2K}{h\lambda}}.$$

Example 2.2.3 A reliability system with redundancies

An electronic system consists of a number of independent and identical components hooked up in parallel. The lifetime of each component has an exponential distribution with mean $1/\mu$. The system is operative only if m or more components are operating. The non-failed units remain in operation when the system as a whole is in a non-operative state. The system availability is increased by periodic maintenance and by putting r redundant components into operation in addition to the minimum number m of components required. Under the periodic maintenance the system is inspected every T time units, where at inspection the failed components are repaired. The repair time is negligible and each repaired component is again as good as new. The periodic inspections provide the only repair opportunities. The following costs are involved. For each component there is a depreciation cost of $I > 0$ per time unit. A fixed cost of $K > 0$ is made for each inspection and there is a repair cost of $R > 0$ for each failed component. How can we choose the number r of redundant components and the time T between two consecutive inspections such that the long-run average cost per time unit is minimal subject to the requirement that the probability of system failure between two inspections is no more than a prespecified value α?

We first derive the performance measures for given values of the parameters r and T. The stochastic process describing the number of operating components is regenerative. Using the lack of memory of the exponential lifetimes of the components, it follows that the process regenerates itself after each inspection. Taking a cycle as the time interval between two inspections, we have

$$E(\text{length of one cycle}) = T.$$

Further, using the fact that a given component fails within a time T with probability $1 - e^{-\mu T}$, it follows that

$$P\{\text{the system as a whole fails between two inspections}\}$$
$$= \sum_{k=r+1}^{m+r} \binom{m+r}{k} (1 - e^{-\mu T})^k e^{-\mu T(m+r-k)}$$

and

$$E(\text{number of components that fail between two inspections})$$
$$= (m + r)(1 - e^{-\mu T}).$$

Hence

$$E(\text{total costs in one cycle}) = (m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R.$$

This gives

$$\text{the long-run average cost per time unit}$$
$$= \frac{1}{T}[(m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R]$$

with probability 1. The optimal values of the parameters r and T are found from the following minimization problem:

$$\text{Minimize } \frac{1}{T}[(m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R]$$
$$\text{subject to } \sum_{k=r+1}^{m+r} \binom{m+r}{k} (1 - e^{-\mu T})^k e^{-\mu T(m+r-k)} \leq \alpha.$$

Using the Lagrange method this problem can be numerically solved.

*Rare events**

In many applied probability problems one has to study rare events. For example, a rare event could be a system failure in reliability applications or buffer overflow in finite-buffer telecommunication problems. Under general conditions it holds that the time until the first occurrence of a rare event is approximately *exponentially* distributed. Loosely formulated, the following result holds. Let $\{X(t)\}$ be a regenerative process having a set B of (bad) states such that the probability q that the process visits the set B during a given cycle is very small. Denote by the random variable U the time until the process visits the set B for the first time. Assuming that the cycle length has a finite and positive mean $E(T)$, it holds that $P\{U > t\} \approx e^{-tq/E(T)}$ for $t \geq 0$; see Keilson (1979) or Solovyez (1971) for

*This section may be skipped at first reading.

a proof. The result that the time until the first occurrence of a rare event in a regenerative process is approximately exponentially distributed is very useful. It gives not only quantitative insight, but it also implies that the computation of the mean of the first-passage time suffices to get the whole distribution.

In the next example we obtain the above result by elementary arguments.

Example 2.2.4 A reliability problem with periodic inspections

High reliability of an electronic system is often achieved by employing redundant components and having periodic inspections. Let us consider a reliability system with two identical units, where one unit is in full operation and the other unit is in warm standby. The operating unit has a constant failure rate of λ_0 and the unit in standby has a constant failure rate of λ_1, where $0 \leq \lambda_1 < \lambda_0$. Upon failure of the operating unit, the standby unit is put into full operation provided the standby is not in the failure state. Failed units are replaced only at the scheduled times $T, 2T, \ldots$ when the system is inspected. The time to replace any failed unit is negligible. A system failure occurs if both units are down. It is assumed that $(\lambda_0 + \lambda_1)T$ is sufficiently small so that a system failure is a rare event. In designing highly reliable systems a key measure of system performance is the probability distribution of the time until the first system failure.

To find the distribution of the time until the first system failure, we first compute the probability q defined by

$$q = P\{\text{system failure occurs between two inspections}\}.$$

To do so, observe that a constant failure rate λ for the lifetime of a unit implies that the lifetime has an exponential distribution with mean $1/\lambda$. Using the fact that the minimum of two independent exponentials with respective means $1/\lambda_0$ and $1/\lambda_1$ is exponentially distributed with mean $1/(\lambda_0+\lambda_1)$, we find by conditioning on the epoch of the first failure of a unit that

$$\begin{aligned} q &= \int_0^T \left\{1 - e^{-\lambda_0(T-x)}\right\} (\lambda_0 + \lambda_1)\, e^{-(\lambda_0+\lambda_1)x}\, dx \\ &= 1 - \frac{(\lambda_0 + \lambda_1)}{\lambda_1} e^{-\lambda_0 T} + \frac{\lambda_0}{\lambda_1} e^{-(\lambda_0+\lambda_1)T}. \end{aligned}$$

Assuming that both units are in good condition at epoch 0, let

$$U = \text{the time until the first system failure.}$$

Since the process describing the state of the two units regenerates itself at each inspection, it follows that

$$P\{U > nT\} = (1 - q)^n, \quad n = 0, 1, \ldots .$$

Assuming that the failure probability q is close to 0, the approximations $(1-q)^n \approx 1 - nq$ and $e^{-nq} \approx 1 - nq$ apply. Thus we find that

$$P\{U > t\} \approx e^{-tq/T}, \quad t \geq 0.$$

In other words, the time until the first system failure is approximately exponentially distributed.

2.3 THE FORMULA OF LITTLE

To introduce the formula of Little, we consider first two illustrative examples. In the first example a hospital admits on average 25 new patients per day. A patient stays on average 3 days in the hospital. What is the average number of occupied beds? Let $\lambda = 25$ denote the average number of new patients who are admitted per day, $W = 3$ the average number of days a patient stays in the hospital and L the average number of occupied beds. Then $L = \lambda W = 25 \times 3 = 75$ beds. In the second example a specialist shop sells on average 100 bottles of a famous Mexican premium beer per week. The shop has on average 250 bottles in inventory. What is the average number of weeks that a bottle is kept in inventory? Let $\lambda = 100$ denote the average demand per week, $L = 250$ the average number of bottles kept in stock and W the average number of weeks that a bottle is kept in stock. Then the answer is $W = L/\lambda = 250/100 = 2.5$ weeks. These examples illustrate Little's formula $L = \lambda W$. The formula of Little is a 'law of nature' that applies to almost any type of queueing system. It relates long-run averages such as the long-run average number of customers in a queue (system) and the long-run average amount of time spent per customer in the queue (system). A queueing system is described by the arrival process of customers, the service facility and the service discipline, to name the most important elements. In formulating the law of Little, there is no need to specify those basic elements. For didactical reasons, however, it is convenient to distinguish between queueing systems with infinite queue capacity and queueing systems with finite queue capacity.

Infinite-capacity queues

Consider a queueing system with infinite queue capacity, that is, every arriving customer is allowed to wait until service can be provided. Define the following random variables:

$L_q(t)$ = the number of customers in the queue at time t (excluding those in service),

$L(t)$ = the number of customers in the system at time t (including those in service),

D_n = the amount of time spent by the nth customer in the queue (excluding service time),

U_n = the amount of time spent by the nth customer in the system (including service time).

Let us assume that each of the stochastic processes $\{L_q(t)\}$, $\{L(t)\}$, $\{D_n\}$ and $\{U_n\}$ is regenerative and has a cycle length with a finite expectation. Then there are constants L_q, L, W_q and W such that the following limits exist and are equal to the respective constants with probability 1:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t L_q(u)\,du = L_q \quad \text{(the long-run average number in queue)},$$

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t L(u)\,du = L \quad \text{(the long-run average number in system)},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} D_k = W_q \quad \text{(the long-run average delay in queue per customer)},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} U_k = W \quad \text{(the long-run average sojourn time per customer)}.$$

Now define the random variable

$$A(t) = \text{the number of customers arrived by time } t,$$

It is also assumed that, for some constant λ,

$$\lim_{t\to\infty} \frac{A(t)}{t} = \lambda \quad \text{with probability 1.}$$

The constant λ gives the long-run average arrival rate of customers. The limit λ exists when customers arrive according to a renewal process (or batches of customers arrive according to a renewal process with independent and identically distributed batch sizes).

The existence of the above limits is sufficient to prove the basic relations

$$L_q = \lambda W_q \tag{2.3.1}$$

and

$$L = \lambda W \tag{2.3.2}$$

These basic relations are the most familiar form of the formula of Little. The reader is referred to Stidham (1974) and Wolff (1989) for a rigorous proof of the formula of Little. Here we will be content to demonstrate the plausibility of this result. The

formula of Little is easiest understood (and reconstructed) when imagining that each customer pays money to the system manager according to some non-discrimination rule. Then it is intuitively obvious that

$$\begin{aligned}&\text{the long-run average reward per time unit earned by the system}\\&\quad=(\text{the long-run average arrival rate of paying customers})\\&\qquad\times(\text{the long-run average amount received per paying customer}).\end{aligned}\tag{2.3.3}$$

In regenerative queueing processes this relation can often be directly proved by using the renewal-reward theorem; see Exercise 2.26. Taking the 'money principle' (2.3.3) as starting point, it is easy to reproduce various representations of Little's law. To obtain (2.3.1), imagine that each customer pays \$1 per time unit while waiting in queue. Then the long-run average amount received per customer equals the long-run average time in queue per customer ($= W_q$). On the other hand, the system manager receives $\$j$ for each time unit that there are j customers waiting in queue. Hence the long-run average reward earned per time unit by the system manager equals the long-run average number of customers waiting in queue ($= L_q$). The average arrival rate of paying customers is obviously given by λ. Applying the relation (2.3.3) gives next the formula (2.3.1). The formula (2.3.2) can be seen by a very similar reasoning: imagine that each customer pays \$1 per time unit while in the system. Another interesting relation arises by imagining that each customer pays \$1 per time unit while in service. Denoting by $E(S)$ the long-run average service time per customer, it follows that

$$\text{the long-run average number of customers in service} = \lambda E(S).\tag{2.3.4}$$

If each customer requires only one server and each server can handle only one customer at a time, this relation leads to

$$\text{the long-run average number of busy servers} = \lambda E(S).\tag{2.3.5}$$

Finite-capacity queues

Assume now there is a maximum on the number of customers allowed in the system. In other words, there are only a finite number of waiting places and each arriving customer finding all waiting places occupied is turned away. It is assumed that a rejected customer has no further influence on the system. Let the rejection probability P_{rej} be defined by

$$P_{rej} = \text{the long-run fraction of customers who are turned away,}$$

assuming that this long-run fraction is well defined. The random variables $L(t)$, $L_q(t)$, D_n and U_n are defined as before, except that D_n and U_n now refer to the queueing time and sojourn time of the nth *accepted* customer. The constants W_q and W now represent the long-run average queueing time per *accepted* customer

and the long-run average sojourn time per *accepted* customer. The formulas (2.3.1), (2.3.2) and (2.3.4) need only slight modification:

$$L_q = \lambda(1 - P_{rej})W_q \quad \text{and} \quad L = \lambda(1 - P_{rej})W, \tag{2.3.6}$$

the long-run average number of customers in service

$$= \lambda(1 - P_{rej})E(S). \tag{2.3.7}$$

Heuristically, these formulas follow by applying the money principle (2.3.3) and taking only the accepted customers as paying customers.

2.4 POISSON ARRIVALS SEE TIME AVERAGES

In the analysis of queueing (and other) problems, one sometimes needs the long-run fraction of time the system is in a given state and sometimes needs the long-run fraction of arrivals who find the system in a given state. These averages can often be related to each other, but in general they are not equal to each other. To illustrate that the two averages are in general not equal to each other, suppose that customers arrive at a service facility according to a deterministic process in which the interarrival times are 1 minute. If the service of each customer is uniformly distributed between $\frac{1}{4}$ minute and $\frac{3}{4}$ minute, then the long-run fraction of time the system is empty equals $\frac{1}{2}$, whereas the long-run fraction of arrivals finding the system empty equals 1. However the two averages would have been the same if the arrival process of customers had been a Poisson process. As a prelude to the generally valid property that Poisson arrivals see time averages, we first analyse two specific problems by the renewal-reward theorem.

Example 2.4.1 A manufacturing problem

Suppose that jobs arrive at a workstation according to a Poisson process with rate λ. The workstation has no buffer to store temporarily arriving jobs. An arriving job is accepted only when the workstation is idle, and is lost otherwise. The processing times of the jobs are independent random variables having a common probability distribution with finite mean β. What is the long-run fraction of time the workstation is busy and what is the long-run fraction of jobs that are lost?

These two questions are easily answered by using the renewal-reward theorem. Let us define the following random variables. For any $t \geq 0$, let

$$I(t) = \begin{cases} 1 & \text{if the workstation is busy at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Also, for any $n = 1, 2, \ldots$, let

$$I_n = \begin{cases} 1 & \text{if the workstation is busy just prior to the } n\text{th arrival,} \\ 0 & \text{otherwise.} \end{cases}$$

The continuous-time process $\{I(t)\}$ and the discrete-time process $\{I_n\}$ are both regenerative. The arrival epochs occurring when the workstation is idle are regeneration epochs for the two processes. Why? Let us say that a cycle starts each time an arriving job finds the workstation idle. The long-run fraction of time the workstation is busy is equal to the expected amount of time the workstation is busy during one cycle divided by the expected length of one cycle. The expected length of the busy period in one cycle equals β. Since the Poisson arrival process is memoryless, the expected length of the idle period during one cycle equals the mean interarrival time $1/\lambda$. Hence, with probability 1,

$$\text{the long-run fraction of time the workstation is busy} = \frac{\beta}{\beta + 1/\lambda}. \tag{2.4.1}$$

The long-run fraction of jobs that are lost equals the expected number of jobs lost during one cycle divided by the expected number of jobs arriving during one cycle. Since the arrival process is a Poisson process, the expected number of (lost) arrivals during the busy period in one cycle equals $\lambda \times E(\text{processing time of a job}) = \lambda\beta$. Hence, with probability 1,

$$\text{the long-run fraction of jobs that are lost} = \frac{\lambda\beta}{1 + \lambda\beta}. \tag{2.4.2}$$

Thus, we obtain from (2.4.1) and (2.4.2) the remarkable result

$$\text{the long-run fraction of arrivals finding the workstation busy} = \text{the long-run fraction of time the workstation is busy.} \tag{2.4.3}$$

Incidentally, it is interesting to note that in this loss system the long-run fraction of lost jobs is insensitive to the form of the distribution function of the processing time but needs only the first moment of this distribution. This simple loss system is a special case of Erlang's loss model to be discussed in Chapter 5.

Example 2.4.2 An inventory model

Consider a single-product inventory system in which customers asking for the product arrive according to a Poisson process with rate λ. Each customer asks for one unit of the product. Each demand which cannot be satisfied directly from stock on hand is lost. Opportunities to replenish the inventory occur according to a Poisson process with rate μ. This process is assumed to be independent of the demand process. For technical reasons a replenishment can only be made when the inventory is zero. The inventory on hand is raised to the level Q each time a replenishment is done. What is the long-run fraction of time the system is out of stock? What is the long-run fraction of demand that is lost?

In the same way as in Example 2.4.1, we define the random variables

$$I(t) = \begin{cases} 1 & \text{if the system is out of stock at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

and

$$I_n = \begin{cases} 1 & \text{if the system is out of stock when the } n\text{th demand occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

The continuous-time process $\{I(t)\}$ and the discrete-time process $\{I_n\}$ are both regenerative. The regeneration epochs are the demand epochs at which the stock on hand drops to zero. Why? Let us say that a cycle starts each time the stock on hand drops to zero. The system is out of stock during the time elapsed from the beginning of a cycle until the next inventory replenishment. This amount of time is exponentially distributed with mean $1/\mu$. The expected amount of time it takes to go from stock level Q to 0 equals Q/λ. Hence, with probability 1,

$$\text{the long-run fraction of time the system is out of stock} = \frac{1/\mu}{1/\mu + Q/\lambda}. \tag{2.4.4}$$

To find the fraction of demand that is lost, note that the expected amount of demand lost in one cycle equals $\lambda \times E(\text{amount of time the system is out of stock during one cycle}) = \lambda/\mu$. Hence, with probability 1,

$$\text{the long-run fraction of demand that is lost} = \frac{\lambda/\mu}{\lambda/\mu + Q}. \tag{2.4.5}$$

Together (2.4.4) and (2.4.5) lead to this remarkable result:

$$\text{the long-run fraction of customers finding the system out of stock} = \text{the long-run fraction of time the system is out of stock.} \tag{2.4.6}$$

The relations (2.4.3) and (2.4.6) are particular instances of the property 'Poisson arrivals see time averages'. Roughly stated, this property expresses that in statistical equilibrium the distribution of the state of the system *just prior* to an arrival epoch is the same as the distribution of the state of the system at an *arbitrary* epoch when arrivals occur according to a Poisson process. An intuitive explanation of the property 'Poisson arrivals see time averages' is that Poisson arrivals occur completely randomly in time; cf. Theorem 1.1.5.

Next we discuss the property of 'Poisson arrivals see time averages' in a broader context. For ease of presentation we use the terminology of Poisson arrivals. However, the results below also apply to Poisson processes in other contexts. For some

specific problem let the continuous-time stochastic process $\{X(t), t \geq 0\}$ describe the evolution of the state of a system and let $\{N(t), t \geq 0\}$ be a renewal process describing arrivals to that system. As examples:

(a) $X(t)$ is the number of customers present at time t in a queueing system.

(b) $X(t)$ describes jointly the inventory level and the prevailing production rate at time t in a production/inventory problem with a variable production rate.

It is assumed that the arrival process $\{N(t), t \geq 0\}$ can be seen as an exogenous factor to the system and is not affected by the system itself. More precisely, the following assumption is made.

Lack of anticipation assumption *For each $u \geq 0$ the future arrivals occurring after time u are independent of the history of the process $\{X(t)\}$ up to time u.*

It is not necessary to specify how the arrival process $\{N(t)\}$ precisely interacts with the state process $\{X(t)\}$. Denoting by τ_n the nth arrival epoch, let the random variable X_n be defined by $X(\tau_n^-)$. In other words,

$$X_n = \text{the state of the system just prior to the } n\text{th arrival epoch.}$$

Let B be any set of states for the $\{X(t)\}$ process. For each $t \geq 0$, define the indicator variable

$$I_B(t) = \begin{cases} 1 & \text{if } X(t) \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Also, for each $n = 1, 2, \ldots$, define the indicator variable $I_n(B)$ by

$$I_n(B) = \begin{cases} 1 & \text{if } X_n \in B, \\ 0 & \text{otherwise.} \end{cases}$$

The technical assumption is made that the sample paths of the continuous-time process $\{I_B(t), t \geq 0\}$ are right-continuous and have left-hand limits. In practical situations this assumption is always satisfied.

Theorem 2.4.1 (Poisson arrivals see time averages) *Suppose that the arrival process $\{N(t)\}$ is a Poisson process with rate λ. Then:*

(a) For any $t > 0$,

$$E[\textit{number of arrivals in } (0, t) \textit{ finding the system in the set } B] = \lambda E\left[\int_0^t I_B(u)\, du\right].$$

(b) With probability 1*, the long-run fraction of arrivals who find the system in the set B of states equals the long-run fraction of time the system is in the set B of states. That is, with probability* 1*,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k(B) = \lim_{t\to\infty} \frac{1}{t} \int_0^t I_B(u)\, du.$$

Proof See Wolff (1982).

It is remarkable in Theorem 2.4.1 that E[number of arrivals in $(0, t)$ finding the system in the set B] is equal to $\lambda \times E$[amount of time in $(0, t)$ that the system is in the set B], although there is dependency between the arrivals in $(0, t)$ and the evolution of the state of the system during $(0, t)$. This result is characteristic for the Poisson process.

The property 'Poisson arrivals see time averages' is usually abbreviated as ***PASTA***. Theorem 2.4.1 has a useful corollary when it is assumed that the continuous-time process $\{X(t)\}$ is a regenerative process whose cycle length has a finite positive mean. Define the random variables T_B and N_B by

T_B = amount of time the process $\{X(t)\}$ is in the set B of states during one cycle,

N_B = number of arrivals during one cycle who find the process $\{X(t)\}$ in the set of B states.

The following corollary will be very useful in the algorithmic analysis of queueing systems in Chapter 9.

Corollary 2.4.2 *If the arrival process* $\{N(t)\}$ *is a Poisson process with rate* λ*, then*

$$E(N_B) = \lambda E(T_B).$$

Proof Denote by the random variables T and N the length of one cycle and the number of arrivals during one cycle. Then, by Theorem 2.2.3,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_B(u)\, du = \frac{E(T_B)}{E(T)} \quad \text{with probability 1}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k(B) = \frac{E(N_B)}{E(N)} \quad \text{with probability 1.}$$

It now follows from part (b) of Theorem 2.4.1 that $E(N_B)/E(N) = E(T_B)/E(T)$. Thus the corollary follows if we can verify that $E(N)/E(T) = \lambda$. To do so, note that the regeneration epochs for the process $\{X(t)\}$ are also regeneration epochs for the Poisson arrival process. Thus, by the renewal-reward theorem, the long-run average number of arrivals per time unit equals $E(N)/E(T)$, showing that $E(N)/E(T) = \lambda$.

To conclude this section, we use the PASTA property to derive in a heuristic way one of the most famous formulas from queueing theory.

2.5 THE POLLACZEK–KHINTCHINE FORMULA

Suppose customers arrive at a service facility according to a Poisson process with rate λ. The service times of the customers are independent random variables having a common probability distribution with finite first two moments $E(S)$ and $E(S^2)$. There is a single server and ample waiting room for arriving customers finding the server busy. Each customer waits until service is provided. The server can handle only one customer at a time. This particular queueing model is abbreviated as the $M/G/1$ queue; see Kendall's notation in Section 9.1. The offered load ρ is defined by

$$\rho = \lambda E(S)$$

and it is assumed that $\rho < 1$. By Little's formula (2.3.5) the load factor ρ can be interpreted as the long-run fraction of time the server is busy. Important performance measures are

L_q = the long-run average number of customers waiting in queue,

W_q = the long-run average time spent per customer in queue.

The Pollaczek–Khintchine formula states that

$$W_q = \frac{\lambda E(S^2)}{2(1-\rho)}. \tag{2.5.1}$$

This formula also implies an explicit expression for L_q by Little's formula

$$L_q = \lambda W_q; \tag{2.5.2}$$

see Section 2.3. The Pollaczek–Khintchine formula gives not only an explicit expression for W_q, but more importantly it gives useful qualitative insights as well. It shows that the average delay per customer in the $M/G/1$ queue uses the service-time distribution only through its first two moments. Denoting by $c_S^2 = \sigma^2(S)/E^2(S)$ the squared coefficient of variation of the service time and using the relation (2.1.10), we can write the Pollaczek–Khintchine formula in the more insightful form

$$W_q = \frac{1}{2}(1 + c_S^2)\frac{\rho E(S)}{1-\rho}. \tag{2.5.3}$$

Hence the Pollaczek–Khintchine formula shows that the average delay per customer decreases according to the factor $\frac{1}{2}(1 + c_S^2)$ when the variability in the service is reduced while the average arrival rate and the mean service time are kept

fixed. Noting that $c_S^2 = 1$ for exponentially distributed service times, the expression (2.5.3) can also be written as

$$W_q = \frac{1}{2}(1 + c_S^2) W_q(\exp), \tag{2.5.4}$$

where $W_q(\exp) = \rho E(S)/(1-\rho)$ denotes the average delay per customer for the case of exponential services. In particular, writing $W_q = W_q(\det)$ for deterministic services ($c_S^2 = 0$), we have

$$W_q(\det) = \frac{1}{2} W_q(\exp). \tag{2.5.5}$$

It will be seen in Chapter 9 that the structural form (2.5.4) is very useful to design approximations in more complex queueing models.

Another important feature shown by the Pollaczek–Khintchine formula is that the average delay and average queue size increase in a *non-linear* way when the offered load ρ increases. A twice as large value for the offered load does not imply a twice as large value for the average delay! On the contrary, the average delay and the average queue size explode when the average arrival rate becomes very close to the average service rate. Differentiation of W_q as a function of ρ shows that the slope of increase of W_q as a function of ρ is proportional to $(1-\rho)^{-2}$. As an illustration a small increase in the average arrival rate when the load $\rho = 0.9$ causes an increase in the average delay 25 times greater than it would cause when the load $\rho = 0.5$. This non-intuitive finding demonstrates the danger of designing a stochastic system with too high a utilization level, since then a small increase in the offered load will in general cause a dramatic degradation in system performance.

We have not yet proved the Pollaczek–Khintchine formula. First we give a heuristic derivation and next we give a rigorous proof.

Heuristic derivation

Tag a customer who arrives when the system has reached statistical equilibrium. Denote its waiting time in queue by the random variable D_{tag}. Heuristically, $E(D_{tag}) = W_q$. By the PASTA property, the expected number of customers in queue seen upon arrival by the tagged customer equals L_q. Noting that ρ is the long-run fraction of time the server is busy, it also follows that the tagged customer finds the server busy upon arrival with probability ρ. Using the result (2.1.8) for the excess variable, it is plausible that the expected remaining service time of the customer seen in service by a Poisson arrival equals $\frac{1}{2}E(S^2)/E(S)$. Putting the pieces together, we find the relation

$$E(D_{tag}) = L_q E(S) + \rho \frac{E(S^2)}{2E(S)}.$$

Substituting $E(D_{tag}) = W_q$ and $L_q = \lambda W_q$, the relation becomes

$$W_q = \lambda E(S) W_q + \frac{\rho E(S^2)}{2E(S)}$$

yielding the Pollaczek–Khintchine formula for W_q.

Rigorous derivation

A rigorous derivation of the Pollaczek–Khintchine formula can be given by using the powerful generating-function approach. Define first the random variables

$L(t) =$ the number of customers present at time t,

$Q_n =$ the number of customers present just after the nth service completion epoch,

$L_n =$ the number of customers present just before the nth arrival epoch.

The processes $\{L(t)\}$, $\{Q_n\}$ and $\{L_n\}$ are regenerative stochastic processes with finite expected cycle lengths. Denote the corresponding limiting distributions by

$$p_j = \lim_{t\to\infty} P\{L(t) = j\},\ q_j = \lim_{n\to\infty} P\{Q_n = j\} \quad \text{and} \quad \pi_j = \lim_{n\to\infty} P\{L_n = j\}$$

for $j = 0, 1, \ldots$. The existence of the limiting distributions can be deduced from Theorem 2.2.4 (the amount of time elapsed between two arrivals that find the system empty has a probability density and the number of customers served during this time has an aperiodic distribution). We omit the details. The limiting probabilities can also be interpreted as long-run averages. For example, q_j is the long-run fraction of customers leaving j other customers behind upon service completion and π_j is the long-run fraction of customers finding j other customers present upon arrival. The following important identity holds:

$$\pi_j = p_j = q_j, \quad j = 0, 1, \ldots . \tag{2.5.6}$$

Since the arrival process is a Poisson process, the equality $\pi_j = p_j$ is readily verified from Theorem 2.4.1. To verify the equality $\pi_j = q_j$, define the random variable $L_n^{(j)}$ as the number of customers over the first n arrivals who see j other customers present upon arrival and define the random variable $Q_n^{(j)}$ as the number of service completion epochs over the first n service completions at which j customers are left behind. Customers arrive singly and are served singly. Thus between any two arrivals that find j other customers present there must be a service completion at which j customers are left behind and, conversely, between any two service completions at which j customers are left behind there must be an arrival that sees j other customers present. By this up- and downcrossing argument, we have for

each j that

$$\left| L_n^{(j)} - Q_n^{(j)} \right| \leq 1, \quad n = 1, 2, \ldots .$$

Consequently, $\pi_j = \lim_{n\to\infty} L_n^{(j)}/n = \lim_{n\to\infty} Q_n^{(j)}/n = q_j$ for all j. We are now ready to prove that

$$\lim_{n\to\infty} E(z^{Q_n}) = \frac{(1-z)q_0 A(z)}{A(z) - z}, \tag{2.5.7}$$

where

$$A(z) = \int_0^\infty e^{-\lambda t(1-z)} b(t)\, dt$$

with $b(t)$ denoting the probability density of the service time of a customer. Before proving this result, we note that the unknown q_0 is determined by the fact that the left-hand side of (2.5.7) equals 1 for $z = 1$. By applying L'Hospital's rule, we find $q_0 = 1 - \rho$, in agreement with Little's formula $1 - p_0 = \rho$. By the bounded convergence theorem in Appendix A,

$$\lim_{n\to\infty} E(z^{Q_n}) = \lim_{n\to\infty} \sum_{j=0}^{\infty} P\{Q_n = j\} z^j = \sum_{j=0}^{\infty} q_j z^j, \quad |z| \leq 1.$$

Hence, by (2.5.6) and (2.5.7),

$$\sum_{j=0}^{\infty} p_j z^j = \frac{(1-\rho)(1-z)A(z)}{A(z) - z}. \tag{2.5.8}$$

Since the long-run average queue size L_q is given by

$$L_q = \sum_{j=1}^{\infty} (j-1) p_j = \sum_{j=0}^{\infty} j p_j - (1 - p_0)$$

(see Exercise 2.28), the Pollaczek–Khintchine formula for L_q follows by differentiating the right-hand side of (2.5.8) and taking $z = 1$ in the derivative. It remains to prove (2.5.7). To do so, note that

$$Q_n = Q_{n-1} - \delta(Q_{n-1}) + A_n, \quad n = 1, 2, \ldots ,$$

where $\delta(x) = 1$ for $x > 0$, $\delta(x) = 0$ for $x = 0$ and A_n is the number of customers arriving during the nth service time. By the law of total probability,

$$P\{A_n = k\} = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} b(t)\, dt, \quad k = 0, 1, \ldots$$

and so

$$\sum_{k=0}^{\infty} P\{A_n = k\} z^k = \int_0^{\infty} e^{-\lambda t(1-z)} b(t)\, dt.$$

Since the random variables $Q_{n-1} - \delta(Q_{n-1})$ and A_n are independent of each other,

$$E(z^{Q_n}) = E(z^{Q_{n-1}-\delta(Q_{n-1})}) E(z^{A_n}). \tag{2.5.9}$$

We have

$$\begin{aligned} E(z^{Q_{n-1}-\delta(Q_{n-1})}) &= P\{Q_{n-1} = 0\} + \sum_{j=1}^{\infty} z^{j-1} P\{Q_{n-1} = j\} \\ &= P\{Q_{n-1} = 0\} + \frac{1}{z}[E(z^{Q_{n-1}}) - P\{Q_{n-1} = 0\}]. \end{aligned}$$

Substituting this in (2.5.9), we find

$$zE(z^{Q_n}) = \Big[E(z^{Q_{n-1}}) - (1-z)P\{Q_{n-1} = 0\}\Big] A(z).$$

Letting $n \to \infty$, we next obtain the desired result (2.5.7). This completes the proof.

Before concluding this section, we give an amusing application of the Pollaczek–Khintchine formula.

Example 2.5.1 Ladies in waiting*

Everybody knows women spend on average more time in the loo than men. As worldwide studies show, women typically take 89 seconds to use the loo—about twice as long as the 39 seconds required by the average man. However, this does not mean that the queue for the women's loo is twice as long as the queue for the men's. The sequence for the women's loo is usually far longer. To explain this using the Pollaczek–Khintchine formula, let us make the following reasonable assumptions:

1. Men and women arrive at the loo according to independent Poisson processes with the same rates.
2. The expected amount of time people spend in the loo is twice as large for women as for men.
3. The coefficient of variation of the time people spend in the loo is larger for women than for men.
4. There is one loo for women only and one loo for men only.

*This application is based on the article 'Ladies Waiting' by Robert Matthews in *New Scientist*, Vol. 167, Issue 2249, 29 July 2000.

Let λ_w and λ_m denote the average arrival rates of women and men. Let μ_w and c_w denote the mean and the coefficient of variation of the amount of time a woman spends in the loo. Similarly, μ_m and c_m are defined for men. It is assumed that $\lambda_w \mu_w < 1$. Using the assumptions $\lambda_w = \lambda_m$, $\mu_w = 2\mu_m$ and $c_w \geq c_m$, it follows from (2.5.2) and the Pollaczek–Khintchine formula (2.5.3) that

$$\text{the average queue size for the women's loo}$$
$$= \frac{1}{2}(1 + c_w^2)\frac{(\lambda_w \mu_w)^2}{1 - \lambda_w \mu_w} \geq \frac{1}{2}(1 + c_m^2)\frac{(2\lambda_m \mu_m)^2}{1 - 2\lambda_m \mu_m}$$
$$\geq 4 \times \frac{1}{2}(1 + c_m^2)\frac{(\lambda_m \mu_m)^2}{1 - \lambda_m \mu_m}.$$

Hence

$$\text{the average queue size for the women's loo}$$
$$\geq 4 \times (\text{the average queue size for the men's loo}).$$

The above derivation uses the estimate $1 - 2\lambda_m \mu_m \leq 1 - \lambda_m \mu_m$ and thus shows that the relative difference actually increases much faster than a factor 4 when the utilization factor $\lambda_w \mu_w$ becomes closer to 1.

*Laplace transform of the waiting-time probabilities**

The generating-function method enabled us to prove the Pollaczek–Khintchine formula for the average queue size. Using Little's formula we next found the Pollaczek–Khintchine formula for the average delay in queue of a customer. The latter formula can also be directly obtained from the Laplace transform of the waiting-time distribution. This Laplace transform is also of great importance in itself. The waiting-time probabilities can be calculated by numerical inversion of the Laplace transform; see Appendix F. A simple derivation can be given for the Laplace transform of the waiting-time distribution in the $M/G/1$ queue when service is in order of arrival. The derivation parallels the derivation of the generating function of the number of customers in the system.

Denote by D_n the delay in queue of the nth arriving customer and let the random variables S_n and τ_n denote the service time of the nth customer and the time elapsed between the arrivals of the nth customer and the $(n+1)$th customer. Since $D_{n+1} = 0$ if $D_n + S_n < \tau_n$ and $D_{n+1} = D_n + S_n - \tau_n$ otherwise, we have

$$D_{n+1} = (D_n + S_n - \tau_n)^+, \quad n = 1, 2, \ldots, \tag{2.5.10}$$

where x^+ is the usual notation for $x = \max(x, 0)$. From the recurrence formula (2.5.10), we can derive that for all s with $\mathrm{Re}(s) \geq 0$ and $n = 1, 2, \ldots$

$$(\lambda - s)E\left(e^{-sD_{n+1}}\right) = \lambda E\left(e^{-sD_n}\right) b^*(s) - sP\{D_{n+1} = 0\}, \tag{2.5.11}$$

*This section can be skipped at first reading.

where $b^*(s) = \int_0^\infty e^{-sx} b(x)\,dx$ denotes the Laplace transform of the probability density $b(x)$ of the service time. To prove this, note that D_n, S_n and τ_n are independent of each other. This implies that, for any $x > 0$,

$$\begin{aligned} &E\left[e^{-s(D_n+S_n-\tau_n)^+} \mid D_n + S_n = x\right] \\ &\quad = \int_0^x e^{-s(x-y)} \lambda e^{-\lambda y}\,dy + \int_x^\infty e^{-s\times 0} \lambda e^{-\lambda y}\,dy \\ &\quad = \frac{\lambda}{\lambda - s}(e^{-sx} - e^{-\lambda x}) + e^{-\lambda x} = \frac{1}{\lambda - s}(\lambda e^{-sx} - s e^{-\lambda x}) \end{aligned}$$

for $s \neq \lambda$ (using L'Hospital's rule it can be seen that this relation also holds for $s = \lambda$). Hence, using (2.5.10),

$$(\lambda - s) E\left(e^{-sD_{n+1}}\right) = \lambda E\left[e^{-s(D_n+S_n)}\right] - s E\left[e^{-\lambda(D_n+S_n)}\right].$$

Since $P\{(D_n + S_n - \tau_n)^+ = 0 \mid D_n + S_n = x\} = e^{-\lambda x}$, we also have

$$P\{D_{n+1} = 0\} = E\left[e^{-\lambda(D_n+S_n)}\right].$$

The latter two relations and $E\left[e^{-s(D_n+S_n)}\right] = E\left(e^{-sD_n}\right) E\left(e^{-sS_n}\right)$ lead to (2.5.11). The steady-state waiting-time distribution function $W_q(x)$ is defined by

$$W_q(x) = \lim_{n\to\infty} P\{D_n \leq x\}, \quad x \geq 0.$$

The existence of this limit can be proved from Theorem 2.2.4. Let the random variable D_∞ have $W_q(x)$ as probability distribution function. Then, by the bounded convergence theorem in Appendix A, $E(e^{-sD_\infty}) = \lim_{n\to\infty} E(e^{-sD_n})$. Using (2.5.6), it follows from $\lim_{n\to\infty} P\{D_{n+1} = 0\} = \pi_0$ and $q_0 = 1 - \rho$ that $\lim_{n\to\infty} P\{D_{n+1} = 0\} = 1 - \rho$. Letting $n \to \infty$ in (2.5.11), we find that

$$E\left(e^{-sD_\infty}\right) = \frac{(1-\rho)s}{s - \lambda + \lambda b^*(s)}. \tag{2.5.12}$$

Noting that $P\{D_\infty \leq x\} = W_q(x)$ and using relation (E.7) in Appendix E, we get from (2.5.12) the desired result:

$$\int_0^\infty e^{-sx} \left\{1 - W_q(x)\right\} dx = \frac{\rho s - \lambda + \lambda b^*(s)}{s(s - \lambda + \lambda b^*(s))}. \tag{2.5.13}$$

Taking the derivative of the right-hand side of (2.5.13) and putting $s = 0$, we obtain

$$\int_0^\infty \left\{1 - W_q(x)\right\} dx = \frac{\lambda E(S^2)}{2(1-\rho)},$$

in agreement with the Pollaczek–Khintchine formula (2.5.1).

Remark 2.5.1 Relation between queue size and waiting time

Let the random variable $L_q^{(\infty)}$ be distributed according to the limiting distribution of the number of customers in queue at an arbitrary point in time. That is, $P\{L_q^{(\infty)} = j\} = p_{j+1}$ for $j \geq 1$ and $P\{L_q^{(\infty)} = 0\} = p_0 + p_1$. Then the generating function of $L_q^{(\infty)}$ and the Laplace transform of the delay distribution are related to each other by

$$E(z^{L_q^{(\infty)}}) = E[e^{-\lambda(1-z)D_\infty}], \quad |z| \leq 1. \tag{2.5.14}$$

A direct probabilistic proof of this important relation can be given. Denote by L_n the number of customers left behind in queue when the nth customer enters service. Since service is in order of arrival, L_n is given by the number of customers arriving during the delay D_n of the nth customer. Since the generating function of a Poisson distributed variable with mean δ is exp $(-\delta\,(1 - z))$, it follows that for any $x \geq 0$ and $n \geq 1$,

$$E(z^{L_n} | D_n = x) = e^{-\lambda x(1-z)}.$$

Hence

$$E(z^{L_n}) = E[e^{-\lambda(1-z)D_n}], \quad n \geq 1. \tag{2.5.15}$$

The limiting distribution of L_n as $n \to \infty$ is the same as the probability distribution of $L_q^{(\infty)}$. This follows from an up- and downcrossing argument: the long-run fraction of customers leaving j other customers behind in queue when entering service equals the long-run fraction of customers finding j other customers in queue upon arrival. Noting that there is a single server and using the PASTA property, it follows that the latter fraction equals p_{j+1} for $j \geq 1$ and $p_0 + p_1$ for $j = 0$. This proves that the limiting distribution of L_n equals the distribution of $L_q^{(\infty)}$. Note that, by Theorem 2.2.4, L_n has a limiting distribution as $n \to \infty$. Letting $n \to \infty$ in (2.5.15), the result (2.5.14) follows.

Letting $w_q\,(x)$ denote the derivative of the waiting-time distribution function $W_q\,(x)$ for $x > 0$, note that for the $M/G/1$ queue the relation (2.5.14) can be restated as

$$p_{j+1} = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} w_q\,(x)\,dx, \quad j = 1, 2, \ldots\ .$$

The relation (2.5.14) applies to many other queueing systems with Poisson arrivals. The importance of (2.5.14) is that this relation enables us to directly obtain the Laplace transform of the waiting-time distribution function from the generating function of the queue size. To illustrate this, note that $E(z^{L_q^{(\infty)}}) = p_0 + \frac{1}{z}[P\,(z) - p_0]$ for the $M/G/1$ queue, where $P(z) = \sum_{j=0}^\infty p_j z^j$ is given by (2.5.8). Using this relation together with (2.5.8) and noting that $A\,(z) = b^*\,(\lambda\,(1 - z))$, it follows from the basic relation (2.5.14) that $E(e^{-sD_\infty})$ is indeed given by (2.5.12).

2.6 A CONTROLLED QUEUE WITH REMOVABLE SERVER*

Consider a production facility at which production orders arrive according to a Poisson process with rate λ. The production times $\tau_1, \tau_2, \ldots$ of the orders are independent random variables having a common probability distribution function F with finite first two moments. Also, the production process is independent of the arrival process. The facility can only work on one order at a time. It is assumed that $E(\tau_1) < 1/\lambda$; that is, the average production time per order is less than the mean interarrival time between two consecutive orders. The facility operates only intermittently and is shut down when no orders are present any more. A fixed set-up cost of $K > 0$ is incurred each time the facility is reopened. Also a holding cost $h > 0$ per time unit is incurred for each order waiting in queue. The facility is only turned on when enough orders have accumulated. The so-called N-policy reactivates the facility as soon as N orders are present. For ease we assume that it takes a zero set-up time to restart production. How do we choose the value of the control parameter N such that the long-run average cost per time unit is minimal?

To analyse this problem, we first observe that for a given N-policy the stochastic process describing jointly the number of orders present and the status of the facility (on or off) regenerates itself each time the facility is turned on. Define a cycle as the time elapsed between two consecutive reactivations of the facility. Clearly, each cycle consists of a busy period B with production and an idle period I with no production. We deal separately with the idle and the busy periods. Using the memoryless property of the Poisson process, the length of the idle period is the sum of N exponential random variables each having mean $1/\lambda$. Hence

$$E(\text{length of the idle period } I) = \frac{N}{\lambda}.$$

Similarly,

$$E(\text{holding cost incurred during } I) = h\left(\frac{N-1}{\lambda} + \cdots + \frac{1}{\lambda}\right).$$

To deal with the busy period, we define for $n = 1, 2, \ldots$ the quantities

$$t_n = \text{the expected time until the facility becomes empty given that at epoch 0 a production starts with } n \text{ orders present,}$$

and

$$h_n = \text{the expected holding costs incurred until the facility becomes empty given that at epoch 0 a production starts with } n \text{ orders present.}$$

These quantities are independent of the control rule considered. In particular, the expected length of a busy period equals t_N and the expected holding costs incurred

*This section contains specialized material and can be skipped at first reading.

during a busy period equals h_N. By the renewal-reward theorem,

$$\text{the long-run average cost per time unit} = \frac{(h/2\lambda)N(N-1) + K + h_N}{N/\lambda + t_N}$$

with probability 1. To find the functions t_n and h_n, we need

a_j = the probability that j orders arrive during the production time of a single order.

Assume for ease that the production time has a probability density $f(x)$. By conditioning on the production time and noting that the number of orders arriving in a fixed time y is Poisson distributed with mean λy, it follows that

$$a_j = \int_0^\infty e^{-\lambda y} \frac{(\lambda y)^j}{j!} f(y)\, dy, \quad j = 0, 1, \ldots .$$

It is readily verified that

$$\sum_{j=1}^{\infty} j a_j = \lambda E(\tau_1) \quad \text{and} \quad \sum_{j=1}^{\infty} j^2 a_j = \lambda^2 E(\tau_1^2) + \lambda E(\tau_1). \tag{2.6.1}$$

We now derive recursion relations for the quantities t_n and h_n. Suppose that at epoch 0 a production starts with n orders present. If the number of new orders arriving during the production time of the first order is j, then the time to empty the system equals the first production time plus the time to empty the system starting with $n - 1 + j$ orders present. Thus

$$t_n = E(\tau_1) + \sum_{j=0}^{\infty} t_{n-1+j} a_j, \quad n = 1, 2, \ldots ,$$

where $t_0 = 0$. Similarly, we derive a recursion relation for the h_n. To do so, note that relation (1.1.10) implies that the expected holding cost for new orders arriving during the first production time τ_1 equals $\frac{1}{2}h\lambda E(\tau_1^2)$. Hence

$$h_n = (n-1)hE(\tau_1) + \frac{1}{2}h\lambda E(\tau_1^2) + \sum_{j=0}^{\infty} h_{n-1+j} a_j, \quad n = 1, 2, \ldots ,$$

where $h_0 = 0$. In a moment it will be shown that t_n is linear in n and h_n is quadratic in n. Substituting these functional forms in the above recursion relations and using (2.6.1), we find after some algebra that for $n = 1, 2, \ldots ,$

$$t_n = \frac{nE(\tau_1)}{1 - \lambda E(\tau_1)}, \tag{2.6.2}$$

$$h_n = \frac{h}{1-\lambda E(\tau_1)}\left[\frac{1}{2}n(n-1)E(\tau_1) + \frac{\lambda n E(\tau_1^2)}{2\{1-\lambda E(\tau_1)\}}\right]. \tag{2.6.3}$$

To verify that t_n is linear in n and h_n is quadratic in n, a brilliant idea due to Takács (1962) is used. First observe that t_n and h_n do not depend on the specific order in which the production orders are coped with during the production process. Imagine now the following production discipline. The n initial orders $O_1, \ldots, O_n$ are separated. Order O_1 is produced first, after which all orders (if any) are produced that have arrived during the production time of O_1, and this way of production is continued until the facility is free of all orders but $O_2, \ldots, O_n$. Next this procedure is repeated with order O_2, etc. Thus we find that $t_n = nt_1$, proving that t_n is linear in n. The memoryless property of the Poisson process is crucial in this argument. Why? The same separation argument is used to prove that h_n is quadratic in n. Since $h_1 + (n-k) \times ht_1$ gives the expected holding cost incurred during the time to free the system of order O_k and its direct descendants until only the orders $O_{k+1}, \ldots, O_n$ are left, it follows that

$$h_n = \sum_{k=1}^{n}\{h_1 + (n-k)ht_1\} = nh_1 + \frac{1}{2}hn(n-1)t_1.$$

Combining the above results we find for the N-policy that

$$\begin{aligned}&\text{the long-run average cost per time unit} \\ &= \frac{\lambda(1-\rho)K}{N} + h\left\{\frac{\lambda^2 E(\tau_1^2)}{2(1-\rho)} + \frac{N-1}{2}\right\},\end{aligned} \tag{2.6.4}$$

where $\rho = \lambda E(\tau_1)$. It is worth noting here that this expression needs only the first two moments from the production time. Also note that, by putting $K = 0$ and $h = 1$ in (2.6.4),

$$\begin{aligned}&\text{the long-run average number of orders waiting in queue} \\ &= \frac{\lambda^2 E(\tau_1^2)}{2(1-\rho)} + \frac{N-1}{2}.\end{aligned}$$

For the special case of $N = 1$ this formula reduces to the famous Pollaczek–Khintchine formula for the average queue length in the standard $M/G/1$ queue; see Section 2.5.

The optimal value of N can be obtained by differentiating the right-hand side of (2.6.4), in which we take N as a continuous variable. Since the average cost is convex in N, it follows that the average cost is minimal for one of the two integers nearest to

$$N^* = \sqrt{\frac{2\lambda(1-\rho)K}{h}}.$$

2.7 AN UP- AND DOWNCROSSING TECHNIQUE

In this section we discuss a generally applicable up- and downcrossing technique that, in conjunction with the PASTA property, can be used to establish relations between customer-average and time-average probabilities in queueing systems. To illustrate this, we consider the so-called $GI/M/1$ queue. In this single-server system, customers arrive according to a renewal process and the service times of the customers have a common exponential distribution. The single server can handle only one customer at a time and there is ample waiting room for customers who find the server busy upon arrival. The service times of the customers are independent of each other and are also independent of the arrival process. Denoting by λ the average arrival rate ($1/\lambda$ = the mean interarrival time) and by β the service rate ($1/\beta$ = the mean service time), it is assumed that $\lambda < \beta$.

The continuous-time stochastic process $\{X(t),\ t \geq 0\}$ and the discrete-time stochastic process $\{X_n,\ n = 1, 2, \dots\}$ are defined by

$$X(t) = \text{the number of customers present at time } t,$$

and

$$X_n = \text{the number of customers present just prior to the } n\text{th arrival epoch.}$$

The stochastic processes $\{X(t)\}$ and $\{X_n\}$ are both regenerative. The regeneration epochs are the epochs at which an arriving customer finds the system empty. It is stated without proof that the assumption of $\lambda/\beta < 1$ implies that the processes have a finite mean cycle length. Thus we can define the time-average and the customer-average probabilities p_j and π_j by

$$p_j = \text{the long-run fraction of time that } j \text{ customers are present}$$

and

$$\pi_j = \text{the long-run fraction of customers who find } j \text{ other customers present upon arrival}$$

for $j = 0, 1, \dots$. Time averages are averages over time, and customer averages are averages over customers. To be precise, $p_j = \lim_{t\to\infty}(1/t)\int_0^t I_j(u)\,du$ and $\pi_j = \lim_{n\to\infty}(1/n)\sum_{k=1}^{n} I_k(j)$, where $I_j(t) = 1$ if j customers are present at time t and $I_j(t) = 0$ otherwise, and $I_n(j) = 1$ if j other customers are present just before the nth arrival epoch and $I_n(j) = 0$ otherwise. The probabilities p_j and π_j are related to each other by

$$\lambda\pi_{j-1} = \beta p_j, \quad j = 1, 2, \dots . \tag{2.7.1}$$

The proof of this result is instructive and is based on three observations. Before giving the three steps, let us say that the continuous-time process $\{X(t)\}$ makes an *upcrossing* from state $j-1$ to state j if a customer arrives and finds $j-1$

other customers present. The process $\{X(t)\}$ makes a *downcrossing* from state j to state $j-1$ if the service of a customer is completed and $j-1$ other customers are left behind.

Observation 1 Since customers arrive singly and are served singly, the long-run average number of upcrossings from $j-1$ to j per time unit equals the long-run average number of downcrossings from j to $j-1$ per time unit. This follows by noting that in any finite time interval the number of upcrossings from $j-1$ to j and the number of downcrossings from j to $j-1$ can differ at most by 1.

Observation 2 The long-run fraction of customers seeing $j-1$ other customers upon arrival is equal to

$$\frac{\text{the long-run average number of upcrossings from } j-1 \text{ to } j \text{ per time unit}}{\text{the long-run average number of arrivals per time unit}}$$

for $j = 1, 2, \ldots$. In other words, the long-run average number of upcrossings from $j-1$ to j per time unit equals $\lambda\pi_{j-1}$.

The latter relation for fixed j is in fact a special case of the Little relation (2.4.1) by assuming that each customer finding $j-1$ other customers present upon arrival pays \$1 (using this reward structure observation 2 can also be obtained directly from the renewal-reward theorem). Observations 1 and 2 do not use the assumption of exponential services and apply in fact to any regenerative queueing process in which customers arrive singly and are served singly.

Observation 3 For exponential services, the long-run average number of downcrossings from j to $j-1$ per time unit equals βp_j with probability 1 for each $j \geq 1$.

The proof of this result relies heavily on the PASTA property. To make this clear, fix j and note that service completions occur according to a Poisson process with rate β as long as the server is busy. Equivalently, we can assume that an exogenous Poisson process generates events at a rate of β, where a Poisson event results in a service completion only when there are j customers present. Thus, by part (a) of Theorem 2.4.1,

$$\beta E[I_j(t)] = E[D_j(t)] \quad \text{for } t > 0 \tag{2.7.2}$$

for any $j \geq 1$, where $I_j(t)$ is defined as the amount of time that j customers are present during $(0, t]$ and $D_j(t)$ is defined as the number of downcrossings from j to $j-1$ in $(0, t]$. Letting the constant d_j denote the long-run average number of downcrossings from j to $j-1$ per time unit, we have by the renewal-reward theorem that $\lim_{t\to\infty} D_j(t)/t = d_j$ with probability 1. Similarly, $\lim_{t\to\infty} I_j(t)/t = p_j$ with probability 1. The renewal-reward theorem also holds in the expected-value version. Thus, for any $j \geq 1$,

$$\lim_{t\to\infty} \frac{E[D_j(t)]}{t} = d_j \quad \text{and} \quad \lim_{t\to\infty} \frac{E[I_j(t)]}{t} = p_j.$$

Hence relation (2.7.2) gives that $d_j = \beta p_j$ for all $j \geq 1$. By observations 1 and 2 we have $d_j = \lambda\pi_{j-1}$. This gives $\lambda\pi_{j-1} = \beta p_j$ for all $j \geq 1$, as was to be proved.

In Chapter 3 the method of embedded Markov chains will be used to derive an explicit expression for the customer-average probabilities π_j.

EXERCISES

2.1 A street lamp is replaced by a new one upon failure and upon scheduled times $T, 2T, \ldots$. There is always a replacement at the scheduled times regardless of the age of the street lamp in use. The lifetimes of the street lamps are independent random variables and have a common Erlang $(2, \mu)$ distribution. What is the expected number of street lamps used in a scheduling interval?

2.2 The municipality of Gotham City has opened a depot for temporarily storing chemical waste. The amount of waste brought in each week has a gamma distribution with given shape parameter α and scale parameter λ. The amounts brought in during the successive weeks are independent of each other.

(a) What is the expected number of weeks until the total amount of waste in the depot exceeds the critical level L?

(b) Give an asymptotic estimate for the expected value from question (a).

2.3 Limousines depart from the railway station to the airport from the early morning till late at night. The limousines leave from the railway station with independent interdeparture times that are uniformly distributed between 10 and 20 minutes. Suppose you plan to arrive at the railway station at 3 o'clock in the afternoon. What are the estimates for the mean and the standard deviation of your waiting time at the railway station until a limousine leaves for the airport?

2.4 Consider the expression (2.1.4) for the renewal function $M(t)$.

(a) Prove that for any $k = 0, 1, \ldots$

$$\sum_{n=k+1}^{\infty} F_n(t) \leq \frac{F_k(t)F(t)}{1 - F(t)}$$

for any t with $F(t) < 1$. (*Hint*: use $P\{X_1 + \cdots + X_n \leq t\} \leq P\{X_1 + \cdots + X_k \leq t\}P\{X_{k+1} \leq t\} \cdots P\{X_n \leq t\}$.)

(b) Conclude that $M(t) < \infty$ for all $t \geq 0$.

2.5 Consider a renewal process with Erlang (r, λ) distributed interoccurrence times. Use the phase method to prove:

(a) For any $t > 0$,

$$P\{N(t) > k\} = \sum_{j=(k+1)r}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \quad k = 0, 1, \ldots .$$

(b) The excess variable γ_t is Erlang (j, λ) distributed with probability

$$p_j(t) = \sum_{k=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{kr-j}}{(kr - j)!}, \quad j = 1, \ldots, r.$$

2.6 Consider a continuous-time stochastic process $\{X(t), t \geq 0\}$ that can assume only the two states 1 and 2. If the process is currently in state i, it moves to the next state after an exponentially distributed time with mean $1/\lambda_i$ for $i = 1, 2$. The next state is state 1 with probability p_1 and state 2 with probability $p_2 = 1 - p_1$ irrespective of the past of the process.

(a) Use the renewal-reward model to find the long-run fraction of time the process $\{X(t)\}$ is in state i for $i = 1, 2$. Does $\lim_{t\to\infty} P\{X(t) = i\}$ exist for $i = 1, 2$? If so, what is the limit?

(b) Consider a renewal process in which the interoccurrence times have an H_2 distribution with density $p_1\lambda_1 e^{-\lambda_1 t} + p_2\lambda_2 e^{-\lambda_2 t}$. Argue that

$$\lim_{t\to\infty} P\{\gamma_t > x\} = \frac{p_1\lambda_2}{p_1\lambda_2 + p_2\lambda_1} e^{-\lambda_1 x} + \frac{p_2\lambda_1}{p_1\lambda_2 + p_2\lambda_1} e^{-\lambda_2 x}, \quad x \geq 0.$$

2.7 Consider a renewal process with Erlang (r, λ) distributed interoccurrence times. Let the probability $p_j(t)$ be defined as in part (b) of Exercise 2.5. Use the renewal-reward model to argue that $\lim_{t\to\infty} p_j(t) = 1/r$ for $j = 1, \ldots, r$ and conclude that

$$\lim_{t\to\infty} P\{\gamma_t > x\} = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{j-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}, \quad x \geq 0.$$

Generalize these results when the interoccurrence time is distributed as an Erlang (j, λ) random variable with probability β_j for $j = 1, \ldots, r$.

2.8 Consider the $E_r/D/\infty$ queueing system with infinitely many servers. Customers arrive according to a renewal process in which the interoccurence times have an Erlang (r,λ) distribution and the service time of each customer is a constant D. Each newly arriving customer gets immediately assigned a free server. Let $p_n(t)$ denote the probability that n servers will be busy at time t. Use an appropriate conditioning argument to verify that

$$\lim_{t\to\infty} p_0(t) = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{j-1} e^{-\mu D} \frac{(\mu D)^k}{k!}$$

$$\lim_{t\to\infty} p_n(t) = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{r-1} e^{-\mu D} \frac{(\mu D)^{r-j+1+(n-1)r+k}}{(r-j+1+(n-1)r+k)!}, \quad n \geq 1.$$

(*Hint*: the only customers present at time t are those customers who have arrived in $(t - D, t]$.)

2.9 The lifetime of a street lamp has a given probability distribution function $F(x)$ with probability density $f(x)$. The street lamp is replaced by a new one upon failure or upon reaching the critical age T, whichever occurs first. A cost of $c_f > 0$ is incurred for each failure replacement and a cost of $c_p > 0$ for each preventive replacement, where $c_p < c_f$. The lifetimes of the street lamps are independent of each other.

(a) Define a regenerative process and specify its regeneration epochs.

(b) Show that the long-run average cost per time unit under the age-replacement rule equals $g(T) = [c_p + (c_f - c_p)F(T)]/\int_0^T \{1 - F(x)\}\,dx$.

(c) Verify that the optimal value of T satisfies $g(T) = (c_f - c_p)r(T)$, where $r(x)$ is the failure rate function of the lifetime.

2.10 Consider the $M/G/\infty$ queue from Section 1.1.3 again. Let the random variable L be the length of a busy period. A busy period begins when an arrival finds the system empty and ends when there are no longer any customers in the system. Use the result (2.2.1) to argue that $E(L) = (e^{\lambda\mu} - 1)/\lambda$.

2.11 Consider an electronic system having n identical components that operate independently of each other. If a component breaks down, it goes immediately into repair. There are ample

repair facilities. Both the running times and the repair times are sequences of independent and identically distributed random variables. It is also assumed that these two sequences are independent of each other. The running time has a positive density on some interval. Denote by α the mean running time and by β the mean repair time.

(a) Prove that

$$\lim_{t\to\infty} P\{k \text{ components are in repair at time } t\} = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$, where $p = \beta/(\alpha + \beta)$.

(b) Argue that the limiting distribution in (a) becomes a Poisson distribution with mean $\lambda\beta$ when $n \to \infty$ and $1/\alpha \to 0$ such that n/α remains equal to the constant λ. Can you explain the similarity of this result with the insensitivity result (1.1.6) for the $M/G/\infty$ queue in Section 1.1.3?

2.12 A production process in a factory yields waste that is temporarily stored on the factory site. The amounts of waste that are produced in the successive weeks are independent and identically distributed random variables with finite first two moments μ_1 and μ_2. Opportunities to remove the waste from the factory site occur at the end of each week. The following control rule is used. If at the end of a week the total amount of waste present is larger than D, then all the waste present is removed; otherwise, nothing is removed. There is a fixed cost of $K > 0$ for removing the waste and a variable cost of $v > 0$ for each unit of waste in excess of the amount D.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

(c) Assuming that D is sufficiently large compared to μ_1, give an approximate expression for the average cost.

2.13 At a production facility orders arrive according to a renewal process with a mean interarrival time $1/\lambda$. A production is started only when N orders have accumulated. The production time is negligible. A fixed cost of $K > 0$ is incurred for each production set-up and holding costs are incurred at the rate of hj when j orders are waiting to be processed.

(a) Define a regenerative stochastic process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

(c) What value of N minimizes the long-run average cost per time unit?

2.14 Consider again Exercise 2.13. Assume now that it takes a fixed set-up time T to start a production. Any new order that arrives during the set-up time is included in the production run. Answer parts (a) and (b) from Exercise 2.13 for the particular case that the orders arrive according to a Poisson process with rate λ.

2.15 How do you modify the expression for the long-run average cost per time unit in Exercise 2.14 when it is assumed that the set-up time is a random variable with finite first two moments?

2.16 Consider Example 1.3.1 again. Assume that a fixed cost of $K > 0$ is incurred for each round trip and that a fixed amount $R > 0$ is earned for each passenger.

(a) Define a regenerative stochastic process and identify its regeneration epochs.

(b) Determine the long-run average net reward per time unit.

(c) Verify that the average reward is maximal for the unique value of T satisfying the equation $e^{-\mu T}(R\lambda T + R\lambda/\mu) = R\lambda/\mu - K$ when $R\lambda/\mu > K$.

2.17 Passengers arrive at a bus stop according to a Poisson process with rate λ. Buses depart from the stop according to a renewal process with interdeparture time A. Using renewal-reward processes, prove that the long-run average waiting time per passenger equals $E(A^2)/2E(A)$. Specify the regenerative process you need to prove this result. Can you give

a heuristic explanation of why the answer for the average waiting time is the same as the average residual life in a renewal process?

2.18 Consider a renewal process in which the interoccurrence times have a positive density on some interval. For any time t let the age variable δ_t denote the time elapsed since the last occurrence of an event. Use the renewal-reward model to prove that $\lim_{t\to\infty} E(\delta_t) = \mu_2/2\mu_1$, where μ_k is the kth moment of the interoccurrence times. (*Hint*: assume a cost at rate x when a time x has elapsed since the last occurrence of an event.)

2.19 A common car service between cities in Israel is a sheroot. A sheroot is a seven-seat cab that leaves from its stand as soon as it has collected seven passengers. Suppose that potential passengers arrive at the stand according to a Poisson process with rate λ. An arriving person who sees no cab at the stand goes elsewhere and is lost for the particular car service. Empty cabs pass the stand according to a Poisson process with rate μ. An empty cab stops only at the stand when there is no other cab.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run fraction of time there is no cab at the stand and determine the long-run fraction of customers who are lost. Explain why these two fractions are equal to each other.

2.20 Big Jim, a man of few words, runs a one-man business. This business is called upon by loan sharks to collect overdue loans. Big Jim takes his profession seriously and accepts only one assignment at a time. The assignments are classified by Jim into n different categories $j = 1, \ldots, n$. An assignment of type j takes him a random number of τ_j days and gives a random profit of ξ_j dollars for $j = 1, \ldots, n$. Assignments of the types $1, \ldots, n$ arrive according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_n$. Big Jim, once studying at a prestigious business school, is a muscleman with brains. He has decided to accept those type j assignments for which $E(\xi_j)/E(\tau_j)$ is at least g^* dollars per day for a carefully chosen value of g^* (in Exercise 7.4 you are asked to use Markov decision theory to determine g^*). Suppose that Big Jim only accepts type j assignments for $j = 1, \ldots, n_0$. An assignment can only be accepted when Big Jim is not at work on another assignment. Assignments that are refused are handled by a colleague of Big Jim.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average pay-off per time unit for Big Jim.

(c) Determine the long-run fraction of time Big Jim is at work and the long-run fraction of the assignments of the types $1, \ldots, n_0$ that are not accepted. Explain why these two fractions are equal to each other.

2.21 Consider the $(S-1, S)$ inventory model with back ordering from Section 1.1.3. What is the long-run fraction of customer demand that is back ordered? What is the long-run average amount of time a unit is kept in stock?

2.22 Consider a machine whose state deteriorates through time. The state of the machine is inspected at fixed times $t = 0, 1, \ldots$. In each period between two successive inspections the machine incurs a random amount of damage. The amounts of damage accumulate. The amounts of damage incurred in the successive periods are independent random variables having a common exponential distribution with mean $1/\alpha$. A compulsory repair of the machine is required when an inspection reveals a cumulative amount of damage larger than a critical level L. A compulsory repair involves a fixed cost of $R_c > 0$. A preventive repair at a lower cost of $R_p > 0$ is possible when an inspection reveals a cumulative amount of damage below or at the level L. The following control limit rule is used. A repair is done at each inspection that reveals a cumulative amount of damage larger than some repair limit z with $0 \le z < L$. It is assumed that each repair takes a negligible time and that after each repair the machine is as good as new.

(a) Define a regenerative process and identify its regeneration epochs.

(b) What is the expected number of periods between two successive repairs? What is the

probability that a repair involves the high repair cost R_c? Give the long-run average cost per time unit.

(c) Verify that the average cost is minimal for the unique solution z to the equation $\alpha z \exp[-\alpha(L - z)] = R_p/(R_c - R_p)$ when $\alpha L > R_p/(R_c - R_p)$.

2.23 A group of N identical machines is maintained by a single repairman. The machines operate independently of each other and each machine has a constant failure rate μ. Repair is done only if the number of failed machines has reached a given critical level R with $1 \leq R \leq N$. Then all failed machines are repaired simultaneously. Any repair takes a negligible time and a repaired machine is again as good as new. The cost of the simultaneous repair of R machines is $K + cR$, where $K, c > 0$. Also there is an idle-time cost of $\alpha > 0$ per time unit for each failed machine.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

2.24 The following control rule is used for a slow-moving expensive product. No more than one unit of the product is kept in stock. Each time the stock drops to zero a replenishment order for one unit is placed. The replenishment lead time is a positive constant L. Customers asking for the product arrive according to a renewal process in which the interarrival times are Erlang (r, λ) distributed. Each customer asks for one unit of the product. Each demand occurring while the system is out of stock is lost.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run fraction of demand that is lost.

(c) Determine the long-run fraction of time the system is out of stock. (*Hint:* use part (b) of Exercise 2.5.)

2.25 Jobs arrive at a station according to a renewal process. The station can handle only one job at a time, but has no buffer to store other jobs. An arriving job that finds the station busy is lost. The handling time of a job has a given probability density $h(x)$. Use renewal-reward theory to verify for this loss system that the long-run fraction of jobs that are rejected is given by $\int_0^\infty M(x)h(x)\,dx$ divided by $1 + \int_0^\infty M(x)h(x)\,dx$, where $M(x)$ is the renewal function in the renewal process describing the arrival of jobs. What is the long-run fraction of time that the station is busy? Simplify the formulas for the cases of deterministic and Poisson arrivals.

2.26 Use the renewal-reward theorem to prove relation (2.3.3) when customers arrive according to a renewal process and the stochastic processes $\{L(t)\}$ and $\{U_n\}$ regenerate themselves each time an arriving customer finds the system empty, where the cycle lengths have finite expectations. For ease assume the case of an infinite-capacity queue. Use the following relations:

(i) the long-run average reward earned per time unit = (the expected reward earned in one cycle)/(expected length of one cycle),

(ii) the long-run average amount paid per customer = (the expected amount earned in one cycle)/(expected number of arrivals in one cycle),

(iii) the long-run average arrival rate = (expected number of arrivals in one cycle)/(expected length of one cycle).

2.27 Let $\{X(t), t \geq 0\}$ be a continuous-time regenerative stochastic process whose state space is a subset of the non-negative reals. The cycle length is assumed to have a finite expectation. Denote by $\overline{P}(y)$ the long-run fraction of time that the process $\{X(t)\}$ takes on a value larger than y. Use the renewal-reward theorem to prove that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t X(u)\,du = \int_0^\infty \overline{P}(y)\,dy \quad \text{with probability 1.}$$

2.28 Consider a queueing system in which the continuous-time process $\{L(t)\}$ describing the number of customers in the system is regenerative, where the cycle length has a finite

expectation. Let p_j denote the long-run fraction of time that j customers are in the system and let L denote the long-run average number of customers in the system. Apply the result of Exercise 2.27 to conclude that $L = \sum_{j=1}^{\infty} j p_j$.

2.29 Verify that the Pollaczek–Khintchine formula for the average waiting time in the $M/G/1$ queue can also be written as

$$W_q = (1 - c_S^2) W_q(\text{det}) + c_S^2 W_q(\text{exp}).$$

This interpolation formula is very useful and goes back to Cox (1955).

2.30 A professional cleaner in the harbour of Rotterdam is faced with the decision to acquire a new clean installation for oil tankers. Oil tankers requiring a clean arrive according to a Poisson process with rate λ. The amount of time needed to clean a tanker has a given probability distribution with mean α and standard deviation β when the standard Fadar installation is used. Cleaning costs at a rate of $c > 0$ are incurred for each time unit this installation is in use. However, it is also possible to buy another installation. An installation that works z times as fast as the standard Fadar installation involves cleaning costs at a rate of cz^2 per time unit. In addition to the cleaning costs, a holding cost at rate of $h > 0$ is incurred for each tanker in the harbour. What is the long-run average cost per time unit as function of z? Assume that the cleaning installation can handle only one tanker at a time and assume that the cleaner has ample berths for tankers.

2.31 Liquid is put into an infinite-capacity buffer at epochs generated by a Poisson process with rate λ. The successive amounts of liquid that are put in the buffer are independent and identically distributed random variables with finite first two moments μ_1 and μ_2. The buffer is emptied at a constant rate of $\sigma > 0$ whenever it is not empty. Use the PASTA property to give an expression for the long-run average buffer content.

2.32 Consider the $M/G/1$ queue with two types of customers. Customers of the types 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. The service times of the customers are independent of each other, where the service times of type i customers are distributed as the random variable S_i having finite first two moments. Customers of type 1 have priority over customers of type 2 when the server is ready to start a new service. It is not allowed to interrupt the service of a type 2 customer when a higher-priority customer arrives. This queueing model is called the *non-pre-emptive priority* $M/G/1$ queue. Letting $\rho_i = \lambda_i E(S_i)$, it is assumed that $\rho_1 + \rho_2 < 1$.

(a) Use Little's formula to argue that the long-run fraction of time the server is servicing type i customers equals ρ_i for $i = 1, 2$. What is the long-run fraction of customers finding the server servicing a type i customer upon arrival?

(b) Extend the heuristic derivation of the Pollaczek–Khintchine formula to show

$$W_{q1} = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1 - \rho_1)} \quad \text{and} \quad W_{q2} = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1 - \rho_1)(1 - \rho_1 - \rho_2)},$$

where W_{qi} is defined as the long-run average waiting time in queue per type i customer for $i = 1, 2$.

(c) Use Little's formula to give a direct argument for the result that the overall average waiting time $W_{q1}\lambda_1/(\lambda_1 + \lambda_2) + W_{q2}\lambda_2/(\lambda_1 + \lambda_2)$ per customer is the same as the average waiting time per customer in the $M/G/1$ queue in which customers are served in order of arrival (view the non-pre-emptive priority rule as a rule that merely changes the order in which the customers are served).

2.33 Customers arrive at a single-server station according to a Poisson process with rate λ. A customer finding the server idle upon arrival gets served immediately, otherwise the customer enters a so-called orbit. A customer in orbit tries whether the server is idle after an

exponentially distributed time with mean $1/\nu$. If the server is idle, the customer gets served, otherwise the customer returns to orbit and tries again after an exponentially distributed time until the server is found free. The customers in orbit act independently of each other. The service times of the customers are independent random variables having the same general probability distribution. Letting the random variable S denote the service time of a customer, it is assumed that $\rho = \lambda E(S)$ is less than 1. For this model, known as the $M/G/1$ queue with *retrials*, define $L(t)$ as the number of customers in the system (service station plus orbit) at time t and define Q_n as the number of customers in orbit just after the nth service completion. Let $p_j = \lim_{t\to\infty} P\{L(t) = j\}$ and $q_j = \lim_{n\to\infty} P\{Q_n = j\}$ for $j \geq 0$.

(a) Use an up- and downcrossing argument to argue that $p_j = q_j$ for all $j \geq 0$.

(b) Letting $Q(z) = \sum_{j=0}^{\infty} q_j z^j$, prove that

$$Q(z) = A(z)\{\lambda R(z) + \nu R'(z)\},$$

where $A(z)$ is the generating function of the number of new customers arriving during the service time S and $R(z)$ is defined by $R(z) = \sum_{j=0}^{\infty} z^j q_j/(\lambda + j\nu)$. (*Hint*: under the condition that $Q_{n-1} = i$ it holds that $Q_n = Q_{n-1} + C_n$ with probability $\lambda/(\lambda + i\nu)$ and $Q_n = Q_{n-1} - 1 + C_n$ with probability $i\nu/(\lambda + i\nu)$, where C_n denotes the number of new customers arriving at the nth service time.)

(c) Prove that

$$Q(z) = \frac{(1-\rho)(1-z)A(z)}{A(z)-z} \exp\left[\frac{\lambda}{\nu}\int_1^z \frac{1-A(u)}{A(u)-u}\,du\right].$$

(*Hint*: use that $Q(z) = \lambda R(z) + \nu z R'(z)$, which follows directly from the definition of $R(z)$.)

(d) Show that the long-run average number of customers in the system is given by

$$L = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} + \frac{\lambda^2 E(S)}{\nu(1-\rho)}.$$

Retrial queues are in general much more difficult to analyse than queues without retrials. The Laplace transform for the waiting-time distribution in the $M/G/1$ queue with retrials is very complex; see also Artalejo *et al.* (2002).

2.34 Consider again the production system from Section 2.6 except that the system is now controlled in a different way when it becomes idle. Each time the production facility becomes empty of orders, the facility is used during a period of fixed length T for some other work in order to utilize the idle time. After this vacation period the facility is reactivated for servicing the orders only when at least one order is present; otherwise the facility is used again for some other work during a vacation period of length T. This utilization of idle time is continued until at least one order is present after the end of a vacation period. This control policy is called the T-policy. The cost structure is the same as in Section 2.6. Use renewal-reward theory to show that $K(1-\lambda\mu_1)(1-e^{-\lambda T})/T + \frac{1}{2}h\lambda T + \frac{1}{2}h\lambda^2\mu_2/(1-\lambda\mu_1)$ gives the long-run average cost per time unit under a T-policy.

2.35 Suppose that, at a communication channel, messages of types 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. Messages of type 1 finding the channel occupied upon arrival are lost, whereas messages of type 2 are temporarily stored in a buffer and wait until the channel becomes available. The channel can transmit only one message at a time. The transmission time of a message of type i has a general probability distribution with mean μ_i and the transmission times are independent of each other. It is assumed that $\lambda_2\mu_2 < 1$. Use the renewal-reward theorem to prove that the long-run fraction of time the channel is busy equals $(\rho_1 + \rho_2)/(1 + \rho_1)$, where $\rho_i = \lambda_i\mu_i$ for $i = 1, 2$.

(*Hint*: use results from Section 2.6 to obtain the expected amount of time elapsed between two arrivals finding the channel free.)

BIBLIOGRAPHIC NOTES

The very readable monograph of Cox (1962) contributed much to the popularization of renewal theory. A good account of renewal theory can also be found in the texts Ross (1996) and Wolff (1989). A basic paper on renewal theory and regenerative processes is that of Smith (1958), a paper which recognized the usefulness of renewal-reward processes in the analysis of applied probability problems. The book of Ross (1970) was influential in promoting the application of renewal-reward processes. The renewal-reward model has many applications in inventory, queueing and reliability. The illustrative queueing example from Section 2.6 is taken from the paper of Yadin and Naor (1963), which initiated the study of control rules for queueing systems. Example 2.2.3 is adapted from the paper of Vered and Yechiali (1979).

The first rigorous proof of $L = \lambda W$ was given by Little (1961) under rather strong conditions; see also Jewell (1967). Under very weak conditions a sample-path proof of $L = \lambda W$ was given by Stidham (1974). The important result that Poisson arrivals see time averages was taken for granted by earlier practitioners. A rigorous proof was given in the paper of Wolff (1982). The derivation of the Laplace transform of the waiting-time distribution in the $M/G/1$ queue is adapted from Cohen (1982) and the relation between this transform and the generating function of the queue size comes from Haji and Newell (1971).

REFERENCES

Artalejo, J.R., Falin, G.I. and Lopez-Herrero, M.J. (2002) A second order analysis of the waiting time in the $M/G/1$ retrial queue. *Asia-Pacific J. Operat. Res.*, **19**, 131–148.

Cohen, J.W. (1982) *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

Cox, D.R. (1955) The statistical analysis of congestion. *J. R. Statist. Soc. A.*, **118**, 324–335.

Cox, D.R. (1962) *Renewal Theory*. Methuen, London.

Haji, R. and Newell, G.F. (1971) A relation between stationary queue and waiting-time distribution. *J. Appl. Prob.*, **8**, 617–620.

Jewell, W.S. (1967) A simple proof of $L = \lambda W$. *Operat. Res.*, **15**, 1109–1116.

Keilson, J. (1979) *Markov Chain Models—Rarity and Exponentiality*. Springer-Verlag, Berlin.

Little, J.D.C. (1961) A proof for the queueing formula $L = \lambda W$. *Operat. Res.*, **9**, 383–387.

Miller, D.R. (1972) Existence of limits in regenerative processes. *Ann. Math. Statist.*, **43**, 1275–1282.

Ross, S.M. (1970) *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

Ross, S.M. (1996) *Stochastic Processes*, 2nd. edn. John Wiley & Sons, Inc., New York.

Smith, W.L. (1958) Renewal theory and its ramifications. *J. R. Statist. Soc. B*, **20**, 243–302.

Solovyez, A.D. (1971) Asymptotic behaviour of the time of first occurrence of a rare event in a regenerating process. *Engineering Cybernetics*, **9**, 1038–1048.

Stidham, S. Jr (1974) A last word on $L = \lambda W$. *Operat. Res.*, **22**, 417–421.

Takács, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, New York.

Vered, G. and Yechiali, U. (1979) Optimal structures and maintenance policies for PABX power systems. *Operat. Res.*, **27**, 37–47.

Wolff, R.W. (1982) Poisson arrivals see time averages. *Operat. Res.*, **30**, 223–231.

Wolff, R.W. (1989) *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs NJ.

Yadin, M. and Naor, P. (1963) Queueing systems with removable service station. *Operat. Res. Quart.*, **14**, 393–405.

CHAPTER 3

Discrete-Time Markov Chains

3.0 INTRODUCTION

The notion of what is nowadays called a Markov chain was devised by the Russian mathematician A.A. Markov when, at the beginning of the twentieth century, he investigated the alternation of vowels and consonants in Pushkin's poem *Onegin*. He developed a probability model in which the outcomes of successive trials are allowed to be dependent on each other such that each trial depends only on its immediate predecessor. This model, being the simplest generalization of the probability model of independent trials, appeared to give an excellent description of the alternation of vowels and consonants and enabled Markov to calculate a very accurate estimate of the frequency at which consonants occur in Pushkin's poem.

The Markov model is no exception to the rule that simple models are often the most useful models for analysing practical problems. The theory of Markov processes has applications to a wide variety of fields, including biology, computer science, engineering and operations research. A Markov process allows us to model the uncertainty in many real-world systems that evolve dynamically in time. The basic concepts of a Markov process are those of a *state* and of a *state transition*. In specific applications the modelling 'art' is to find an adequate state description such that the associated stochastic process indeed has the Markovian property that the knowledge of the present state is sufficient to predict the future stochastic behaviour of the process. In this chapter we consider discrete-time Markov processes in which state transitions only occur at *fixed* times. Continuous-time Markov processes in which the state can change at any time are the subject of Chapter 4. The discrete-time Markov chain model is introduced in Section 3.1. In this section considerable attention is paid to the modelling aspects. Most students find the modelling more difficult than the mathematics. Section 3.2 deals with the n-step transition probabilities and absorption probabilities. The main interest, however, is in the long-run behaviour of the Markov chain. In Section 3.3 we discuss both the existence of an equilibrium distribution and the computation of this distribution.

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

Several applications will be discussed as well. For didactical reasons not all of the results that are stated in Section 3.3 are proved in this section. Some of the proofs are deferred to a later section. In Section 3.4 we discuss computational methods for solving the equilibrium equations of the Markov chain. In particular, we give a simple but powerful method for computing the equilibrium distribution of an infinite-state Markov chain whose state probabilities exhibit a geometric tail behaviour. Section 3.5 deals with theoretical issues such as the state classification for Markov chains and proofs of the ergodic theorems used in earlier sections.

3.1 THE MODEL

A discrete-time Markov chain is a stochastic process which is the simplest generalization of a sequence of independent random variables. A Markov chain is a random sequence in which the dependency of the successive events goes back only one unit in time. In other words, the future probabilistic behaviour of the process depends only on the present state of the process and is not influenced by its past history. This is called the *Markovian* property. Despite its very simple structure the Markov chain model is extremely useful in a wide variety of practical probability problems. Let us first give an illustrative example.

Example 3.1.1 The drunkard's random walk

A drunkard starts a random walk in the middle of a square; see Figure 3.1.1. He performs a sequence of independent unit steps. Each step has equal probability $\frac{1}{4}$ of going north, south, east or west as long as the drunkard has not reached the edge of the square. The drunkard never leaves the square. Should he reach the boundary of the square, his next step is equally likely to be in one of the three remaining directions if he is not at a corner point, and is equally likely to be in two remaining directions otherwise. What stochastic process describes the drunkard's walk? What is the expected number of steps he needs to return to his starting point?

For $n = 0, 1, \ldots$, we define the random variable

$$X_n = \text{the position of the drunkard just after the } n\text{th step}$$

with the convention $X_0 = (0, 0)$. Let us say that the process $\{X_n\}$ is in state (x, y) when the current position of the drunkard is described by point (x, y). Then $\{X_n, n = 0, 1, \ldots\}$ is a discrete-time stochastic process with state space

$$I = \{(x, y) \mid x, y \text{ integer}, \ -N \leq x, y \leq N\}.$$

The successive states of the drunkard's process are not independent of each other, but are dependent. However, the dependence goes only one step back. The next position of the drunkard depends only on the current position and is not influenced by the earlier positions in the path of the drunkard. In other words, the drunkard's

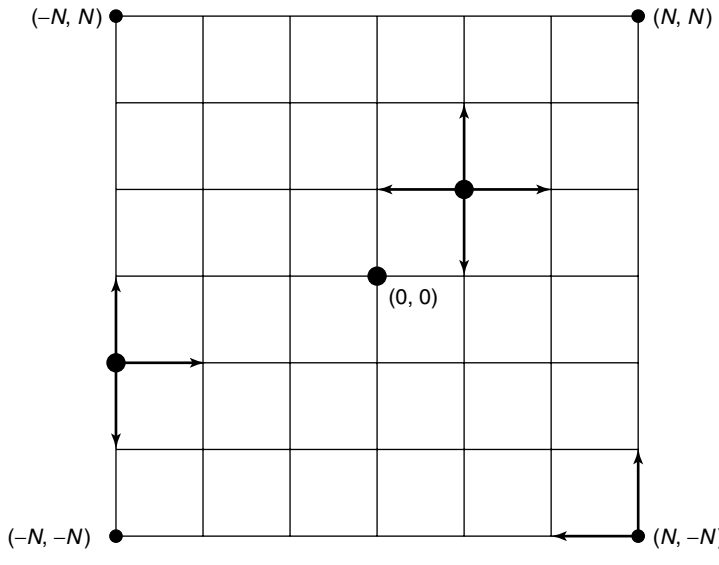


Figure 3.1.1 The drunkard's random walk

process $\{X_n\}$ has the Markovian property. We are now ready to give the general definition of a Markov chain.

Let $\{X_n, n = 0, 1, \dots\}$ be a sequence of random variables with state space I. We interpret the random variable X_n as the state of some dynamic system at time n. The set of possible values of the process is denoted by I and is assumed to be finite or countably infinite.

Definition 3.1.1 *The stochastic process* $\{X_n, n = 0, 1, \dots\}$ *with state space* I *is called a discrete-time Markov chain if, for each* $n = 0, 1, \dots,$

$$P\{X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n\} = P\{X_{n+1} = i_{n+1} \mid X_n = i_n\} \quad (3.1.1)$$

for all possible values of $i_0, \dots, i_{n+1} \in I$.

In the following, we consider only Markov chains with time-homogeneous transition probabilities; that is, we assume that

$$P\{X_{n+1} = j \mid X_n = i\} = p_{ij}, \quad i, j \in I,$$

independently of the time parameter n. The probabilities p_{ij} are called the *one-step transition probabilities* and satisfy

$$p_{ij} \geq 0, \quad i, j \in I, \quad \text{and} \quad \sum_{j \in I} p_{ij} = 1, \quad i \in I.$$

The Markov chain $\{X_n, n = 0, 1, \dots\}$ is completely determined by the probability distribution of the initial state X_0 and the one-step transition probabilities p_{ij}. In applications of Markov chains the art is:

(a) to choose the state variable(s) such that the Markovian property (3.1.1) holds,

(b) to determine the one-step transition probabilities p_{ij}.

Once this (difficult) modelling step is done, the rest is simply a matter of applying the theory that will be developed in the next sections. The student cannot be urged strongly enough to try the problems at the end of this chapter to acquire skills to model new situations. Let us return to the drunkard's walk.

Example 3.1.1 (continued) The drunkard's random walk

In this example we have already defined the state variable as the position of the drunkard. The process $\{X_n\}$ with X_n denoting the state just after the nth step of the drunkard is indeed a discrete-time Markov chain. The one-step transition probabilities are as follows. For any interior state (x, y) with $-N < x, y < N$, we have

$$p_{(x,y)(v,w)} = \begin{cases} \frac{1}{4} & \text{for } (v, w) = (x+1, y), (x-1, y), (x, y+1), (x, y-1), \\ 0 & \text{otherwise.} \end{cases}$$

For any boundary state (x, N) with $-N < x < N$, we have

$$p_{(x,y)(v,w)} = \begin{cases} \frac{1}{3} & \text{for } (v, w) = (x+1, N), (x-1, N), (x, N-1), \\ 0 & \text{otherwise.} \end{cases}$$

For the boundary state $(x, -N)$ with $-N < x < N$, (N, y) and $(N, -y)$ with $-N < y < N$, the one-step transition probabilities follow similarly. For the corner point $(x, y) = (N, N)$, we have

$$p_{(x,y)(v,w)} = \begin{cases} \frac{1}{2} & \text{for } (v, w) = (N-1, N), (N, N-1), \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for the corner points $(x, y) = (-N, N)$, $(-N, -N)$ and $(N, -N)$.

A variant of the drunkard's random walk problem is the problem in which the drunkard never chooses the same direction as was chosen in the previous step. Then we have to augment the state with an extra state variable in order to satisfy the Markovian property. The state of the drunkard after each step is now defined as (x, y, z), where (x, y) denotes the position of the drunkard and $z \in \{N, S, W, L\}$ denotes the direction of the last step. Letting X_n be the state of the drunkard's process just after the nth step (with the convention $X_0 = (0, 0)$), the stochastic process $\{X_n\}$ is a discrete-time Markov chain. It is left to the reader to write down the one-step transition probabilities of this process.

Example 3.1.2 A stock-control problem

The Johnson hardware shop carries adjustable-joint pliers as a regular stock item. The demand for this tool is stable over time. The total demand during a week has a Poisson distribution with mean λ. The demands in the successive weeks are independent of each other. Each demand that occurs when the shop is out of stock is lost. The owner of the shop uses a so-called periodic review (s, S) control rule for stock replenishment of the item. The inventory position is only reviewed at the beginning of each week. If the stock on hand is less than the reorder point s, the inventory is replenished to the order-up point S; otherwise, no ordering is done. Here s and S are given integers with $0 \leq s \leq S$. The replenishment time is negligible. What is the average ordering frequency and what is the average amount of demand that is lost per week?

These questions can be answered by the theory of Markov chains. In this example we take as state variable the stock on hand just prior to review. Let

$$X_n = \text{the stock on hand at the beginning of the } n\text{th week just prior to review,}$$

then the stochastic process $\{X_n\}$ is a discrete-time Markov chain with the finite state space $I = \{0, 1, \ldots, S\}$. It will be immediately clear that the Markovian property (3.1.1) is satisfied: the stock on hand at the beginning of the current week and the demand in the coming week determine the stock on hand at the beginning of the next week. It is not relevant how the stock level fluctuated in the past. To find the one-step transition probabilities $p_{ij} = P\{X_{n+1} = j \mid X_n = i\}$ we have to distinguish the cases $i \geq s$ and $i < s$. In the first case the stock on hand just *after* review equals i, while in the second case the stock on hand just after review equals S. For state $i \geq s$, we have

$$\begin{aligned} p_{ij} &= P\{\text{the demand in the coming week is } i - j\} \\ &= e^{-\lambda} \frac{\lambda^{i-j}}{(i-j)!}, \quad j = 1, \ldots, i. \end{aligned}$$

Note that this formula does not hold for $j = 0$. Then we have for $i \geq s$,

$$\begin{aligned} p_{i0} &= P\{\text{the demand in the coming week is } i \text{ or more}\} \\ &= \sum_{k=i}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1 - \sum_{k=0}^{i-1} e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

The other p_{ij} are zero for $i \geq s$. Similarly, we find for $i < s$

$$\begin{aligned} p_{ij} &= P\{\text{the demand in the coming week is } S - j\} \\ &= e^{-\lambda} \frac{\lambda^{S-j}}{(S-j)!}, \quad j = 1, \ldots, S, \end{aligned}$$

$$p_{i0} = P\{\text{the demand in the coming week is } S \text{ or more}\}$$
$$= \sum_{k=S}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1 - \sum_{k=0}^{S-1} e^{-\lambda} \frac{\lambda^k}{k!}.$$

The following example illustrates the powerful technique of *embedded Markov chains*. Many stochastic processes can be analysed by using properly chosen embedded stochastic processes that are discrete-time Markov chains. A classic example is the single-server $M/G/1$ queue with Poisson arrivals and general service times. The embedded process describing the number of customers left behind at the service completion epochs is a discrete-time Markov chain; see also Section 2.5. Another example is provided by the 'dual' queue with general interarrival times and exponential service times.

Example 3.1.3 The $GI/M/1$ queue

Customers arrive at a single-server station according to a renewal process, that is, the interarrival times of the customers are independent and identically distributed random variables. It is assumed that the interarrival time has a probability density $a(t)$. A customer who finds upon arrival that the server is idle enters service immediately; otherwise the customer waits in line. The service times of the successive customers are independent random variables having a common exponential distribution with mean $1/\mu$. The service times are also independent of the arrival process. A customer leaves the system upon service completion. This queueing system is usually abbreviated as the $GI/M/1$ queue. For any $t \geq 0$, define the random variable $X(t)$ by

$$X(t) = \text{the number of customers present at time } t.$$

The continuous-time stochastic process $\{X(t), t \geq 0\}$ does not possess the Markovian property that the future behaviour of the process depends only on its present state. Clearly, to predict the future behaviour of the process, the knowledge of the number of customers present does not suffice in general but the knowledge of the time elapsed since the last arrival is required too. Note that, by the memoryless property of the exponential distribution, the elapsed service time of the service in progress (if any) is not relevant. However, we can find an embedded Markov chain for the continuous-time process $\{X(t)\}$. Consider the process embedded at the epochs when customers arrive. At these epochs the time elapsed since the last arrival is known and equals zero. Define for $n = 0, 1, \ldots,$

$$X_n = \text{the number of customers present just prior to the } n\text{th arrival epoch}$$

with $X_0 = 0$ by convention. The embedded stochastic process $\{X_n, n = 0, 1, \ldots\}$ is a discrete-time Markov chain, since the exponential services are memoryless. This Markov chain has the countably infinite state space $I = \{0, 1, \ldots\}$. To find the one-step transition probabilities p_{ij} of the Markov chain, denote by A_n the

time between the arrival epochs of the nth and $(n+1)$th customer and let C_n denote the number of customers served during the interarrival time A_n. Note that $X_{n+1} = X_n + 1 - C_n$. The probability distribution of C_n obviously depends on X_n (= the number of customers seen by the nth arrival). The easiest way to find the probability distribution of C_n is to use the observation that service completions occur according to a Poisson process with rate μ as long as the server is busy. This observation is a consequence of the assumption of exponentially distributed service times and the relation between the Poisson process and the exponential distribution. By conditioning on the interarrival time A_n and using the law of total probability, we find for each state i that

$$\begin{aligned} p_{ij} &= P\{X_{n+1} = j \mid X_n = i\} \\ &= \int_0^\infty P\{i+1-j \text{ service completions during } A_n \mid A_n = t\} a(t)\, dt \\ &= \int_0^\infty e^{-\mu t} \frac{(\mu t)^{i+1-j}}{(i+1-j)!} a(t)\, dt, \quad 1 \le j \le i+1. \end{aligned} \tag{3.1.2}$$

This formula does not hold for $j = 0$. Why not? The probability p_{i0} is easiest to compute from

$$p_{i0} = 1 - \sum_{j=1}^{i+1} p_{ij}, \quad i = 0, 1, \ldots .$$

Obviously, $p_{ij} = 0$ for $j > i+1$ for each state i.

3.2 TRANSIENT ANALYSIS

This section deals with the transient analysis of the Markov chain $\{X_n, n = 0, 1, \ldots\}$ with state space I and one-step transition probabilities p_{ij} for $i, j \in I$. We first show how the one-step transition probabilities determine the probability of going from state i to state j in the next n steps. The n-step transition probabilities are defined by

$$p_{ij}^{(n)} = P\{X_n = j \mid X_0 = i\}, \quad i, j \in I$$

for any $n = 1, 2, \ldots$ Note that $p_{ij}^{(1)} = p_{ij}$. It is convenient to define

$$p_{ij}^{(0)} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

Theorem 3.2.1 (Chapman–Kolmogoroff equations) *For all* $n, m = 0, 1, \ldots,$

$$p_{ij}^{(n+m)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)}, \quad i, j \in I. \tag{3.2.1}$$

Proof A formal proof is as follows. By conditioning on the state of the Markov chain at time $t = n$, we find

$$\begin{aligned} P\{X_{n+m} = j \mid X_0 = i\} &= \sum_{k \in I} P\{X_{n+m} = j \mid X_0 = i, X_n = k\} P\{X_n = k \mid X_0 = i\} \\ &= \sum_{k \in I} P\{X_{n+m} = j \mid X_n = k\} P\{X_n = k \mid X_0 = i\} \\ &= \sum_{k \in I} P\{X_m = j \mid X_0 = k\} P\{X_n = k \mid X_0 = i\}, \end{aligned}$$

which verifies (3.2.1). Note that the second equality uses the Markovian property and the last equality uses the assumption of time homogeneity.

The theorem states that the probability of going from i to j in $n + m$ steps is obtained by summing the probabilities of the mutually exclusive events of going first from state i to some state k in n steps and then going from state k to state j in m steps. This explanation is helpful to memorize the equation (3.2.1). In particular, we have for any $n = 1, 2, \ldots,$

$$p_{ij}^{(n+1)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}, \quad i, j \in I. \tag{3.2.2}$$

Hence the n-step transition probabilities $p_{ij}^{(n)}$ can be recursively computed from the one-step transition probabilities p_{ij}. In fact the $p_{ij}^{(n)}$ are the elements of the n-fold matrix product $\mathbf{P}^n$, where $\mathbf{P}$ denotes the matrix whose (i, j)th element is the one-step transition probability p_{ij}. If the state space I is finite, the probabilities $p_{ij}^{(n)}$ can also be found by computing the eigenvalues and the eigenvectors of the matrix $\mathbf{P}$.

Example 3.2.1 The weather as Markov chain

On the Island of Hope the weather each day is classified as sunny, cloudy or rainy. The next day's weather depends only on the weather of the present day and not on the weather of the previous days. If the present day is sunny, the next day will be sunny, cloudy or rainy with respective probabilities 0.70, 0.10 and 0.20. The transition probabilities are 0.50, 0.25 and 0.25 when the present day is cloudy and they are 0.40, 0.30 and 0.30 when the present day is rainy. An interesting question is how often the weather is sunny, cloudy and rainy over a long period of time.

Let us first answer a simpler question, namely what the probability is of sunny weather three days later when the present day is rainy. To answer this question, we define a Markov chain $\{X_n\}$ with three states 1, 2 and 3. The process is in state 1 when the weather is sunny, in state 2 when the weather is cloudy and in state 3 when the weather is rainy. The matrix $\mathbf{P}$ of one-step transition probabilities p_{ij} is

given by

$$\mathbf{P} = \begin{pmatrix} 0.70 & 0.10 & 0.20 \\ 0.50 & 0.25 & 0.25 \\ 0.40 & 0.30 & 0.30 \end{pmatrix}.$$

To obtain the probability of having sunny weather three days from now, we need the matrix product $\mathbf{P}^3$:

$$\mathbf{P}^3 = \begin{pmatrix} 0.6015000 & 0.1682500 & 0.2302500 \\ 0.5912500 & 0.1756250 & 0.2331250 \\ 0.5855000 & 0.1797500 & 0.2347500 \end{pmatrix}.$$

This matrix shows that it will be sunny three days from now with probability 0.5855 when the present day is rainy. You could also ask: what is the probability distribution of the weather after many days? Intuitively you expect that this probability distribution does not depend on the present weather. This is indeed confirmed by the calculations:

$$\mathbf{P}^5 = \begin{pmatrix} 0.5963113 & 0.1719806 & 0.2317081 \\ 0.5957781 & 0.1723641 & 0.2318578 \\ 0.5954788 & 0.1725794 & 0.2319418 \end{pmatrix}$$

$$\mathbf{P}^{12} = \begin{pmatrix} 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \end{pmatrix} = \mathbf{P}^{13} = \mathbf{P}^{14} = \ldots .$$

In this example the n-step transition probability $p_{ij}^{(n)}$ converges for $n \to \infty$ to a limit which is independent of the initial state i. You see that the weather after many days will be sunny, cloudy or rainy with respective probabilities 0.5960, 0.1722 and 0.2318. Intuitively it will be clear that these probabilities also give the proportions of time the weather is sunny, cloudy and rainy over a long period. The limiting behaviour of the n-step transition probabilities is the subject of Section 3.3.

3.2.1 Absorbing States

A useful Markov chain model is the model with one or more absorbing states. A state is absorbing if the process cannot leave this state once it entered this state.

Definition 3.2.1 *A state i is said to be an absorbing state if $p_{ii} = 1$.*

The next example shows the usefulness of the Markov model with absorbing states.

Example 3.2.2 Success runs in roulette

A memorable event occurred in the casino of Monte Carlo on the evening of 18 August 1913. The roulette ball hit a red number 26 times in a row. In European

roulette the wheel has 37 compartments numbered $0, 1, \ldots, 36$, where the odd numbers are black and the even numbers except for the zero are red. An interesting question that naturally arises is: what is the probability that during the next m spins of the wheel there will be some sequence of r *consecutive* spins that all result either in r black numbers or in r red numbers for a given value of r?

This question can be answered by Markov chain theory. The idea is to define a Markov chain with $r+1$ states including an absorbing state. The process is said to be in state 0 when the last spin of the wheel resulted in a zero, while the process is said to be in state i with $1 \leq i < r$ when the *same* colour (red or black) appeared in the last i spins but this colour did not appear in the spin preceding the last i spins. The process is said to be in state r when the last r spins of the wheel have resulted in the same colour. The state r is taken as an absorbing state; imagine that the wheel sticks to the colour of the success run once a success run of length r has occurred. A success run of length r is said to occur when state r is reached. Denote by X_n the state of the process after the nth spin of the wheel, with $X_0 = 0$ by convention. The stochastic process $\{X_n\}$ is a discrete-time Markov chain. Its one-step transition probabilities are given by

$$\begin{aligned} p_{00} &= \frac{1}{37}, \quad p_{01} = \frac{36}{37}, \\ p_{i,i+1} &= p_{i1} = \frac{18}{37}, \; p_{i0} = \frac{1}{37} \quad \text{for } i = 1, \ldots, r-1 \\ p_{rr} &= 1. \end{aligned}$$

The other p_{ij} are zero. Since state r is absorbing, it is not possible that the process has visited state r before time t when the process is in some state $i \neq r$ at time t. Hence

$$\begin{aligned} &P\{\text{more than } m \text{ spins are needed to get a success run of length } r\} \\ &\quad = P\{X_k \neq r \text{ for } k = 1, \ldots, m \mid X_0 = 0\} \\ &\quad = P\{X_m \neq r \mid X_0 = 0\} = 1 - P\{X_m = r \mid X_0 = 0\} \\ &\quad = 1 - p_{0r}^{(m)}. \end{aligned}$$

The desired probability that a success run of length r will occur during the first m spins of the wheel is thus $p_{0r}^{(m)}$. How can we calculate this probability for $r = 26$ when N is of order 8 million (a rough estimate for the number of spins of the roulette wheel in Monte Carlo between the date of the founding of the casino and the date of 18 August 1913)? It is not advised to multiply the 27×27 matrix $\mathbf{P} = (p_{ij})$ 8 million times by itself. A more clever computation is based on

$$\mathbf{P}^2 = \mathbf{P} \times \mathbf{P}, \mathbf{P}^4 = \mathbf{P}^2 \times \mathbf{P}^2, \mathbf{P}^8 = \mathbf{P}^4 \times \mathbf{P}^4, \text{ etc.}$$

Taking $k = 23$, we have 2^k is about 8 million. Hence it suffices to do 23 matrix multiplications to get $p_{0,26}^{(m)}$ for $m = 2^{23}$. This gives the probability 0.061. Another

approach to analysing success runs is given in Appendix C and uses generating functions.

Example 3.2.3 A coin-tossing surprise

A fair coin is repeatedly flipped until the last three tosses either show the combination *TTH* or the combination *THH*. Here *H* means that the outcome of a toss is a head and *T* that it is a tail. What is the probability that the combination *TTH* occurs before the combination *THH*?

To answer this question, we define a Markov chain with eight states, including two absorbing states. Let state 0 mean the beginning of a game, state 1 = the first toss is *H*, state 2 = the first toss is *T*, state 3 = the last two tosses show *HH*, state 4 = the last two tosses show *HT*, state 5 = the last two tosses show *TT*, state 6 = the last two tosses show *TH*, state 7 = the last three tosses show *TTH* and state 8 = the last three tosses show *THH*. The states 7 and 8 are taken absorbing. It is implicit in the definition of the states 3, 4, 5, 6 that the combinations *TTH* and *THH* have not appeared before. The Markov chain that describes the evolution of the state of the system has the one-step transition probabilities

$$p_{01} = p_{02} = \tfrac{1}{2},\ p_{13} = p_{14} = \tfrac{1}{2},\ p_{25} = p_{26} = \tfrac{1}{2},$$
$$p_{33} = p_{34} = \tfrac{1}{2},\ p_{45} = p_{46} = \tfrac{1}{2},\ p_{55} = p_{57} = \tfrac{1}{2},$$
$$p_{63} = p_{68} = \tfrac{1}{2},\ p_{77} = 1,\ p_{88} = 1,\ \text{the other } p_{ij} = 0.$$

The Markov chain will ultimately be absorbed in one of the states 7 and 8 (this fact can formally be proved by proceeding as in the proof of Theorem 3.2.2 below and replacing the states 7 and 8 by a single absorbing state). Denote by f_i the probability that the Markov chain is ultimately absorbed in state 7 starting from state i. The probability f_0 gives the desired probability that the combination *TTH* occurs before the combination *THH*. The probabilities $f_0, \ldots, f_6$ satisfy a system of linear equations. The equation for f_i follows by conditioning on the next state after the current state i. This gives

$$f_0 = \tfrac{1}{2}f_1 + \tfrac{1}{2}f_2, \qquad f_1 = \tfrac{1}{2}f_3 + \tfrac{1}{2}f_4, \quad f_2 = \tfrac{1}{2}f_5 + \tfrac{1}{2}f_6,$$
$$f_3 = \tfrac{1}{2}f_3 + \tfrac{1}{2}f_4, \qquad f_4 = \tfrac{1}{2}f_5 + \tfrac{1}{2}f_6,$$
$$f_5 = \tfrac{1}{2}f_5 + \tfrac{1}{2} \times 1, \qquad f_6 = \tfrac{1}{2}f_3 + \tfrac{1}{2} \times 0.$$

The solution of these equations is $(f_0, \ldots, f_6) = (\frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, 1, \frac{1}{3})$. The desired probability is thus $\frac{2}{3}$. A surprising result for many people. Can you give a simple explanation why the sought probability is not equal to $\frac{1}{2}$?

3.2.2 Mean First-Passage Times

Example 3.1.1 asked how to find the expected number of steps the drunkard needs to return to his starting point. More generally, consider a Markov chain $\{X_n\}$ for which

(a) the state space I is finite,

(b) there is some state r such that for each state $i \in I$ there is an integer $n(= n_i)$ such that $p_{ir}^{(n)} > 0$.

What is the mean return time from state r to itself? Let

$$\tau = \min\{n \geq 1 \mid X_n = r\},$$

To calculate $\mu_{rr} = E(\tau | X_0 = r)$, we need the mean visit times

$$\mu_{ir} = E(\tau \mid X_0 = i)$$

for each state $i \neq r$. By conditioning on the next state after state r,

$$\mu_{rr} = 1 + \sum_{j \in I, j \neq r} p_{rj} \mu_{jr}. \tag{3.2.3}$$

The μ_{ir} with $i \neq r$ are found by solving a system of linear equations. For notational convenience, number the states as $1, \ldots, N$ and let state r be numbered as N.

Theorem 3.2.2 *The mean visit times μ_{iN} for $i \neq N$ are the unique solution to the linear equations*

$$\mu_{iN} = 1 + \sum_{j=1}^{N-1} p_{ij} \mu_{jN}, \quad i = 1, \ldots, N-1. \tag{3.2.4}$$

Proof The equation for μ_{iN} follows by conditioning on the next state visited after state i. To prove that the linear equations have a unique solution we use the trick of making state N absorbing for a modified Markov chain. Let $\widehat{\mathbf{P}} = (\widehat{p}_{ij})$, $i, j \in I$ be the Markov matrix obtained by replacing the Nth row in the matrix $\mathbf{P} = (p_{ij})$, $i, j \in I$ by $(0, 0, \ldots, 1)$. The mean first passage times μ_{jN} for $j = 1, \ldots, N-1$ are not changed by making state N absorbing. Denote by $\mathbf{Q} = (q_{ij})$ the $(N-1) \times (N-1)$ submatrix that results by omitting the Nth row and the Nth column in the matrix $\mathbf{P}$. Let the vectors $\mu = (\mu_{1N}, \ldots, \mu_{N-1,N})$ and $\mathbf{e} = (1, \ldots, 1)$. Then we can write (3.2.4) in matrix notation as

$$\mu = \mathbf{e} + \mathbf{Q}\mu. \tag{3.2.5}$$

Since state N is absorbing for the Markov matrix $\widehat{\mathbf{P}}$, we have for each $n \geq 1$ that

$$q_{ij}^{(n)} = \widehat{p}_{ij}^{(n)}, \quad i, j = 1, \ldots, N-1, \tag{3.2.6}$$

where the $q_{ij}^{(n)}$ and the $\widehat{p}_{ij}^{(n)}$ are the elements of the n-fold matrix products $\mathbf{Q}^n$ and $\widehat{\mathbf{P}}^n$. State N can be reached from each starting state $i \neq N$ under the Markov matrix $\widehat{\mathbf{P}}$, since by assumption (b) $\widehat{p}_{iN}^{(n)} \geq p_{iN}^{(n)} > 0$ for some $n \geq 1$. Further, state N is absorbing under $\widehat{\mathbf{P}}$. This implies that

$$\lim_{n\to\infty} \widehat{p}_{ij}^{(n)} = 0 \quad \text{for all } i, j = 1, \ldots, N-1,$$

as a special case of Lemma 3.2.3 below. Hence, by (3.2.6), $\lim_{n\to\infty} \mathbf{Q}^n = \mathbf{0}$. By a standard result from linear algebra, it now follows that (3.2.5) has the unique solution

$$\mu = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{e}. \tag{3.2.7}$$

This completes the proof that the linear equations (3.2.4) have a unique solution.

Example 3.1.1 (continued) The drunkard's random walk

The drunkard moves over a square with the corner points (N, N), $(-N, N)$, $(-N, -N)$ and $(-N, N)$. It is interesting to see how the mean return time to the starting point depends on N. Let $\mu_{00}(N)$ denote the expected number of steps the drunkard needs to return to the starting point $(0, 0)$. For fixed N the mean return time $\mu_{00}(N)$ can be computed by solving a system of linear equations of the form (3.2.4) and next using (3.2.3). Table 3.2.1 gives the values of $\mu_{00}(N)$ for several values of N. The computations indicate that $\mu_{00}(N) \to \infty$ as $N \to \infty$. This result is indeed true and can be theoretically proved by the theory of Markov chains; see for example Feller (1950).

3.2.3 Transient and Recurrent States

Many applications of Markov chains involve chains in which some of the states are absorbing and the other states are transient. An absorbing state is a special case of a recurrent state. To define the concepts of transient states and recurrent states, we need first to introduce the first-passage time probabilities. Let $\{X_n\}$ be a discrete-time Markov chain with state space I (finite or countably infinite) and one-step transition probabilities p_{ij}, $i, j \in I$. For any $n = 1, 2, \ldots$, let the *first-passage time probability* $f_{ij}^{(n)}$ be defined by

$$f_{ij}^{(n)} = P\{X_n = j, X_k \neq j \text{ for } 1 \leq k \leq n-1 \mid X_0 = i\}, \quad i, j \in I. \tag{3.2.8}$$

Table 3.2.1 The mean return time to the origin

N	1	2	5	10	25	50
$\mu_{00}(N)$	6	20	110	420	2550	10 100

In other words, $f_{ij}^{(n)}$ is the probability that the first transition of the process into state j is at time $t = n$ when the process starts in state i. Next define the probabilities f_{ij} by

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}. \tag{3.2.9}$$

Then $f_{ij} = P\{X_n = j \text{ for some } n \geq 1 \mid X_0 = i\}$ denotes the probability that the process *ever* makes a transition into state j when the process starts in state i.

Definition 3.2.2 *A state i is said to be transient if $f_{ii} < 1$ and is said to be recurrent if $f_{ii} = 1$.*

Denoting for each state $i \in I$ the probability Q_{ii} by

$$Q_{ii} = P\{X_n = i \text{ for infinitely many values of } n \mid X_0 = i\},$$

it is not difficult to verify that $Q_{ii} = 0$ if i is transient and $Q_{ii} = 1$ if i is recurrent. A useful characterization of a transient state is given by the result that a state i is *transient* if and only if

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty. \tag{3.2.10}$$

To see this, fix $i \in I$ and define the indicator variable I_n as $I_n = 1$ if $X_n = i$ and $I_n = 0$ otherwise. Then $\sum_{n=1}^{\infty} I_n$ represents the number of visits of the Markov chain to state i over the epochs $t = 1, 2, \ldots$. Since $E(I_n \mid X_0 = i) = P\{X_n = i \mid X_0 = i\} = p_{ii}^{(n)}$, it follows that

$$E\left(\sum_{n=1}^{\infty} I_n \mid X_0 = i\right) = \sum_{n=1}^{\infty} E(I_n \mid X_0 = i) = \sum_{n=1}^{\infty} p_{ii}^{(n)}, \tag{3.2.11}$$

where the interchange of expectation and summation is justified by the nonnegativity of the I_n. On the other hand, letting $N = \sum_{n=1}^{\infty} I_n$, the distribution of the number of visits to state i satisfies $P\{N \geq k \mid X_0 = i\} = (f_{ii})^k$ for $k \geq 0$ and so, by the well-known relation $E(N) = \sum_{j=0}^{\infty} P\{N > j\}$, we find

$$E\left(\sum_{n=1}^{\infty} I_n \mid X_0 = i\right) = \sum_{k=1}^{\infty} (f_{ii})^k.$$

Hence $E\left(\sum_{n=1}^{\infty} I_n \mid X_0 = i\right) = \infty$ when $f_{ii} = 1$ and equals $f_{ii}/(1 - f_{ii}) < \infty$ otherwise. This result and (3.2.11) prove that state i is transient only if (3.2.10) holds.

Lemma 3.2.3 *Suppose that state j is transient. Then, for any state $i \in I$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Proof By (3.2.10), $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ and thus $\lim_{n\to\infty} p_{jj}^{(n)} = 0$. Take now a starting state i with $i \neq j$. By conditioning on the first epoch at which the process makes a transition into state j, we obtain the useful relation

$$p_{ij}^{(n)} = \sum_{k=1}^{n} p_{jj}^{(n-k)} f_{ij}^{(k)}, \quad n = 1, 2, \ldots . \tag{3.2.12}$$

Since $\lim_{n\to\infty} p_{jj}^{(n)}$ exists and $\sum_{k=1}^{\infty} f_{ij}^{(k)} = f_{ij} < \infty$, it follows from the bounded convergence theorem in Appendix A that

$$\lim_{n\to\infty} p_{ij}^{(n)} = f_{ij} \lim_{n\to\infty} p_{jj}^{(n)}. \tag{3.2.13}$$

Since $\lim_{n\to\infty} p_{jj}^{(n)} = 0$, the lemma now follows.

The limiting behaviour of $p_{ij}^{(n)}$ as $n \to \infty$ for a recurrent state j will be discussed in Section 3.3. It will be seen that this limit does not always exist. For a recurrent state j an important concept is the *mean recurrence time* μ_{jj} which is defined by

$$\mu_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}. \tag{3.2.14}$$

In other words, μ_{jj} is the expected number of transitions needed to return from state j to itself. A recurrent state j is said to be *positive recurrent* if $\mu_{jj} < \infty$ and is said to be *null-recurrent* if $\mu_{jj} = \infty$. In Section 3.5 it will be seen that null-recurrency can only occur in Markov chains with an infinite state space. To illustrate this, consider the Markov chain $\{X_n\}$ describing the drunkard's walk on an infinite square in Example 3.1.1 ($N = \infty$). It can be shown for this infinite-state random walk that each state (x, y) is recurrent, but the mean recurrence time of each state is ∞ so that all states are null-recurrent. The same holds for the infinite-state Markov chain describing the symmetric random walk on the integers ($p_{i,i+1} = p_{i,i-1} = \frac{1}{2}$ for any integer i). However, for the symmetric random walk on an infinite lattice in three or more dimensions, the corresponding Markov chain has the property that all states are transient (in three dimensions, the probability of ever returning to the origin when starting there equals 0.3405). These remarkable results will not be proved here, but are mentioned to show that Markov chains with an infinite state space are intrinsically more complex than finite-state Markov chains.

3.3 THE EQUILIBRIUM PROBABILITIES

This section deals with the long-run behaviour of the Markov chain $\{X_n\}$. In particular, we discuss the characterization of the equilibrium distribution of the process and a formula for the long-run average cost per time unit when a cost structure is imposed on the Markov chain. In this section the emphasis is on giving insights into the long-run behaviour of the Markov chain. Most of the proofs are deferred to Section 3.5.

3.3.1 Preliminaries

A natural question for a Markov chain $\{X_n\}$ is whether the n-step probabilities $p_{ij}^{(n)}$ always have a limit as $n \to \infty$. The answer to this question is negative as shown by the following counterexample. Consider a Markov chain with state space $I = \{1, 2\}$ and one-step transition probabilities p_{ij} with $p_{12} = p_{21} = 1$ and $p_{11} = p_{22} = 0$. In this example the n-step transition probabilities $p_{ij}^{(n)}$ alternate between 0 and 1 for $n = 1, 2, \ldots$ and hence have no limit as $n \to \infty$. The reason is the *periodicity* in this Markov chain example. In our treatment of Markov chains we will not give a detailed discussion on the relation between the limiting behaviour of the $p_{ij}^{(n)}$ and the issue of periodicity. The reason is that our treatment of Markov chains emphasizes the study of long-run averages. As explained in Section 2.2, the long-run average behaviour of a stochastic process is in general much easier to handle than its limiting behaviour. More importantly, long-run averages are usually required in the analysis of practical applications. In the next theorem we prove that for each Markov chain $\{X_n\}$ the Cesaro limit of the n-step transition probabilities always exists.

Theorem 3.3.1 *For all $i, j \in I$, $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} p_{ij}^{(k)}$ always exists. For any $j \in I$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{jj}^{(k)} = \begin{cases} \frac{1}{\mu_{jj}} & \text{if state } j \text{ is recurrent,} \\ 0 & \text{if state } j \text{ is transient,} \end{cases} \tag{3.3.1}$$

where μ_{jj} denotes the mean recurrence time from state j to itself. Also,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = f_{ij} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{jj}^{(k)} \tag{3.3.2}$$

for any $i, j \in I$, where f_{ij} is the probability that the process ever makes a transition into state j when the process starts in state i.

Proof For a transient state j we have by Lemma 3.2.3 that $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for all $i \in I$. Using the well-known result that the Cesaro limit is equal to the ordinary limit whenever the latter limit exists, the results (3.3.1) and (3.3.2) follow

for transient states j. Fix now a recurrent state j. By the definition of recurrence, we have $f_{jj} = 1$. The times between successive visits to state j are independent and identically distributed random variables with mean μ_{jj}. In other words, visits of the Markov chain to state j can be seen as renewals. Denote by $N(t)$ the number of visits of the Markov chain to state j during the first t transition epochs. Then, by Lemma 2.2.2,

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mu_{jj}} \quad \text{with probability 1.} \tag{3.3.3}$$

This limiting result holds for both $\mu_{jj} < \infty$ and $\mu_{jj} = \infty$. In other words, the long-run average number of transitions to state j per time unit equals $1/\mu_{jj}$ with probability 1 when the process starts in state j. Define the indicator variable

$$I_k = \begin{cases} 1 & \text{if the process visits state } j \text{ at time } k, \\ 0 & \text{otherwise.} \end{cases}$$

Since $N(n) = I_1 + \cdots + I_n$, we can rewrite (3.3.3) as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k = \frac{1}{\mu_{jj}} \quad \text{with probability 1.} \tag{3.3.4}$$

Obviously,

$$E(I_k \mid X_0 = j) = P\{X_k = j \mid X_0 = j\} = p_{jj}^{(k)}.$$

Noting that $(1/n)\sum_{k=1}^{n} I_k$ is bounded by 1 and using the bounded convergence theorem from Appendix A, it follows from (3.3.4) that

$$\frac{1}{\mu_{jj}} = E\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k \mid X_0 = j\right) = \lim_{n\to\infty} E\left(\frac{1}{n} \sum_{k=1}^{n} I_k \mid X_0 = j\right)$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E\,(I_k \mid X_0 = j) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{jj}^{(k)}.$$

It remains to prove that (3.3.2) holds for any state $i \neq j$. To do so, we use the relation (3.2.12) which was derived in the proof of Lemma 3.2.3. Averaging this relation over $n = 1, \ldots, m$, interchanging the order of summation and letting $m \to \infty$, the relation (3.3.2) follows in the same way as (3.2.13).

Another natural question is under which condition the effect of the initial state of the process fades away as time increases so that $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} p_{ij}^{(k)}$ does not depend on the initial state $X_0 = i$ for each $j \in I$. We need some condition as the following example shows. Take a Markov chain with state space $I = \{1, 2\}$ and the one-step transition probabilities p_{ij} with $p_{11} = p_{22} = 1$ and $p_{12} = p_{21} = 0$. In this example $p_{11}^{(n)} = 1$ and $p_{21}^{(n)} = 0$ for all $n \geq 1$ so that $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} p_{i1}^{(k)}$

depends on the initial state i. The reason is that in this Markov chain example there are two disjoint closed sets of states.

Definition 3.3.1 *A non-empty set C of states is said to be closed if*

$$p_{ij} = 0 \quad \text{for } i \in C \text{ and } j \notin C,$$

that is, the process cannot leave the set C once the process is in the set C.

For a finite-state Markov chain having no two disjoint closed sets it is proved in Theorem 3.5.7 that $f_{ij} = 1$ for all $i \in I$ when j is a recurrent state. For such a Markov chain it then follows from (3.3.2) that $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} p_{ij}^{(k)}$ does not depend on the initial state i when j is recurrent. This statement is also true for a transient state j, since then the limit is always equal to 0 for all $i \in I$ by Lemma 3.2.3. For the case of an infinite-state Markov chain, however, the situation is more complex. That is why we make the following assumption.

Assumption 3.3.1 *The Markov chain $\{X_n\}$ has some state r such that $f_{ir} = 1$ for all $i \in I$ and $\mu_{rr} < \infty$.*

In other words, the Markov chain has a regeneration state r that is ultimately reached from each initial state with probability 1 and the number of steps needed to return from state r to itself has a finite expectation. The assumption is satisfied in most practical applications. For a finite-state Markov chain the Assumption 3.3.1 is automatically satisfied when the Markov chain has no two disjoint closed sets; see Theorem 3.5.7. The state r from Assumption 3.3.1 is a positive recurrent state. Assumption 3.3.1 implies that the set of recurrent states is not empty and that there is a single closed set of recurrent states. Moreover, by Lemma 3.5.8 we have for any recurrent state j that $f_{ij} = 1$ for all $i \in I$ and $\mu_{jj} < \infty$. Summarizing, under Assumption 3.3.1 we have both for a finite-state and an infinite-state Markov chain that $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} p_{ij}^{(k)}$ does not depend on the initial state i for all $j \in I$. In the next subsection it will be seen that the Cesaro limits give the equilibrium distribution of the Markov chain.

3.3.2 The Equilibrium Equations

We first give an important definition for a Markov chain $\{X_n\}$ with state space I and one-step transition probabilities p_{ij}, $i, j \in I$.

Definition 3.3.2 *A probability distribution $\{\pi_j,\ j \in I\}$ is said to be an equilibrium distribution for the Markov chain $\{X_n\}$ if*

$$\pi_j = \sum_{k\in I} \pi_k p_{kj}, \quad j \in I. \tag{3.3.5}$$

An explanation of the term equilibrium distribution is as follows. Suppose that the initial state of the process $\{X_n\}$ is chosen according to

$$P\{X_0 = j\} = \pi_j, \quad j \in I.$$

Then, for each $n = 1, 2, \ldots,$

$$P\{X_n = j\} = \pi_j, \quad j \in I.$$

In other words, starting the process according to the equilibrium distribution leads to a process that operates in an equilibrium mode. The proof is simple and is based on induction. Suppose that $P\{X_m = j\} = \pi_j$, $j \in I$ for some $m \geq 0$. Then

$$\begin{aligned} P\{X_{m+1} = j\} &= \sum_{k \in I} P\{X_{m+1} = j \mid X_m = k\} P\{X_m = k\} \\ &= \sum_{k \in I} p_{kj} \pi_k = \pi_j, \quad j \in I. \end{aligned}$$

An important question is: does the Markov chain have an equilibrium distribution, and if it has, is this equilibrium distribution unique? The answer to this question is positive when Assumption 3.3.1 is satisfied.

Theorem 3.3.2 *Suppose that the Markov chain $\{X_n\}$ satisfies Assumption 3.3.1. Then the Markov chain $\{X_n\}$ has a unique equilibrium distribution $\{\pi_j, j \in I\}$. For each state j,*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \pi_j \tag{3.3.6}$$

independently of the initial state i. Moreover, let $\{x_j, j \in I\}$ with $\sum_{j \in I} |x_j| < \infty$ be any solution to the equilibrium equations

$$x_j = \sum_{k \in I} x_k p_{kj}, \quad j \in I. \tag{3.3.7}$$

Then, for some constant c, $x_j = c\pi_j$ for all $j \in I$.

The proof of this important ergodic theorem is given in Section 3.5. It follows from Theorem 3.3.2 that the *equilibrium probabilities* π_j are the unique solution to the *equilibrium equations* (3.3.5) in conjunction with the normalizing equation

$$\sum_{j \in I} \pi_j = 1. \tag{3.3.8}$$

Interpretation of the π_j

Using elementary results from renewal theory, we have already seen from the proof of Theorem 3.3.1 that for any state j,

$$\begin{aligned} &\text{the long-run average number of visits to state } j \\ &\text{per time unit} = \pi_j \quad \text{with probability 1} \end{aligned} \tag{3.3.9}$$

when the process starts in state j. Under Assumption 3.3.1, the interpretation (3.3.9) can easily be shown to hold for each starting state $i \in I$ (this is obvious for a transient state j and, by Lemma 3.5.8, a recurrent state j will be reached from each initial state $X_0 = i$ after finitely many transitions with probability 1). The proof of Theorem 3.3.1 also showed that

$$\pi_j = \frac{1}{\mu_{jj}} \quad \text{for each recurrent state } j, \tag{3.3.10}$$

where μ_{jj} is the mean recurrence time from state j to itself. The interpretation (3.3.9) is most useful for our purposes. Using this interpretation, we can also give a physical interpretation of the equilibrium equation (3.3.5). Each visit to state j means a transition to state j (including self-transitions) and subsequently a transition from state j. Thus

$$\begin{aligned} &\text{the long-run average number of transitions from state } j \\ &\text{per time unit} = \pi_j \end{aligned}$$

and

$$\begin{aligned} &\text{the long-run average number of transitions from state } k \text{ to state } j \\ &\text{per time unit} = \pi_k p_{kj}. \end{aligned}$$

This latter relation gives

$$\begin{aligned} &\text{the long-run average number of transitions to state } j \\ &\text{per time unit} = \sum_{k \in I} \pi_k p_{kj}. \end{aligned}$$

By physical considerations, the long-run average number of transitions *to* state j per time unit must be equal to the long-run average number of transitions *from* state j per time unit. Why? Hence the equilibrium equations express that the long-run average number of transitions *from* state j per time unit equals the long-run average number of transitions *to* state j per time unit for all $j \in I$. The simplest way to memorize the equilibrium equations is provided by the following heuristic. Suppose that $\lim_{n\to\infty} p_{ij}^{(n)}$ exists so that $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$. Next apply

the heuristic reasoning

$$\pi_j = P\{X_\infty = j\} = \sum_{k \in I} P\{X_\infty = j \mid X_{\infty-1} = k\} P\{X_{\infty-1} = k\}$$

$$= \sum_{k \in I} p_{kj} \pi_k, \quad j \in I. \tag{3.3.11}$$

Example 3.2.1 (continued) The weather as Markov chain

In this example the three-state Markov chain $\{X_n\}$ has no two disjoint closed sets and thus has a unique equilibrium distribution. The equilibrium probabilities π_1, π_2 and π_3 can be interpreted as the fractions of time the weather is sunny, cloudy or rainy over a very long period of time. The probabilities π_1, π_2 and π_3 are the unique solution to the equilibrium equations

$$\pi_1 = 0.70\pi_1 + 0.50\pi_2 + 0.40\pi_3$$

$$\pi_2 = 0.10\pi_1 + 0.25\pi_2 + 0.30\pi_3$$

$$\pi_3 = 0.20\pi_1 + 0.25\pi_3 + 0.30\pi_3$$

together with the normalizing equation $\pi_1 + \pi_2 + \pi_3 = 1$. To get a square system of linear equations, it is permitted to delete one of the equilibrium equations. The solution is

$$\pi_1 = 0.5960, \pi_2 = 0.1722, \pi_3 = 0.2318$$

in accordance with earlier calculations in Section 3.2.

Example 3.1.2 (continued) A stock-control problem

In this example the Markov chain $\{X_n\}$ describing the stock on hand just prior to review has a finite state space and has no two disjoint closed sets (e.g. state 0 can be reached from each other state). Hence the Markov chain has a unique equilibrium distribution. The equilibrium probability π_j denotes the long-run fraction of weeks for which the stock on hand at the end of the week equals j for $j = 0, 1, \ldots, S$. Thus

$$\text{the long-run average frequency of ordering} = \sum_{j=0}^{s-1} \pi_j$$

$$\text{the long-run average stock on hand at the end of the week} = \sum_{j=0}^{S} j\pi_j$$

with probability 1. Using the expressions for the p_{ij} given in Section 3.1, we obtain for the π_j the equilibrium equations

$$\pi_0 = \left(1 - \sum_{\ell=0}^{S-1} e^{-\lambda}\frac{\lambda^\ell}{\ell!}\right)(\pi_0 + \cdots + \pi_{s-1}) + \sum_{k=s}^{S}\left(1 - \sum_{\ell=0}^{k-1} e^{-\lambda}\frac{\lambda^\ell}{\ell!}\right)\pi_k,$$

$$\pi_j = \sum_{k=0}^{s-1} e^{-\lambda}\frac{\lambda^{S-j}}{(S-j)!}\pi_k + \sum_{k=s}^{S} e^{-\lambda}\frac{\lambda^{k-j}}{(k-j)!}\pi_k, \quad 1 \le j \le s-1,$$

$$\pi_j = \sum_{k=0}^{s-1} e^{-\lambda}\frac{\lambda^{S-j}}{(S-j)!}\pi_k + \sum_{k=j}^{S} e^{-\lambda}\frac{\lambda^{k-j}}{(k-j)!}\pi_k, \quad s \le j \le S.$$

These equations together with the normalizing equation $\sum_{k=0}^{S}\pi_k = 1$ determine uniquely the equilibrium probabilities π_j, $j = 0, 1, \ldots, S$. If one of the equilibrium equations is omitted to obtain a square system of linear equations, the solution of the resulting system is still uniquely determined.

Example 3.1.3 (continued) The $GI/M/1$ queue

In this example the Markov chain $\{X_n\}$ describing the number of customers present just prior to arrival epochs has the infinite state space $I = \{0, 1, \ldots\}$. In order to ensure that Assumption 3.3.1 is satisfied, we have to assume that the arrival rate of customers is less than the service rate. Thus, denoting by λ the reciprocal of the mean interarrival time, it is assumed that

$$\lambda < \mu. \tag{3.3.12}$$

We omit the proof that under this condition Assumption 3.3.1 is satisfied (with state 0 as regeneration state r). In the $GI/M/1$ queueing example the equilibrium probability π_j can be interpreted as the long-run fraction of customers who see j other customers present upon arrival for $j = 0, 1, \ldots$. In particular, $1 - \pi_0$ is the long-run fraction of customers who have to wait in queue. Using the specification of the p_{ij} given in Section 3.1, we obtain the equilibrium equations

$$\pi_j = \sum_{k=j-1}^{\infty}\pi_k \int_0^\infty e^{-\mu t}\frac{(\mu t)^{k+1-j}}{(k+1-j)!}a(t)\,dt, \quad j \ge 1. \tag{3.3.13}$$

The equilibrium equation for π_0 is omitted since it is not needed. An explicit solution for the π_j can be given. This solution is

$$\pi_j = (1-\eta)\eta^j, \quad j = 0, 1, \ldots \tag{3.3.14}$$

where η is the unique solution of the equation

$$\eta - \int_0^\infty e^{-\mu(1-\eta)t}a(t)\,dt = 0 \tag{3.3.15}$$

on the interval (0, 1). Using the condition (3.3.12), it is readily verified that the equation (3.3.15) has a unique solution on (0, 1). The result (3.3.14) can be proved in several ways. A direct way is to try a solution of the form $\pi_j = \gamma \eta^j$, $j \geq 0$ for constants $\gamma > 0$ and $0 < \eta < 1$ and substituting this form into (3.3.13). By doing so, one then finds that η satisfies the equation (3.3.15). The constant γ follows from $\sum_{j=0}^{\infty} \pi_j = 1$. More sophisticated proofs of result (3.3.14) are given in Sections 3.4.2 and 3.5.2.

3.3.3 The Long-run Average Reward per Time Unit

A very useful applied probability model is the Markov chain model on which a reward or cost structure is imposed. Suppose that a reward $f(j)$ is earned each time the Markov chain visits state j for $j \in I$. The ergodic theorem shows how to compute the long-run average reward per time unit in terms of the equilibrium probabilities π_j. In addition to Assumption 3.3.1 involving the regeneration state r, we need the following assumption.

Assumption 3.3.2 *(a) The total reward earned between two visits of the Markov chain to state r has a finite expectation and* $\sum_{j \in I} |f(j)| \pi_j < \infty$.

(b) For each initial state $X_0 = i$ *with* $i \neq r$, *the total reward earned until the first visit of the Markov chain to state r is finite with probability* 1.

This assumption is automatically satisfied when the Markov chain has a finite state space and satisfies Assumption 3.3.1.

Theorem 3.3.3 *Suppose the Markov chain* $\{X_n\}$ *satisfies Assumptions 3.3.1 and 3.3.2. Then the long-run average reward per time unit is*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{j \in I} f(j) \pi_j \quad \textit{with probability } 1$$

for each initial state $X_0 = i$.

Intuitively this theorem is obvious by noting that the long-run average number of visits to state j per time unit equals π_j with probability 1 for each state $j \in I$. A formal proof of Theorem 3.3.3 is given in Section 3.5.2.

Remark 3.3.1 A useful modification of Theorem 3.3.3

In Theorem 3.3.3 the renewal function refers to an *immediate* reward $f(j)$ that is earned each time the Markov chain visits state j. However, in practical applications it happens often that rewards are *gradually* earned during the time between the state transitions of the Markov chain. Define for those situations the reward function $f(j)$ by

$$f(j) = \text{the expected reward earned until the next state transition when a state transition has just occurred to state } j.$$

Then it remains true that the long-run average reward per time unit is $\sum_{j \in I} f(j)\pi_j$ with probability 1. This can be directly seen from the proof of Theorem 3.3.3 that is given in Section 3.5.2. This proof uses the idea that the long-run average reward per time unit equals

$$\frac{E(\text{reward earned in one cycle})}{E(\text{length of one cycle})}$$

with probability 1, where a cycle is defined as the time elapsed between two successive visits to a given recurrent state. The expression for E(reward earned during one cycle) is not affected whether $f(j)$ represents an immediate reward or an expected reward.

Example 3.2.1 (continued) A stock-control problem

Suppose that the following costs are made in the stock-control problem. A fixed ordering cost of $K > 0$ is incurred each time the stock is ordered up to level S. In each week a holding cost of $h > 0$ is charged against each unit that is still in stock at the end of the week. A penalty cost of $b > 0$ is incurred for each demand that is lost. Denoting by $c(j)$ the expected costs incurred in the coming week when the current stock on hand is j just prior to review, it follows that

$$c(j) = K + h\sum_{k=0}^{S-1}(S-k)\,e^{-\lambda}\frac{\lambda^k}{k!} + b\sum_{k=S+1}^{\infty}(k-S)\,e^{-\lambda}\frac{\lambda^k}{k!}, \quad 0 \le j < s,$$

$$c(j) = h\sum_{k=0}^{j-1}(j-k)\,e^{-\lambda}\frac{\lambda^k}{k!} + b\sum_{k=j+1}^{\infty}(k-j)\,e^{-\lambda}\frac{\lambda^k}{k!}, \quad s \le j \le S.$$

The long-run average cost per week equals $\sum_{j=0}^{S} c(j)\pi_j$ with probability 1. In evaluating this expression, it is convenient to replace $\sum_{k=j+1}^{\infty}(j-k)\,e^{-\lambda}\lambda^k/k!$ by $j - \lambda - \sum_{k=0}^{j}(j-k)\,e^{-\lambda}\lambda^k/k!$ in the expression for $c(j)$. Note that by taking $b = 1$ and $K = h = 0$, the long-run average cost per week reduces to the long-run average demand lost per week. Dividing this average by the average weekly demand λ we get the long-run fraction of demand that is lost.

Example 3.3.1 An insurance problem

A transport firm has effected an insurance contract for a fleet of vehicles. The premium payment is due at the beginning of each year. There are four possible premium classes with a premium payment of P_i in class i, where $P_{i+1} < P_i$ for $i = 1, 2, 3$. The size of the premium depends on the previous premium and the claim history during the past year. If no damage is claimed in the past year and the previous premium is P_i, the next premium payment is P_{i+1} (with $P_5 = P_4$, by convention), otherwise the highest premium P_1 is due. Since the insurance

contract is for a whole fleet of vehicles, the transport firm has obtained the option to decide only at the end of the year whether the accumulated damage during that year should be claimed or not. If a claim is made, the insurance company compensates the accumulated damage minus an own risk which amounts to r_i for premium class i. The total damages in the successive years are independent random variables having a common probability distribution function $G(s)$ with density $g(s)$. What is a reasonable claim strategy and what is the long-run average cost per year?

An obvious claim strategy is the rule characterized by four parameters $\alpha_1, \ldots, \alpha_4$. If the current premium class is class i, then the transport firm claims at the end of the year only damages larger than α_i, otherwise nothing is claimed. Consider now a given claim rule $(\alpha_1, \ldots, \alpha_4)$ with $\alpha_i > r_i$ for $i = 1, \ldots, 4$. For this rule the average cost per year can be obtained by considering the stochastic process which describes the evolution of the premium class for the transport firm. Let

$$X_n = \text{the premium class for the firm at the beginning of the } n\text{th year.}$$

Then the stochastic process $\{X_n\}$ is a Markov chain with four possible states $i = 1, \ldots, 4$. The one-step transition probabilities p_{ij} are easily found. A one-step transition from state i to state 1 occurs only if at the end of the present year a damage is claimed, otherwise a transition from state i to state $i+1$ occurs (with state 5 $\equiv$ state 4). Since for premium class i only cumulative damages larger than α_i are claimed, it follows that

$$p_{i1} = 1 - G(\alpha_i), \quad i = 1, \ldots, 4,$$
$$p_{i,i+1} = G(\alpha_i), \quad i = 1, 2, 3 \quad \text{and} \quad p_{44} = G(\alpha_4).$$

The other one-step transition probabilities p_{ij} are equal to zero. The Markov chain has no two disjoint closed sets. Hence the equilibrium probabilities π_j, $1 \leq j \leq 4$, are the unique solution to the equilibrium equations

$$\pi_4 = G(\alpha_3)\pi_3 + G(\alpha_4)\pi_4,$$
$$\pi_3 = G(\alpha_2)\pi_2,$$
$$\pi_2 = G(\alpha_1)\pi_1,$$
$$\pi_1 = \{1 - G(a_1)\}\pi_1 + \{1 - G(\alpha_2)\}\pi_2 + \{1 - G(\alpha_3)\}\pi_3 + \{1 - G(\alpha_4)\}\pi_4$$

together with the normalizing equation $\pi_1+\pi_2+\pi_3+\pi_4 = 1$. These linear equations can be solved recursively. Starting with $\overline{\pi}_4 := 1$, we recursively compute $\overline{\pi}_3$, $\overline{\pi}_2$ and $\overline{\pi}_1$ from the first three equations. Next we obtain the true values of the π_j from $\pi_j := \overline{\pi}_j / \sum_{k=1}^{4} \overline{\pi}_k$. Denote by $c(j)$ the expected costs incurred during a year in which premium P_j is paid. Then by Theorem 3.3.3 we have that the long-run

Table 3.3.1 The optimal claim limits and the minimal costs

	Gamma			Lognormal		
	$c_D^2 = 1$	$c_D^2 = 4$	$c_D^2 = 25$	$c_D^2 = 1$	$c_D^2 = 4$	$c_D^2 = 25$
α_1^*	5908	6008	6280	6015	6065	6174
α_2^*	7800	7908	8236	7931	7983	8112
α_3^*	8595	8702	9007	8717	8769	8890
α_4^*	8345	8452	8757	8467	8519	8640
g^*	9058	7698	6030	9174	8318	7357

average cost per year is

$$g(\alpha_1, \ldots, \alpha_4) = \sum_{j=1}^{4} c(j)\pi_j$$

with probability 1. The one-year cost $c(j)$ consists of the premium P_j and any damages not compensated that year by the insurance company. By conditioning on the cumulative damage in the coming year, it follows that

$$c(j) = P_j + \int_0^{\alpha_j} s g(s)\, ds + r_j[1 - G(\alpha_j)].$$

The optimal claim limits follow by minimizing the function $g(\alpha_1, \ldots, \alpha_4)$ with respect to the parameters $\alpha_1, \ldots, \alpha_4$. Efficient numerical procedures are widely available to minimize a function of several variables. Table 3.3.1 gives for a number of examples the optimal claim limits $\alpha_1^*, \ldots, \alpha_4^*$ together with the minimal average cost g^*. In all examples we take

$$\begin{array}{llll} P_1 = 10\,000, & P_2 = 7500, & P_3 = 6000, & P_4 = 5000, \\ r_1 = 1500, & r_2 = 1000, & r_3 = 750, & r_4 = 500. \end{array}$$

The average damage size is 5000 in each example; the squared coefficient of variation of the damage size D takes three values: $c_D^2 = 1$, 4 and 25. To see the effect of the shape of the probability density of the damage size on the claim limits, we take the gamma distribution and the lognormal distribution both having the same first two moments. In particular, the minimal average cost becomes increasingly sensitive to the distributional form of the damage size D when c_D^2 gets larger. Can you explain why the minimal average cost per year decreases when the variability of the claims increases?

3.4 COMPUTATION OF THE EQUILIBRIUM PROBABILITIES

In this section it is assumed that the Markov chain $\{X_n\}$ satisfies Assumption 3.3.1. The Markov chain then has a unique equilibrium distribution $\{\pi_j, j \in I\}$. The π_j

are determined up to a multiplicative constant by the equilibrium equations

$$\pi_j = \sum_{k \in I} \pi_k p_{kj}, \quad j \in I. \tag{3.4.1}$$

The multiplicative constant is determined by the normalizing equation

$$\sum_{j \in I} \pi_j = 1. \tag{3.4.2}$$

In Section 3.4.1 we consider the case of a finite space I and discuss several methods to compute the equilibrium probabilities π_j. The infinite-state model is dealt with in Section 3.4.2. It is shown that brute-force truncation is not necessary to get a finite system of linear equations when the state space $I = \{0, 1, \dots\}$ and the state probabilities π_j exhibit a geometric tail behaviour as $j \to \infty$. For this situation, which naturally arises in many applications, an elegant computational method for the state probabilities can be given. Markov chains with a multidimensional state space are prevalent in stochastic networks and in such applications it often happens that the equilibrium probabilities are known up to a multiplicative constant. If the number of states is too large for a direct computation of the multiplicative constant, the Metropolis–Hastings algorithm and the Gibbs sampler may be used to obtain the equilibrium probabilities. These powerful methods are discussed in Section 3.4.3.

3.4.1 Methods for a Finite-State Markov Chain

In general there are two methods to solve the Markov chain equations:

(a) direct methods,

(b) iterative methods.

To discuss these methods, let us assume that the states of the Markov chain are numbered or renumbered as $1, \dots, N$.

Direct methods

A convenient direct method is a Gaussian elimination method such as the Gauss–Jordan method. This reliable method is recommended as long as the dimension N of the system of linear equations does not exceed the order of thousands. The computational effort of Gaussian elimination is proportional to N^3. Reliable and ready-to-use codes for Gaussian elimination methods are widely available. A Gaussian elimination method requires that the whole coefficient matrix is stored, since this matrix must be updated at each step of the algorithm. This explains why a Gaussian elimination method suffers from computer memory problems when N

gets large. In some applications the transition probabilities p_{ij} have the property that for each state i the probability $p_{ij} = 0$ for $j \leq i-2$ (or $p_{ij} = 0$ for $j \geq i+2$). Then the linear equations are of the Hessenberg type. Linear equations of the Hessenberg type can be efficiently solved by a special code using the very stable QR method. In solving the Markov chain equations (3.4.1) and (3.4.2) by a direct method, one of the equilibrium equations is omitted to obtain a square system of linear equations.

Iterative method of successive overrelaxation

Iterative methods have to be used when the size of the system of linear equations gets large. In specific applications an iterative method can usually avoid computer memory problems by exploiting the (sparse) structure of the application. An iterative method does not update the matrix of coefficients each time. In applications these coefficients are usually composed from a few constants. Then only these constants have to be stored in memory when using an iterative method. In addition to the advantage that the coefficient matrix need not be stored, an iterative method is easy to program for specific applications.

The iterative method of successive overrelaxation is a suitable method for solving the linear equations of large Markov chains. The well-known Gauss–Seidel method is a special case of the method of successive overrelaxation. The iterative methods generate a sequence of vectors $\mathbf{x}^{(0)} \rightarrow \mathbf{x}^{(1)} \rightarrow \mathbf{x}^{(2)} \rightarrow \ldots$ converging towards a solution of the equilibrium equations (3.4.1). The normalization is done at the end of the calculations. To apply successive overrelaxation, we first rewrite the equilibrium equations (3.4.1) in the form

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} x_j, \quad i = 1, \ldots, N,$$

where

$$a_{ij} = \frac{p_{ji}}{1 - p_{ii}}, \quad i, j = 1, \ldots, N, \; j \neq i.$$

The standard successive overrelaxation method uses a fixed relaxation factor ω for speeding up the convergence. The method starts with an initial approximation vector $\mathbf{x}^{(0)} \neq 0$. In the kth iteration of the algorithm an approximation vector $\mathbf{x}^{(k)}$ is found by a recursive computation of the components $x_i^{(k)}$ such that the calculation of the new estimate $x_i^{(k)}$ uses both the new estimates $x_j^{(k)}$ for $j < i$ and the old estimates $x_j^{(k-1)}$ for $j > i$. The steps of the algorithm are as follows:

Step 0. Choose a non-zero vector $\mathbf{x}^{(0)}$. Let $k := 1$.

Step 1. Calculate successively for $i = 1, \ldots, N$ the component $x_i^{(k)}$ from

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \omega\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} + \sum_{j=i+1}^{N} a_{ij}x_j^{(k-1)}\right).$$

Step 2. If the stopping criterion

$$\sum_{i=1}^{N}\left|x_i^{(k)} - x_i^{(k-1)}\right| \leq \varepsilon \sum_{i=1}^{N}\left|x_i^{(k)}\right|$$

is satisfied with $\varepsilon > 0$, a prespecified accuracy number, then go to step 3. Otherwise $k := k+1$ and go to step 1.
Step 3. Calculate the solution to (3.4.1) and (3.4.2) from

$$x_i^* = \frac{x_i^{(k)}}{\sum_{j=1}^{N} x_j^{(k)}}, \quad 1 \leq i \leq N.$$

The specification of the tolerance number ε typically depends on the particular problem considered and the accuracy required in the final answers. In addition to the stopping criterion, it may be helpful to use an extra accuracy check for the equilibrium probabilities of the underlying Markov chain. An extra accuracy check may prevent a decision upon a premature termination of the algorithm when the tolerance number ε is not chosen sufficiently small. Notice that the normalizing equation (3.4.2) is used only at the very end of the algorithm. In applying successive overrelaxation it is highly recommended that all of the equilibrium equations (3.4.1) are used rather than omitting one redundant equation and substituting the normalizing equation (3.4.2) for it.

The convergence speed of the successive overrelaxation method may dramatically depend on the choice of the relaxation factor ω, and even worse the method may diverge for some choices of ω. A suitable value of ω has to be determined experimentally. Usually $1 \leq \omega \leq 2$. The choice $\omega = 1.2$ is often recommended. The optimal value of the relaxation factor ω depends on the structure of the particular problem considered. It is pointed out that the iteration method with $\omega = 1$ is the well-known *Gauss–Seidel method*. This method is convergent in all practical cases. The ordering of the states may also have a considerable effect on the convergence speed of the successive overrelaxation algorithm. In general one should order the states such that the upper diagonal part of the matrix of coefficients is as sparse as possible. In specific applications the transition structure of the Markov chain often suggests an appropriate ordering of the states.

Krylov iteration method

The Gauss–Seidel iteration method can further be refined to obtain orthogonal basis vectors for a so-called Krylov space. The construction of an appropriate Krylov

basis is strongly dependent of the structure of the system of linear equations to be solved and is typically a matter of experimentation. However, it is worthwhile to try such an experimentation when an extremely large but structured system of linear equations has to be solved many times. Enormous reductions in computing times can be achieved by Krylov iteration methods; see Stewart (1994).

Recursive method

The linear equations (3.4.1) and (3.4.2) become a Hessenberg system when the p_{ij} have the property that for each state $i = 1, \ldots, N$,

$$p_{ij} = 0 \quad \text{for all } j \leq i - 2. \tag{3.4.3}$$

In this special case the equilibrium probabilities π_j can also be computed by a simple recursion scheme. To obtain this recursion scheme, we extend the 'rate out = rate in' principle discussed in Section 3.3. For each set A of states with $A \neq I$, we have that the long-run average number of transitions per time unit from a state inside A to a state outside A equals the long-run average number of transitions per time unit from a state outside A to a state inside A.

Under the property (3.4.3) the set $A = \{i, i+1, \ldots, N\}$ with $i \neq 1$ can be left only through state i. Applying the 'rate out = rate in' principle to this set A, we find

$$p_{i,i-1}\pi_i = \sum_{k=1}^{i-1} \pi_k \left(\sum_{j=i}^{N} p_{kj} \right), \quad i = 2, \ldots, N. \tag{3.4.4}$$

This recursion starts with the value of π_1. Since the equilibrium equations determine the probabilities π_j up to a multiplicative constant, it is no problem that the value of π_1 is not known beforehand. We initialize the recursion with an arbitrary non-zero value for π_1 and normalize at the end of the recursion. In applying (3.4.4) it is no restriction to assume that $p_{i,i-1} > 0$ for all $i \geq 2$.

Algorithm

Step 0. Initialize $\overline{\pi}_1 := 1$.
Step 1. Compute successively $\overline{\pi}_2, \ldots, \overline{\pi}_N$ from (3.4.4).
Step 2. Normalize the π_i according to

$$\pi_i = \overline{\pi}_i / \sum_{k=1}^{N} \overline{\pi}_k, \quad i = 1, 2, \ldots, N.$$

The recursion scheme (3.4.4) involves no subtractions and is thus numerically stable. However, very large numbers $\overline{\pi}_i$ may build up when N is large. In those situations it is recommended to do a renormalization at intermediate steps of the recursion. The recursion method can also be used for a Markov chain with an

infinite state space $I = \{1, 2, \dots\}$ and one-step transition probabilities p_{ij} satisfying (3.4.3). Then a truncation integer N must be used.

3.4.2 Geometric Tail Approach for an Infinite State Space

Many applications of Markov chains involve an *infinite* state space. What one usually does to solve numerically the infinite set of equilibrium equations is to approximate the infinite-state Markov model by a truncated model with finitely many states so that the probability mass of the deleted states is very small. Indeed, for a finite-state truncation with a sufficiently large number of states, the difference between the two models will be negligible from a computational point of view. However, such a truncation often leads to a finite but very large system of linear equations whose numerical solution will be quite time-consuming, although an arsenal of good methods is available to solve the equilibrium equations of a finite Markov chain. Moreover, it is somewhat disconcerting that we need a brute-force approximation to solve the infinite-state model numerically. Usually we introduce infinite-state models to obtain mathematical simplification, and now in its numerical analysis using a brute-force truncation we are proceeding in the reverse direction. Fortunately, many applications allow for a much simpler and more satisfactory approach to solving the infinite set of state equations. Under rather general conditions the state probabilities exhibit a *geometric tail behaviour* that can be exploited to reduce the infinite system of state equations to a finite set of linear equations. The geometric tail approach results in a finite system of linear equations whose size is usually much smaller than the size of the finite system obtained from a brute-force truncation. It is a robust approach that is easy to use by practitioners.

Consider a discrete-time Markov chain whose state space is one-dimensional and is given by

$$I = \{0, 1, \dots\}.$$

Let us assume that the equilibrium probabilities π_j, $j \in I$, exhibit the geometric tail behaviour

$$\pi_j \sim \gamma \eta^j \quad \text{as } j \to \infty \tag{3.4.5}$$

for some constants $\gamma > 0$ and $0 < \eta < 1$. Here $f(x) \sim g(x)$ as $x \to \infty$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Below we will discuss conditions under which (3.4.5) holds. First we demonstrate how the geometric tail behaviour can be exploited to reduce the infinite system of state equations to a finite system of linear equations. It will be seen below that the decay factor η in (3.4.5) can usually be computed beforehand by solving a non-linear equation in a single variable. Solving a non-linear equation in a single variable is standard fare in numerical analysis. In most applications it is not possible to compute the constant γ beforehand. Fortunately, we do not need the constant γ in our approach. The asymptotic expansion is only used by

$$\lim_{j\to\infty} \frac{\pi_j}{\pi_{j-1}} = \eta.$$

In other words, for a sufficiently large integer M,

$$\pi_j \approx \pi_M \eta^{j-M}, \quad j \geq M.$$

Replacing π_j by $\pi_M \eta^{j-M}$ for $j \geq M$ in equations (3.4.1) and (3.4.2) leads to the following finite set of linear equations:

$$\pi_j = \sum_{k=0}^{M} a_{jk}\pi_k, \quad j = 0, 1, \ldots, M-1,$$

$$\sum_{j=0}^{M-1} \pi_j + \frac{\pi_M}{1-\eta} = 1,$$

where for any $j = 0, 1, \ldots, M-1$ the coefficients a_{jk} are given by

$$a_{jk} = \begin{cases} p_{kj}, & k = 0, 1, \ldots, M-1, \\ \sum_{i=M}^{\infty} \eta^{i-M} p_{ij}, & k = M. \end{cases}$$

How large an M should be chosen has to be determined experimentally and depends, of course, on the required accuracy in the calculated values of the equilibrium probabilities. However, empirical investigations show that in specific applications *remarkably small* values of M are already good enough for practical purposes. We found in all practical examples that the system of linear equations is non-singular, irrespective of the value chosen for M. An appropriate value of M is often in the range 1–200 when a reasonable accuracy (perhaps seven-digit accuracy) is required for the equilibrium probabilities. A Gaussian elimination method is a convenient method for solving linear equations of this size. Fast and reliable codes for Gaussian elimination are widely available. The geometric tail approach combines effectivity with simplicity.

Conditions for the geometric tail behaviour

A useful but technical condition for (3.4.5) to hold can be given in terms of the generating function $\sum_{j=0}^{\infty} \pi_j z^j$ of the equilibrium probabilities π_j. In many applications the following condition is satisfied.

Condition A *(a) The generating function* $\sum_{j=0}^{\infty} \pi_j z^j$ *for* $|z| \leq 1$ *has the form*

$$\sum_{j=0}^{\infty} \pi_j z^j = \frac{N(z)}{D(z)}, \tag{3.4.6}$$

where $N(z)$ and $D(z)$ are functions that have no common zeros. The functions $N(z)$ and $D(z)$ are analytic functions that can be analytically continued outside the unit circle $|z| \leq 1$.

(b) Letting $R > 1$ be the largest number such that both functions $N(z)$ and $D(z)$ are analytic in the region $|z| < R$ in the complex plane, the equation

$$D(x) = 0 \tag{3.4.7}$$

has a smallest root x_0 on the interval $(1, R)$.

In specific applications the denominator $D(z)$ in (3.4.6) is usually a nice function that is *explicitly* given (this is usually not true for the numerator $N(z)$). It is only the denominator $D(z)$ that is needed for our purposes. Theorem C.1 in Appendix C shows that under Condition A plus some secondary technical conditions the state probabilities π_j allow for the asymptotic expansion (3.4.5) with

$$\eta = \frac{1}{x_0}. \tag{3.4.8}$$

Condition A is a condition that seems not to have a probabilistic interpretation. Next we give a probabilistic condition for (3.4.5) to hold. This condition is in terms of the one-step transition probabilities p_{ij} of the Markov chain.

Condition B *(a) There is an integer $r \geq 0$ such that p_{ij} depends on i and j only through $j - i$ when $i \geq r$ and $j \geq 1$.*

(b) There is an integer $s \geq 1$ such that

$$p_{ij} = 0 \quad \text{for } j > i + s \text{ and } i \geq 0.$$

(c) Letting α_{j-i} denote p_{ij} for $i \geq r$ and $1 \leq j \leq i + s$, the constants α_k satisfy

$$\alpha_s > 0 \quad \textit{and} \quad \sum_{k=-\infty}^{s} k\alpha_k < 0.$$

Under Condition B the equilibrium equation for π_j has the form

$$\pi_j = \sum_{k=j-s}^{\infty} \alpha_{j-k}\pi_k \quad \text{for } j \geq r + s.$$

This is a homogeneous linear difference equation with constant coefficients. A standard method to solve such a linear difference equation is the method of particular solutions. Substituting a solution of the form $\pi_j = w^j$ in the equilibrium equations for the π_j with $j \geq r + s$, we find the so-called characteristic equation

$$w^s - \sum_{\ell=0}^{\infty} \alpha_{s-\ell} w^\ell = 0. \tag{3.4.9}$$

This equation can be shown to have s roots in the interior of the unit circle $|w| \leq 1$. Assume now that the roots $w_1, \ldots, w_s$ are distinct (as is typically the case in applications). Then, by a standard result from the theory of linear difference equations, there are constants $c_1, \ldots, c_s$ such that

$$\pi_j = \sum_{k=1}^{s} c_k w_k^j \quad j \geq r. \tag{3.4.10}$$

The root w_k having the largest modulus must be real and positive. Why? Denoting this root by η, the asymptotic expansion (3.4.5) then follows.

Example 3.1.3 (continued) The $GI/M/1$ queue

The Markov chain $\{X_n\}$ describing the number of customers present just prior to arrival epochs satisfies Condition B with

$$r = 0 \quad \text{and} \quad s = 1,$$

as directly follows from the one-step transition probabilities p_{ij} given in (3.1.2). The constants α_k are given by

$$\alpha_k = \int_0^\infty e^{-\mu t} \frac{(\mu t)^{1-k}}{(1-k)!} a(t)\, dt, \quad k \leq 1.$$

It is directly verified that $\alpha_1 > 0$ and $\sum_{k=-\infty}^{1} k\alpha_k = 1 - \mu/\lambda < 0$. Thus we can directly conclude from (3.4.10) that the equilibrium probabilities π_j are of the form $\gamma\eta^j$ for all $j \geq 0$ for constants $\gamma > 0$ and $0 < \eta < 1$. The characteristic equation (3.4.9) coincides with the equation (3.3.15).

Next we give an application in which Condition A is used to establish the asymptotic expansion (3.4.5).

Example 3.4.1 A discrete-time queueing model

Messages arrive at a communication system according to a Poisson process with rate λ. The messages are temporarily stored in a buffer which is assumed to have infinite capacity. There are c transmission channels. At fixed clock times $t = 0, 1, \ldots$ messages are taken out of the buffer and are synchronously transmitted. Each channel can only transmit one message at a time. The transmission time of a message is one time slot. Transmission of messages can only start at the clock times $t = 0, 1, \ldots$. It is assumed that

$$\lambda < c,$$

that is, the arrival rate of messages is less than the transmission capacity.

To analyse this queueing model, define the random variable X_n by

$$X_n = \text{the number of messages in the buffer (excluding any message in transmission) just prior to clock time } t = n.$$

Then $\{X_n, n = 0, 1, \dots\}$ is a discrete-time Markov chain with the infinite state space $I = \{0, 1, \dots\}$. The one-step transition probabilities are given by

$$p_{ij} = e^{-\lambda}\frac{\lambda^j}{j!}, \quad 0 \le i < c \text{ and } j = 0, 1, \dots$$

$$p_{ij} = e^{-\lambda}\frac{\lambda^{j-i+c}}{(j-i+c)!}, \quad i \ge c \text{ and } j = i-c, i-c+1, \dots .$$

By the assumption $\lambda < c$ the Markov chain can be shown to satisfy Assumption 3.3.1. Hence the equilibrium probabilities π_j, $j = 0, 1, \dots$ exist and are the unique solution to the equilibrium equations

$$\pi_j = e^{-\lambda}\frac{\lambda^j}{j!}\sum_{k=0}^{c-1}\pi_k + \sum_{k=c}^{c+j} e^{-\lambda}\frac{\lambda^{j-k+c}}{(j-k+c)!}\pi_k, \quad j = 0, 1, \dots$$

in conjunction with the normalizing equation $\sum_{j=0}^{\infty}\pi_j = 1$. Multiplying both sides of the equilibrium equation for π_j by z^j and summing over j, we find

$$\begin{aligned}\sum_{j=0}^{\infty}\pi_j z^j &= \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!}z^j\sum_{k=0}^{c-1}\pi_k + \sum_{j=0}^{\infty} z^j\sum_{k=c}^{c+j} e^{-\lambda}\frac{\lambda^{j-k+c}}{(j-k+c)!}\pi_k \\ &= e^{-\lambda(1-z)}\sum_{k=0}^{c-1}\pi_k + \sum_{k=c}^{\infty}\pi_k z^{k-c}\sum_{j=k-c}^{\infty} e^{-\lambda}\frac{\lambda^{j-k+c}}{(j-k+c)!}z^{j-k+c} \\ &= e^{-\lambda(1-z)}\left[\sum_{k=0}^{c-1}\pi_k + z^{-c}\left(\sum_{k=0}^{\infty}\pi_k z^k - \sum_{k=0}^{c-1}\pi_k z^k\right)\right].\end{aligned}$$

This gives

$$\sum_{j=0}^{\infty}\pi_j z^j = \frac{e^{-\lambda(1-z)}\left[\sum_{k=0}^{c-1}\left(z^c - z^k\right)\pi_k\right]}{z^c - e^{-\lambda(1-z)}}, \quad |z| \le 1.$$

The generating function $\sum_{j=0}^{\infty}\pi_j z^j$ is the ratio of two functions $N(z)$ and $D(z)$. Both functions can be analytically continued to the whole complex plane. The denominator $D(z)$ is indeed a nice function in an explicit form (the function $N(z)$ involves the unknowns $\pi_0, \dots, \pi_{c-1}$). Denote by x_0 the unique solution of the

equation

$$x^c - e^{-\lambda(1-x)} = 0$$

on the interval $(1, \infty)$ and let $\eta = 1/x_0$. Then it can be verified from Theorem C.1 in Appendix C that

$$\pi_j \sim \gamma \eta^j \quad \text{as } j \to \infty$$

for some constant $\gamma > 0$. Thus the geometric approach enables us to compute the π_j by solving a finite and relatively small system of linear equations.

3.4.3 Metropolis—Hastings Algorithm

In the context of stochastic networks, we will encounter in Chapter 5 Markov chains with a multidimensional state space and having the feature that the equilibrium probabilities are known up to a multiplicative constant. However, the number of possible states is enormous so that a direct calculation of the normalization constant is not practically feasible. This raises the following question. Suppose that $\overline{\pi}_1, \ldots, \overline{\pi}_N$ are given positive numbers with a finite sum $S = \sum_{i=1}^N \overline{\pi}_i$. How do we construct a Markov chain whose equilibrium probabilities are given by $\overline{\pi}_j/S$ for $j = 1, \ldots, N$? For ease of presentation, we restrict ourselves to $N < \infty$. To answer the question, we need the concept of a reversible Markov chain. Let $\{X_n\}$ be a Markov chain with a finite state space I and one-step transition probabilities p_{ij}. It is assumed that $\{X_n\}$ has no two disjoint closed sets. Then the Markov chain has a unique equilibrium distribution $\{\pi_j\}$. Assume now that a non-null vector (g_j), $j \in I$ exists such that

$$g_j p_{jk} = g_k p_{kj}, \quad j, k \in I. \tag{3.4.11}$$

Then, for some constant $c \neq 0$,

$$g_j = c\pi_j. \tag{3.4.12}$$

The proof is simple. Fix $j \in I$ and sum both sides of (3.4.11) over k. This gives

$$g_j = \sum_{k \in I} g_k p_{kj}, \quad j \in I.$$

These equations are exactly the equilibrium equations of the Markov chain $\{X_n\}$. Hence, by Theorem 3.3.2, we have that (3.4.12) holds. By (3.4.11) and (3.4.12),

$$\pi_j p_{jk} = \pi_k p_{kj}, \quad j, k \in I. \tag{3.4.13}$$

A Markov chain $\{X_n\}$ having this property is called a *reversible* Markov chain. The property (3.4.13) states that the long-run average number of transitions from state j to state k per time unit is equal to the long-run average number of transitions from state k to state j per time unit for all $j, k \in I$.

Let us return to the problem of constructing a Markov chain with equilibrium probabilities $\{\pi_j = \overline{\pi}_j/S,\ j = 1, \ldots, N\}$ when $\overline{\pi}_1, \ldots, \overline{\pi}_N$ are given positive numbers with a finite sum S. To do so, choose any Markov matrix $M = (m_{ij})$, $i, j = 1, \ldots, N$ with positive elements m_{ij}. Next construct a Markov chain $\{X_n\}$ with state space $I = \{1, \ldots, N\}$ and one-step transition probabilities

$$p_{ij} = \begin{cases} m_{ij}\alpha_{ij}, & j \neq i, \\ m_{ii}\alpha_{ii} + \sum_{k=1}^{N} m_{ik}(1 - \alpha_{ik}), & j = i, \end{cases}$$

where the α_{ij} are appropriately chosen numbers between 0 and 1 with $\alpha_{ii} = 1$ for $i = 1, \ldots, N$. The state transitions of the Markov chain $\{X_n\}$ are governed by the following rule: if the current state of the Markov chain $\{X_n\}$ is i, then a candidate state k is generated according to the probability distribution $\{m_{ij}, j = 1, \ldots, N\}$. The next state of the Markov chain $\{X_n\}$ is chosen equal to the candidate state k with probability α_{ik} and is chosen equal to the current state i with probability $1 - \alpha_{ik}$. By an appropriate choice of the α_{ij}, we have

$$\overline{\pi}_j p_{jk} = \overline{\pi}_k p_{kj}, \quad j, k = 1, \ldots, N, \tag{3.4.14}$$

implying that the Markov chain $\{X_n\}$ has the equilibrium distribution

$$\pi_j = \overline{\pi}_j / \sum_{k=1}^{N} \overline{\pi}_k, \quad j = 1, \ldots, N. \tag{3.4.15}$$

It is left to the reader to verify that (3.4.14) holds for the choice

$$\alpha_{ij} = \min\left(\frac{\overline{\pi}_j m_{ji}}{\overline{\pi}_i m_{ij}}, 1\right), \quad i, j = 1, \ldots, N \tag{3.4.16}$$

(use that $\alpha_{ji} = 1$ if $\alpha_{ij} = \overline{\pi}_j m_{ji}/\overline{\pi}_i m_{ij}$). Note that the sum $S = \sum_{k=1}^{N} \overline{\pi}_k$ is not needed to define the Markov chain $\{X_n\}$.

Summarizing, the following algorithm generates a sequence of successive states of a Markov chain $\{X_n\}$ whose equilibrium distribution is given by (3.4.15).

Metropolis—Hastings algorithm

Step 0. Choose a Markov matrix $M = (m_{ij})$, $i, j = 1, \ldots, N$ with positive elements. Let $X_0 := i$ for some $1 \leq i \leq N$ and let $n := 0$.
Step 1. Generate a candidate state Y from the probability distribution $P\{Y = j\} = m_{X_n, j}$ for $j = 1, \ldots, N$. If $Y = k$, then set X_{n+1} equal to k with probability $\alpha_{X_n, k}$ and equal to X_n with probability $1 - \alpha_{X_n, k}$, where the α_{ij} are given by (3.4.16).
Step 2. $n := n + 1$ and repeat step 1.

For the generated sequence of successive states $X_0, X_1, \ldots$, it holds that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} f(X_k) = \sum_{j=1}^{N} f(j)\pi_j \quad \text{with probability 1}$$

for any given function f. Thus the Metropolis—Hastings algorithm can be used to find performance measures of the Markov chain $\{X_n\}$ such as the long-run average cost per time unit when a cost structure is imposed on the Markov chain.

The most widely used version of the Metropolis—Hastings algorithm is the Gibbs sampler. Suppose that $(N_1, \ldots, N_d)$ is a d-dimensional stochastic vector whose probability distribution

$$p(x_1, \ldots, x_d) = P\{N_1 = x_1, \ldots, N_d = x_d\}$$

is known up to a multiplicative constant. This situation will be encountered in Section 5.6 in the context of a closed queueing network. In this particular application the univariate conditional distribution

$$P\{N_k = x_k | N_j = x_j \text{ for } j = 1, \ldots, d \text{ with } j \neq k\} \tag{3.4.17}$$

is explicitly known for each $k = 1, \ldots, d$. In order to apply the Gibbs sampler, it is required that the univariate conditional distributions in (3.4.17) are known. The Gibbs sampler generates a sequence of successive states $(x_1, \ldots, x_d)$ from a Markov chain whose equilibrium distribution is given by $p(x_1, \ldots, x_d)$.

Gibbs sampler

Step 0. Choose an initial state $\mathbf{x} = (x_1, \ldots, x_d)$.
Step 1. For the current state $\mathbf{x}$ choose a coordinate which is equally likely to be any of the coordinates $1, \ldots, d$. If coordinate k is chosen, then generate a random variable Y whose probability distribution is given by

$$P\{Y = y\} = P\{X_k = y | X_j = x_j \text{ for } j = 1, \ldots, d \text{ with } j \neq k\}.$$

If $Y = y$, let the candidate state $\mathbf{y} = (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_d)$.
Step 2. The next state $\mathbf{x} = (x_1, \ldots, x_d)$ is set equal to $\mathbf{y}$. Repeat step 1 with this new state $\mathbf{x}$.

The Gibbs sampler uses the Metropolis—Hastings algorithm with the choice

$$m_{\mathbf{x},\mathbf{y}} = \frac{1}{d} P\{X_k = y | X_j = x_j \text{ for } j = 1, \ldots, d \text{ with } j \neq k\}$$

for the Markov matrix M. It is not difficult to verify that for this choice the acceptance probability $\alpha_{\mathbf{x},\mathbf{y}}$ is given by

$$\alpha_{\mathbf{x},\mathbf{y}} = \min\left(\frac{p(\mathbf{y})p(\mathbf{x})}{p(\mathbf{x})p(\mathbf{y})}, 1\right) = 1.$$

Hence the candidate state is always accepted as the next state of the Markov chain.

3.5 THEORETICAL CONSIDERATIONS

In this section we give some background material. First the state classification of Markov chains is discussed. Next we prove the results that were used earlier in the analysis of the long-run behaviour of Markov chains.

3.5.1 State Classification

The concepts of a transient state and a recurrent state were introduced in Section 3.2 and the following lemma was proved for the Markov chain $\{X_n\}$.

Lemma 3.5.1 *A state i is transient only if $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$ and a state i is recurrent only if $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$.*

To analyse the transient states and recurrent states in more detail, we need the concept of accessibility.

Definition 3.5.1 *State j is said to be accessible from state i if $p_{ij}^{(n)} > 0$ for some $n \geq 0$. Two states i and j are said to communicate if j is accessible from i and i is accessible from j.*

Since $p_{ii}^{(0)} = 1$ by definition, we always have that any state i is accessible from itself. It is convenient to write $i \to j$ if state j is accessible from state i. The concept of communication enables us to split up the state space in a natural way into disjoint closed sets of recurrent states and a set of transient states (for the finite-state Markov chain an algorithm is given at the end of this subsection). Recall that a non-empty set C of states is called a closed set if $p_{ij} = 0$ for $i \in C$ and $j \notin C$. That is, the Markov chain cannot leave the set C once it is in the set C. By definition the state space I is always a closed set. A closed set C is called *irreducible* when the set C contains no smaller closed set.

Lemma 3.5.2 *Let C be a closed set of states. The set C is irreducible if and only if all states in C communicate with each other.*

Proof For each $i \in C$, define the set $S(i)$ by

$$S(i) = \{j \mid i \to j\}.$$

The set $S(i)$ is not empty since $i \to i$. Since the set C is closed, we have $S(i) \subseteq C$. First suppose that C is irreducible. The 'only if' part of the lemma then follows by showing that $S(i) = C$ for all i. To do so, it suffices to show that $S(i)$ is closed. Assume now to the contrary that $S(i)$ is not closed. Then there is a state $r \in S(i)$ and a state $s \notin S(i)$ with $p_{rs} > 0$. Since $r \in S(i)$ we have $p_{ir}^{(n)} > 0$ for some $n \geq 0$ and so $p_{is}^{(n+1)} \geq p_{ir}^{(n)} p_{rs} > 0$; use relation (3.2.2). The inequality $p_{is}^{(n+1)} > 0$ contradicts the fact that $s \notin S(i)$. This completes the proof of the 'only if' part of

the lemma. To prove the other part, assume to the contrary that C is not irreducible. Then there is a closed set $S \subseteq C$ with $S \neq C$. Choose $i \in S$ and let the set $S(i)$ be as above. Since S is closed, we have $S(i) \subseteq S$. Hence $S(i) \neq C$, which contradicts the assumption that all states in C communicate.

We are now able to prove the following interesting theorem.

Theorem 3.5.3 *(a) Let C be an irreducible set of states. Then either all states in C are recurrent or all states in C are transient.*

(b) Let C be an irreducible set consisting of recurrent states. Then $f_{ij} = 1$ for all $i, j \in C$. Moreover, either $\mu_{jj} < \infty$ for all $j \in C$ or $\mu_{jj} = \infty$ for all $j \in C$.

Proof (a) By Lemma 3.5.1, state i is transient if and only if $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$. Choose now $i, j \in C$ with $j \neq i$. By Lemma 3.5.2 we have that the states i and j communicate. Hence there are integers $v \geq 1$ and $w \geq 1$ such that $p_{ij}^{(v)} > 0$ and $p_{ji}^{(w)} > 0$. Next observe that for any $n \geq 0$,

$$p_{ii}^{(n+v+w)} \geq p_{ij}^{(v)} p_{jj}^{(n)} p_{ji}^{(w)} \quad \text{and} \quad p_{jj}^{(n+v+w)} \geq p_{ji}^{(w)} p_{ii}^{(n)} p_{ij}^{(v)}. \tag{3.5.1}$$

These inequalities imply that $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ if and only if $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$. This proves part (a). In fact the proof shows that $i \to j$ and $j \to i$ implies that both states i and j are recurrent or that both states i and j are transient.

(b) Since the states of C are recurrent, we have by definition that $f_{ii} = 1$ for all $i \in C$. Choose now $i, j \in C$ with $j \neq i$. By Lemma 3.5.2 $j \to i$. Hence there is an integer $m \geq 1$ with $p_{ji}^{(m)} > 0$. Let r be the smallest integer $m \geq 1$ for which $p_{ji}^{(m)} > 0$. Then

$$1 - f_{jj} = P\{X_n \neq j \text{ for all } n \geq 1 \mid X_0 = j\} \geq p_{ji}^{(r)}(1 - f_{ij}).$$

Since $f_{jj} = 1$, we get from this inequality that $f_{ij} = 1$. The inequalities in (3.5.1) imply that the sequence $\{p_{ii}^{(k)}, k \geq 1\}$ has a positive Cesaro limit if and only if the sequence $\{p_{jj}^{(k)}, k \geq 1\}$ has a positive Cesaro limit. It now follows from (3.3.1) in Theorem 3.3.1 that $\mu_{jj} < \infty$ if and only if $\mu_{ii} < \infty$.

Theorem 3.5.4 *Let R be the set of recurrent states of the Markov chain. Suppose that the set R is not empty. Then*

(a) the set R is a closed set,

(b) the set R can be uniquely split into disjoint irreducible subsets $R_1, R_2, \ldots$ (called recurrent subclasses).

Proof (a) Choose any state $r \in R$. Let s be any state such that $p_{rs} > 0$. The set R is closed if we can show that $s \in R$. Since state r is recurrent and state s is accessible from state r, state r must also be accessible from state s. If not, there

would be a positive probability of never returning to state r, contradicting the fact that state r is recurrent. Hence there is a positive integer v such that $p_{sr}^{(v)} > 0$. For any integer k,

$$p_{ss}^{(v+k+1)} \geq p_{sr}^{(v)} p_{rr}^{(k)} p_{rs},$$

implying that $\sum_{n=1}^{\infty} p_{ss}^{(n)} \geq p_{sr}^{(v)} p_{rs} \sum_{k=1}^{\infty} p_{rr}^{(k)}$. Since state r is recurrent, it now follows from Lemma 3.5.1 that state s is recurrent. Hence $s \in R$.

(b) We first observe that the following two properties hold:

(P1) If state i communicates with state j and state i communicates with state k, then the states j and k communicate.

(P2) If state j is recurrent and state k is accessible from state j, then state j is accessible from state k.

The first property is obvious. The second property was in fact proved in part (a). Define now for each $i \in R$ the set $C(i)$ as the set of all states j that communicate with state i. The set $C(i)$ is not empty since i communicates with itself by definition. Further, by part (a), $C(i) \subseteq R$. To prove that the set $C(i)$ is closed, let $j \in C(i)$ and let k be any state with $p_{jk} > 0$. Then we must verify that $i \to k$ and $k \to i$. From $i \to j$ and $j \to k$ it follows that $i \to k$. Since $j \to i$, the relation $k \to i$ follows when we can verify that $k \to j$. The relation $k \to j$ follows directly from property P2, since j is recurrent by the proof of part (a) of Theorem 3.5.3. Moreover, the foregoing arguments show that any two states in $C(i)$ communicate. It now follows from Lemma 3.5.2 that $C(i)$ is an irreducible set. Also, using the properties P1 and P2, it is readily verified that $C(i) = C(j)$ if i and j communicate and that $C(i) \cap C(j)$ is empty otherwise. This completes the proof of part (b).

Definition 3.5.2 *Let i be a recurrent state. The period of state i is said to be d if d is the greatest common divisor of the indices $n \geq 1$ for which $p_{ii}^{(n)} > 0$. A state i with period $d = 1$ is said to be aperiodic.*

Lemma 3.5.5 *(a) Let C be an irreducible set consisting of recurrent states. Then all states in C have the same period.*

(b) If state i is aperiodic, then there is an integer n_0 such that $p_{ii}^{(n)} > 0$ for all $n \geq n_0$.

Proof (a) Denote by $d(k)$ the period of state $k \in C$. Choose $i, j \in C$ with $j \neq i$. By Lemma 3.5.2 we have $i \to j$ and $j \to i$. Hence there are integers $v, w \geq 1$ such that $p_{ij}^{(v)} > 0$ and $p_{ji}^{(w)} > 0$. Let n be any positive integer with $p_{jj}^{(n)} > 0$. Then the first inequality in (3.5.1) implies that $p_{ii}^{(n+v+w)} > 0$ and so $n+v+w$ is divisible by $d(i)$. Thus we find that n is divisible by $d(i)$ whenever $p_{jj}^{(n)} > 0$. This implies that $d(i) \leq d(j)$. For reasons of symmetry, $d(j) \leq d(i)$. Hence $d(i) = d(j)$ which verifies part (a).

(b) Let $A = \{n \geq 1 \mid p_{ii}^{(n)} > 0\}$. The index set A is closed in the sense that $n + m \in A$ when $n \in A$ and $m \in A$. This follows from $p_{ii}^{(n+m)} \geq p_{ii}^{(n)} p_{ii}^{(m)}$. Since

state i is aperiodic, there are integers $a \in A$ and $b \in A$ whose greatest common divisor is equal to 1. An elementary result in number theory states that there exist integers r and s such that gcd $(a, b) = ar + bs$. The integers r and s are not necessarily non-negative. Let p and q be any positive integers such that both p and q are larger than $a \times \max(|r|, |s|)$. Take $m = pa + qb$. Since $m + a = (p+1)a + qb$, part (b) of the lemma follows by proving that $m + k \in A$ for $k = 0, \ldots, a - 1$. We then have $p_{ii}^{(n)} > 0$ for all $n \geq m$. Noting that $ar + bs = 1$, it follows that $m + k = pa + qb + k(ar + bs) = (p + kr)a + (q + ks)b$. The integers $p + kr$ and $q + ks$ are positive. Hence, by the closedness of A, the integers $(p + kr)a$ and $(q + ks)b$ belong to A and so the integer $m + k \in A$ for any $k = 0, \ldots, a - 1$.

Finite state space

There are a number of basic results that hold for finite-state Markov chains but not for Markov chains with infinitely many states. In an infinite-state Markov chain it may happen that there is no recurrent state, as is demonstrated by the Markov chain example with state space $I = \{1, 2, \ldots\}$ and one-step transition probabilities with $p_{i,i+1} = 1$ for all $i \geq 1$. In this example all states are transient. The next lemma shows that a finite-state Markov chain always has recurrent states.

Lemma 3.5.6 *Each finite closed set of states has at least one recurrent state.*

Proof Let C be a closed set of states. Then, for any $i \in C$,

$$\sum_{j \in C} p_{ij}^{(n)} = 1, \quad n = 1, 2, \ldots . \tag{3.5.2}$$

Assume now that all states $j \in C$ are transient. In Lemma 3.2.3 it was shown that $\lim_{n \to \infty} p_{ij}^{(n)} = 0$ for all $i \in I$ if state j is transient. Let $n \to \infty$ in (3.5.2). By the finiteness of C, it is permissible to interchange the order of limit and summation. Hence we obtain the contradiction $0 = 1$ when all states in C are transient. This ends the proof.

In most applications the Markov chain has no two disjoint closed sets (usually there is a state that is accessible from any other state). The next theorem summarizes a number of useful results for *finite-state* Markov chains having no two disjoint closed sets.

Theorem 3.5.7 *Let $\{X_n\}$ be a finite-state Markov chain. Suppose that the Markov chain has no two disjoint closed sets. Denote by R the set of recurrent states. Then*

(a) $f_{ij} = 1$ *for all* $i \in I$ *and* $j \in R$.

(b) $\mu_{ij} < \infty$ *for all* $i \in I$ *and* $j \in R$, *where the mean first-passage times* μ_{ij} *are defined by* $\mu_{ij} = \sum_{n=1}^{\infty} n f_{ij}^{(n)}$.

(c) If the recurrent states are aperiodic, then there is an integer $v \geq 1$ *such that* $p_{ij}^{(v)} > 0$ *for all* $i \in I$ *and* $j \in R$.

Proof Since the Markov chain has no two disjoint closed sets, the closed set R of recurrent states is irreducible by Theorem 3.5.4. Hence, by Lemma 3.5.2, any two states in R communicate with each other. This implies that for any $i, j \in R$ there is an integer $n \geq 1$ such that $p_{ij}^{(n)} > 0$. Next we prove that for any $i \in I$ and $j \in R$ there is an integer $n \geq 1$ such that $p_{ij}^{(n)} > 0$. To verify this, assume to the contrary that there is a transient state $i \in I$ such that no state $j \in R$ is accessible from i. Then there is a closed set that contains i and is disjoint from R. This contradicts the assumption that the Markov chain has no two disjoint closed sets. Hence for any transient state $i \in R$ there is a state $j \in R$ that is accessible from i. Thus any state $j \in R$ is accessible from any $i \in I$, since any two states in R communicate with each other.

To verify parts (b) and (c), define under the condition $X_0 = i$ the random variable N_{ij} by

$$N_{ij} = \min\{n \geq 1 \mid X_n = j\}.$$

Fix now $j \in R$. For each $i \in I$, let r_i be the smallest positive integer n for which $p_{ij}^{(n)} > 0$. Define

$$r = \max_{i \in I} r_i \quad \text{and} \quad \rho = \min_{i \in I} p_{ij}^{(r_i)}.$$

Since I is finite, we have $r < \infty$ and $\rho > 0$. Next observe that

$$P\{N_{ij} > r\} \leq P\{N_{ij} > r_i\} = 1 - p_{ij}^{(r_i)} \leq 1 - \rho, \quad i \in I.$$

Thus, for any $i \in I$,

$$P\{N_{ij} > kr\} \leq (1 - \rho)^k, \quad k = 0, 1, \ldots .$$

Since the probability $P\{N_{ij} > n\}$ is decreasing in n and converges to 0 as $n \to \infty$, it follows from $1 - f_{ij} = \lim_{n \to \infty} P\{N_{ij} > n\}$ that $f_{ij} = 1$. Since $P\{N_{ij} > n\}$ is decreasing in n, we also obtain

$$\mu_{ij} = \sum_{n=0}^{\infty} P\{N_{ij} > n\} = 1 + \sum_{k=1}^{\infty} \sum_{\ell=r(k-1)+1}^{rk} P\{N_{ij} > \ell\}$$

$$\leq 1 + \sum_{k=1}^{\infty} r(1 - \rho)^k,$$

showing that $\mu_{ij} < \infty$. This completes the proof of part (b).

It remains to prove (c). Fix $i \in I$ and $j \in R$. As shown above, there is an integer $v \geq 1$ such that $p_{ij}^{(v)} > 0$. By part (b) of Lemma 3.5.5 there is an integer $n_0 \geq 1$ such that $p_{jj}^{(n)} > 0$ for all $n \geq n_0$. Hence, by $p_{ij}^{(v+n)} \geq p_{ij}^{(v)} p_{jj}^{(n)}$, it follows that

$p_{ij}^{(n)} > 0$ for all $n \geq v + n_0$. Using the finiteness of I, part (c) of the theorem now follows.

Appendix: The Fox—Landi algorithm for state classification

In a finite-state Markov chain the state space can be uniquely split up into a finite number of disjoint recurrent subclasses and a (possibly empty) set of transient states. A recurrent subclass is a closed set in which all states communicate. To illustrate this, consider a Markov chain with five states and the following matrix $\mathbf{P} = (p_{ij})$ of one-step transition probabilities:

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 & 0 \\ 0.1 & 0 & 0.2 & 0.3 & 0.4 \\ 0 & 0.4 & 0.3 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

For such small examples, a state diagram is useful for doing the state classification. The state diagram uses a Boolean representation of the p_{ij}. An arrow is drawn from state i to state j only if $p_{ij} > 0$. The state diagram is given in Figure 3.5.1. By inspection it is seen that the set of transient states is $T = \{3, 4\}$ and the set of recurrent states is $R = \{1, 2, 5\}$. The set R of recurrent states can be split into two disjoint recurrent subclasses $R_1 = \{1, 2\}$ and $R_2 = \{5\}$. State 5 is absorbing.

This example was analysed by visual inspection. In general it is possible to give a systematic procedure for identifying the transient states and the recurrent subclasses in a finite-state Markov chain. The Fox—Landi algorithm (Fox and Landi 1968) first transforms the one-step transition matrix $P = (p_{ij})$ into a Boolean matrix $B = (b_{ij})$ by

$$b_{ij} = \begin{cases} 1 & \text{if } p_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

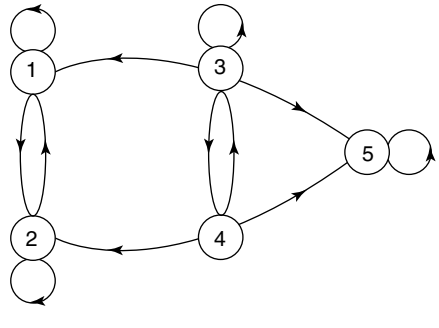


Figure 3.5.1 The state diagram for a Markov chain

The states are numbered or renumbered as $i = 1, \ldots, N$. The algorithm uses the following four rules:

(a) State i is absorbing if and only if $b_{ii} = 1$ and $b_{ij} = 0$ for $j \neq i$.

(b) If state j is absorbing and $b_{ij} = 1$, then state i is transient.

(c) If state j is transient and $b_{ij} = 1$, then state i is transient.

(d) If state i communicates with state j and state j communicates with state k, then state i communicates with state k.

The goal of the algorithm is to find all recurrent subclasses and the set of transient states. The algorithm rules (a), (b), (c) and (d). In particular, make repeated use of rule (d) is used to reduce the size of the Boolean matrix B whenever possible. The algorithm works using the following steps:

Step 1. Initialize the set $T(i) := \{i\}$ for any state i. Find all absorbing states by using rule (a) and classify $T(i) = \{i\}$ as a recurrent subclass for each absorbing state i. Classify any state i such that $b_{ij} = 1$ for some absorbing state j as a transient state.
Step 2. If all states are classified, then stop; otherwise, go to step 3.
Step 3. Take an unclassified state i_0. Since state i_0 is not absorbing, there is another state i_1 (say) that can be reached from state i_0 in one step (i.e. $b_{i_0 i_1} = 1$). Continuing in this way, construct a chain of states $i_0, i_1, \ldots$ until one of the following two exclusive possibilities occurs:

- A transient state i_s is found. Then all states in $T(i_0) \cup T(i_1) \cup \ldots \cup T(i_{s-1})$ are classified as transient according to rule (c).
- A state i_s is found that was already encountered during the development of the chain, i.e. $i_s = i_r$ for some $r < s$. Go to step 4.

Step 4. The circuit of communicating states $i_r, \ldots, i_s$ is replaced by a single aggregated state i_r and the Boolean matrix B is adjusted accordingly. This is done as follows:

- Replace column i_r by the union of the columns $i_r, \ldots, i_{s-1}$ and replace row i_r by the union of the rows $i_r, \ldots, i_{s-1}$ (the union of two Boolean vectors x and y to a Boolean vector z is defined by $z_i = 0$ if $x_i = y_i = 0$ and $z_i = 1$ otherwise).
- Delete the row i_k and the column i_k for $k = r + 1, \ldots, s - 1$.
- Let $T(i_r) := T(i_r) \cup T(i_{r+1}) \cup \ldots \cup T(i_{s-1})$.

Having done this, there are two possibilities:

- State i_r is absorbing for the new Boolean matrix B. Then $T(i_r)$ is classified as a recurrent subclass of states. Classify any state that can reach the set $T(i_r)$ in one step as a transient state (rule (b)). Go to step 2.

- State i_r is not absorbing. Then there exists a state j with $b_{i_r j} = 1$. Go to step 3 and continue the chain $i_0, \ldots, i_r$ for the new Boolean matrix.

3.5.2 Ergodic Theorems

The theoretical analysis of Markov chains is much more subtle for the case of infinitely many states than for the case of finitely many states. A finite-state Markov chain is always a regenerative process with a finite mean cycle length. This is not true for infinite-state Markov chains. Recall the example with $I = \{1, 2, \ldots\}$ and $p_{i,i+1} = 1$ for all $i \in I$ and recall the example of the symmetric random walk with $I = \{0, \pm 1, \pm 2, \ldots\}$ and $p_{i,i+1} = p_{i,i-1} = \frac{1}{2}$ for all i. In the first example the Markov chain is not regenerative, while in the other example the Markov chain is regenerative but has an infinite mean cycle length. In practical applications these pathological situations occur very rarely. Typically there is a positive recurrent state that will ultimately be reached from any other state with probability one. We therefore restrict our theoretical analysis to Markov chains which satisfy Assumption 3.3.1. Let R denote the set of recurrent states of the Markov chain $\{X_n\}$. We first prove the following lemma.

Lemma 3.5.8 *Suppose that the Markov chain $\{X_n\}$ satisfies Assumption 3.3.1. Then the set R is not empty and is an irreducible set consisting of positive recurrent states. For any $j \in R$, it holds that $f_{ij} = 1$ for all $i \in I$ and $\mu_{jj} < \infty$.*

Proof The regeneration state r from Assumption 3.3.1 is recurrent and so R is not empty. Since $f_{ir} = 1$ for all $i \in I$, the Markov chain $\{X_n\}$ has no two disjoint closed sets. Hence, by Theorem 3.5.4, the set R is an irreducible set of recurrent states. Since $\mu_{rr} < \infty$, it follows from part (b) of Theorem 3.5.3 that $\mu_{jj} < \infty$ for all $j \in R$. In other words, each state $j \in R$ is positive recurrent. Also, by part (b) of Theorem 3.5.3, $f_{rj} = 1$ for all $j \in R$. Together with the assumption $f_{ir} = 1$ for all i this implies $f_{ij} = 1$ for all i when $j \in R$. This ends the proof.

Define now the probabilities π_j by

$$\pi_j = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} p_{jj}^{(k)}, \quad j \in I. \tag{3.5.3}$$

In Theorem 3.3.1 it was shown that these limits exist. Under Assumption 3.3.1, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \pi_j, \quad i, j \in I \tag{3.5.4}$$

and

$$\pi_j = \frac{1}{\mu_{jj}} > 0, \quad j \in R \tag{3.5.5}$$

(all states in R are positive recurrent). These results follow directly from Theorem 3.3.1 by noting that $\pi_j = 0$ when j is transient and $f_{ij} = 1$ for all $i \in I$ when j is recurrent. We are now able to prove a main result.

Theorem 3.5.9 *Suppose that the Markov chain $\{X_n\}$ satisfies Assumption 3.3.1. Then the probabilities π_j, $j \in I$ defined by (3.5.3) constitute the unique equilibrium distribution of the Markov chain. Moreover, letting $\{x_j, j \in I\}$ with $\sum_j |x_j| < \infty$ be any solution to the equilibrium equations*

$$x_j = \sum_{k \in I} x_k p_{kj}, \quad j \in I, \tag{3.5.6}$$

it holds that, for some constant c, $x_j = c\pi_j$ for all $j \in I$.

Proof We first show that the π_j satisfy (3.5.6) and

$$\sum_{j \in I} \pi_j = 1. \tag{3.5.7}$$

To do so, we use the relation (3.2.1) for the n-step transition probabilities. Averaging this relation over n, we obtain for any $m \geq 1$

$$\begin{aligned} \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n+1)} &= \frac{1}{m} \sum_{n=1}^{m} \sum_{k \in I} p_{ik}^{(n)} p_{kj} \\ &= \sum_{k \in I} \left(\frac{1}{m} \sum_{n=1}^{m} p_{ik}^{(n)} \right) p_{kj}, \quad j \in I, \end{aligned} \tag{3.5.8}$$

where the interchange of the order of summation is justified by the non-negativity of the terms. Next let $m \to \infty$ in (3.5.8). On the right-hand side of (3.5.8) it is not allowed to interchange limit and summation (except when I is finite). However, we can apply Fatou's lemma from Appendix A. Using (3.5.4), we find

$$\pi_j \geq \sum_{k \in I} \pi_k p_{kj}, \quad j \in I.$$

Next we conclude that the equality sign must hold in this relation for each $j \in I$, otherwise we would obtain the contradiction

$$\sum_{j \in I} \pi_j > \sum_{j \in I} \left(\sum_{k \in I} \pi_k p_{kj} \right) = \sum_{k \in I} \pi_k \sum_{j \in I} p_{kj} = \sum_{k \in I} \pi_k.$$

We have now verified that the π_j satisfy the equilibrium equations (3.5.6). The equation (3.5.7) cannot be directly concluded from $\sum_{j \in I} p_{ij}^{(n)} = 1$ for all $n \geq 1$.

However, by letting $m \to \infty$ in

$$1 = \frac{1}{m} \sum_{n=1}^{m} \left(\sum_{j \in I} p_{ij}^{(n)} \right) = \sum_{j \in I} \left(\frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)} \right)$$

and using Fatou's lemma from Appendix A, we can conclude that

$$\sum_{j \in I} \pi_j \leq 1. \tag{3.5.9}$$

Since the set R of recurrent states is not empty, we have by (3.5.5) that

$$\sum_{j \in I} \pi_j > 0. \tag{3.5.10}$$

Next we prove that the solution to the equilibrium equations (3.5.6) is uniquely determined up to a multiplicative constant. As a by-product of this proof we will find that $\sum_{j \in I} \pi_j$ must be equal to 1. Let $\{x_j\}$ with $\sum |x_j| < \infty$ be any solution to the equation (3.5.6). Substituting this equation into itself, we find

$$\begin{aligned} x_j &= \sum_{k \in I} \left(\sum_{\ell \in I} x_\ell p_{\ell k} \right) p_{kj} = \sum_{\ell \in I} x_\ell \sum_{k \in I} p_{\ell k} p_{kj} \\ &= \sum_{\ell \in I} x_\ell p_{\ell j}^{(2)}, \quad j \in I, \end{aligned}$$

where the interchange of the order of summation in the second equality is justified by Theorem A.1 in Appendix A. By repeated substitution we find $x_j = \sum_{\ell \in I} x_\ell p_{\ell j}^{(n)}$, $j \in I$ for all $n \geq 1$. Averaging this equation over n, we find after an interchange of the order of summation (again justified by Theorem A.1 in Appendix A) that

$$x_j = \sum_{\ell \in I} x_\ell \left(\frac{1}{m} \sum_{n=1}^{m} p_{\ell j}^{(n)} \right), \quad j \in I \text{ and } m \geq 1.$$

Letting $m \to \infty$ and using (3.5.4) together with the bounded convergence theorem from Appendix A, it follows that

$$x_j = \pi_j \sum_{\ell \in I} x_\ell, \quad j \in I.$$

This proves that any solution to (3.5.6) is uniquely determined up to a multiplicative constant. Summing both sides of the latter equation over j, we find

$$\sum_{j \in I} x_j = \left(\sum_{j \in I} \pi_j \right) \left(\sum_{\ell \in I} x_\ell \right).$$

Taking $x_j = \pi_j$ for all j and using (3.5.10), it follows that $\sum_{j \in I} \pi_j = 1$. This ends the proof.

Though we are mainly concerned with the Cesaro limit of the n-step transition probabilities, we also state a result about the ordinary limit. If the regeneration state r from Assumption 3.3.1 is aperiodic, then by Theorem 2.2.4, $\lim_{n\to\infty} p_{rj}^{(n)}$ exists for all j. From this result it is not difficult to obtain that

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \quad i, j \in I \tag{3.5.11}$$

when the positive recurrent state r from Assumption 3.3.1 is *aperiodic*.

Before giving the remaining proof of Theorem 3.3.2, we give an interesting interpretation of the ratio π_i/π_j for two recurrent states i and j.

Lemma 3.5.10 *Suppose that the Markov chain $\{X_n\}$ satisfies Assumption 3.3.1. Then for any two recurrent states s and ℓ*

$$E(\textit{number of visits to state } \ell \textit{ between two successive visits to state } s) = \frac{\pi_\ell}{\pi_s}.$$

Proof Fix states $\ell, s \in R$. The Markov chain can be considered as a regenerative process with the epochs at which the process visits state s as regeneration epochs. Defining a cycle as the time elapsed between two successive visits to state s, it follows from the definition of the mean recurrence time μ_{ss} that

$$E(\text{length of one cycle}) = \mu_{ss}.$$

By Lemma 3.5.8 the mean cycle length μ_{ss} is finite. Imagine that the Markov chain earns a reward of 1 each time the process visits state ℓ. Assuming that the process starts in state s, we have by the renewal-reward theorem from Chapter 2 that

$$\begin{aligned} &\text{the long-run average reward per time unit} \\ &\quad = \frac{E(\text{reward earned during one cycle})}{E(\text{length of one cycle})} \\ &\quad = \frac{1}{\mu_{ss}} E(\text{number of visits to state } \ell \text{ in one cycle}) \end{aligned} \tag{3.5.12}$$

with probability 1. On the other hand,

$$\begin{aligned} &\text{the long-run average reward per time unit} \\ &\quad = \text{the long-run average number of visits to state } \ell \text{ per time unit.} \end{aligned}$$

In the proof of Theorem 3.3.1 we have seen that

$$\begin{aligned} &\text{the long-run average number of visits to state } \ell \text{ per time unit} \\ &\quad = \pi_\ell \quad \text{with probability 1} \end{aligned} \tag{3.5.13}$$

when $X_0 = \ell$. However, this result also holds when the Markov chain starts in state s. To see this, define the indicator variable I_k equal to 1 if $X_k = \ell$ and I_k equal to 0 otherwise. Let $\omega = (s, i_1, i_2, \dots)$ be any realization of the Markov chain with i_k denoting the realized state at the kth state transition. Since $f_{s\ell} = 1$, we have for almost all ω that there is a finite integer $t = t(\omega)$ such that $i_t = \ell$. Hence, for $n > t(\omega)$,

$$\frac{1}{n}\sum_{k=1}^{n} I_k(\omega) = \frac{1}{n}\sum_{k=1}^{t(\omega)} I_k(\omega) + \frac{1}{n}\sum_{k=t(\omega)+1}^{n} I_k(\omega).$$

Letting $n \to \infty$, the first term on the right-hand side of this equation converges to zero and the second term converges to π_ℓ. This proves that (3.5.13) also holds when $X_0 = s$. Together (3.5.12), (3.5.13) and the relation $1/\mu_{ss} = \pi_s$ yield

$$\pi_\ell = \pi_s E(\text{number of visits to state } \ell \text{ in one cycle}),$$

which proves the desired result.

In Example 3.1.3, dealing with the $GI/M/1$ queue, we tried a solution of the form $\pi_j = \gamma\tau^j$, $j \geq 0$ for the equilibrium distribution of the Markov chain $\{X_n\}$ describing the number of customers present just prior to the arrival epochs. This geometric form can be proved by using Lemma 3.5.10. Since the arrival rate is less than the service rate, Assumption 3.3.1 is satisfied with the regeneration state 0. Since any two states of the Markov chain $\{X_n\}$ communicate, it follows from Lemma 3.5.2 and Theorem 3.5.3 that the state space I is an irreducible set consisting of (positive) recurrent states. Hence, by Lemma 3.5.10, we have for the $GI/M/1$ queue that

$$E(\text{number of visits to state } j+1 \text{ between two successive returns to state } j) = \frac{\pi_{j+1}}{\pi_j} \quad \text{for } j = 0, 1, \dots . \tag{3.5.14}$$

Some reflections show that the left-hand side of this equation is independent of j by the memoryless property of the exponential distribution for the service times. Hence, for some constant η, $\pi_{j+1}/\pi_j = \eta$ for all $j \geq 0$ showing that $\pi_j = \pi_0\eta^j$ for $j \geq 0$.

Next we prove Theorem 3.3.3. The proof is very similar to that of Lemma 3.5.10. Assume that the Markov chain earns a reward $f(j)$ each time it visits state j.

Theorem 3.5.11 *Suppose that the Markov chain $\{X_n\}$ satisfies the Assumptions 3.3.1 and 3.3.2. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_k) = \sum_{j\in I} f(j)\pi_j \quad \text{with probability 1}$$

for each initial state $X_0 = i$.

Proof Assume first that the initial state of the process is the regeneration state r from Assumptions 3.3.1 and 3.3.2. The Markov chain can be seen as a regenerative process with the epochs at which the process visits state r as regeneration epochs. Define a cycle as the time elapsed between two successive visits to state r. The expected cycle length equals the mean recurrence time μ_{rr} and is finite. By the renewal-reward theorem from Chapter 2,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \frac{E(\text{reward earned during one cycle})}{E(\text{length of one cycle})}$$

with probability 1. Lemma 3.5.10 states that E (number of visits to state j in one cycle) $= \pi_j/\pi_r$ for any recurrent state j. This relation is also valid for a transient state j, since a transient state is not accessible from a recurrent state and $\pi_j = 0$ for j transient. Hence

$$E(\text{reward earned during one cycle}) = \sum_{j\in I} f(j)\frac{\pi_j}{\pi_r}.$$

Since $E(\text{length of one cycle}) = \mu_{rr} = 1/\pi_r$ by (3.5.5), the assertion of the theorem is now proved when $X_0 = r$. Take next any initial state $X_0 = i$. As in the proof of Lemma 3.5.10, let $\omega = (i_0, i_1, i_2, \dots)$ be any realization of the Markov chain with $i_0 = i$ and let i_k denote the realized state at the kth state transition. Since $f_{ir} = 1$, we have for almost all ω that there is a finite integer $t = t(\omega)$ such that $i_t = r$. Hence

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k(\omega)) = \frac{1}{n}\sum_{k=1}^{t(\omega)} f(X_k(\omega)) + \frac{1}{n}\sum_{k=t(\omega)+1}^{n} f(X_k(\omega)).$$

Letting $n \to \infty$, it follows from part (b) of Assumption 3.3.2 that the first term on the right-hand side of the equation tends to zero, while by the above proof the second term converges to $\sum_{j\in I} f(j)\pi_j$. This completes the proof.

Markov's proof and exponential convergence

It is interesting to examine the original proof of Markov (1906) for the existence of a limiting distribution in a finite-state Markov chain. The proof is not just of historical interest and the ideas it uses are still very much alive. The proof also establishes the rate of convergence to the limiting distribution. An aperiodic finite-state Markov chain with no two disjoint closed sets is assumed. The Markov chain is said to be *aperiodic* when the period of the recurrent states is equal to 1; see Lemma 3.5.5.

Theorem 3.5.12 *Let $\{X_n\}$ be a finite-state Markov chain with no two disjoint closed sets. Suppose that the Markov chain is aperiodic. Then there exists a probability distribution $\{\pi_j, j \in I\}$ and numbers $\alpha > 0$ and $0 < \beta < 1$ such that for all*

$i, j \in I$,

$$|p_{ij}^{(n)} - \pi_j| \leq \alpha\beta^n, \quad n = 1, 2, \ldots .$$

In particular,

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \quad i, j \in I.$$

Proof Let s be any recurrent state of the Markov chain. Since the Markov chain is aperiodic, we have by part (c) of Theorem 3.5.7 that there exists an integer $\nu \geq 1$ and a number $\rho > 0$ such that

$$p_{is}^{(\nu)} \geq \rho \quad i \in I.$$

For any $j \in I$, define the sequences $\{M_j^{(n)}, n \geq 0\}$ and $\{m_j^{(n)}, n \geq 0\}$ by

$$M_j^{(n)} = \max_{i\in I} p_{ij}^{(n)} \quad \text{and} \quad m_j^{(n)} = \min_{i\in I} p_{ij}^{(n)}.$$

Note that $M_j^{(0)} = 1$ and $m_j^{(0)} = 0$. Applying relation (3.2.2), we find

$$M_j^{(n+1)} = \max_{i\in I} \sum_{k\in I} p_{ik} p_{kj}^{(n)} \leq \max_{i\in I} \sum_{k\in I} p_{ik} M_j^{(n)} = M_j^{(n)} \max_{i\in I} \sum_{k\in I} p_{ik},$$

and so, for any $j \in I$,

$$M_j^{(n+1)} \leq M_j^{(n)}, \quad n = 0, 1, \ldots .$$

Similarly, we find for any $j \in I$ that

$$m_j^{(n+1)} \geq m_j^{(n)}, \quad n = 0, 1, \ldots .$$

Since the sequences $\{M_j^{(n)}\}$ and $\{m_j^{(n)}\}$ are bounded and monotone, they have finite limits. Next we establish the inequality

$$0 \leq M_j^{(n)} - m_j^{(n)} \leq (1 - \rho)[M_j^{(n-\nu)} - m_j^{(n-\nu)}], \quad n \geq \nu \tag{3.5.15}$$

for any $j \in I$. Suppose for the moment that we have proved this inequality. A repeated application of the inequality shows that

$$0 \leq M_j^{(n)} - m_j^{(n)} \leq (1 - \rho)^{[n/\nu]}(M_j^{(0)} - m_j^{(0)}), \quad n = 0, 1, \ldots , \tag{3.5.16}$$

where $[x]$ denotes the largest integer contained in x. Here we used the fact that $M_j^{(n)} - m_j^{(n)}$ is decreasing in n. By (3.5.16), we have that the limits of the monotone sequences $\{M_j^{(n)}\}$ and $\{m_j^{(n)}\}$ coincide. Denote the common limit by π_j. Hence

$$\lim_{n\to\infty} M_j^{(n)} = \lim_{n\to\infty} m_j^{(n)} = \pi_j.$$

Using the inequalities $m_j^{(n)} \le p_{ij}^{(n)} \le M_j^{(n)}$ and $m_j^{(n)} \le \pi_j \le M_j^{(n)}$, we find

$$|p_{ij}^{(n)} - \pi_j| \le M_j^{(n)} - m_j^{(n)}, \quad n = 0, 1, \ldots \tag{3.5.17}$$

for any $i, j \in I$. Together the inequalities (3.5.16) and (3.5.17) yield the assertion of the theorem except that we have still to verify that $\{\pi_j\}$ represents a probability distribution. Obviously, the π_j are non-negative. Since $\sum_{j\in I} p_{ij}^{(n)} = 1$ for all n and $p_{ij}^{(n)} \to \pi_j$ as $n \to \infty$, we obtain from the finiteness of I that the π_j sum to 1.

It remains to verify (3.5.15). To do so, fix $j \in I$ and $n \ge \nu$. Let x and y be the states for which $M_j^{(n)} = p_{xj}^{(n)}$ and $m_j^{(n)} = p_{yj}^{(n)}$. Then

$$\begin{aligned} 0 \le M_j^{(n)} - m_j^{(n)} &= p_{xj}^{(n)} - p_{yj}^{(n)} = \sum_{k\in I} p_{xk}^{(\nu)} p_{kj}^{(n-\nu)} - \sum_{k\in I} p_{yk}^{(\nu)} p_{kj}^{(n-\nu)} \\ &= \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\} p_{kj}^{(n-\nu)} \\ &= \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\}^+ \; p_{kj}^{(n-\nu)} - \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\}^- \; p_{kj}^{(n-\nu)}, \end{aligned}$$

where $a^+ = \max(a, 0)$ and $a^- = -\min(a, 0)$. Hence, by $a^+, a^- \ge 0$,

$$\begin{aligned} 0 \le M_j^{(n)} - m_j^{(n)} &\le \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\}^+ M_j^{(n-\nu)} - \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\}^- m_j^{(n-\nu)} \\ &= \sum_{k\in I} \{p_{xk}^{(\nu)} - p_{yk}^{(\nu)}\}^+ [M_j^{(n-\nu)} - m_j^{(n-\nu)}], \end{aligned}$$

where the last equality uses the fact that $\sum_k a_k^+ = \sum_k a_k^-$ if $\sum_k a_k = 0$. Using the relation $(a - b)^+ = a - \min(a, b)$, we next find

$$0 \le M_j^{(n)} - m_j^{(n)} \le \left[1 - \sum_{k\in I} \min(p_{xk}^{(\nu)}, p_{yk}^{(\nu)})\right] \left[M_j^{(n-\nu)} - m_j^{(n-\nu)}\right].$$

Since $p_{is}^{(\nu)} \ge \rho$ for all i, we find

$$1 - \sum_{k\in I} \min(p_{xk}^{(\nu)}, p_{yk}^{(\nu)}) \le 1 - \min(p_{xs}^{(\nu)}, p_{ys}^{(\nu)}) \le 1 - \rho,$$

which implies the inequality (3.5.15). This completes the proof.

Exponential convergence of the n-step transition probabilities does not hold in general for an infinite-state Markov chain. Strong recurrence conditions should be imposed to establish exponential convergence in infinite-state Markov chains.

EXERCISES

3.1 A production machine has two crucial parts which are subject to failures. The two parts are identical. The machine works as long as one of the two parts is functioning. A repair is done when both parts have failed. A repair takes one day and after each repair the system is as good as new. An inspection at the beginning of each day reveals the exact condition of each part. If at the beginning of a day both parts are in good condition, then at the end of the day both parts are still in good condition with probability 0.50, one of them is broken down with probability 0.25 and both are broken down with probability 0.25. If at the beginning of the day only one part is in good condition, this part is still in good condition at the end of the day with probability 0.50. Define a Markov chain to describe the functioning of the machine and specify the one-step transition probabilities.

3.2 To improve the reliability of a production system, two identical production machines are connected in parallel. For the production process only one of the machines is used; the other machine is standby. At the end of the day the used machine is inspected. Regardless how long the machine has already been in uninterrupted use, the probability that an inspection reveals the necessity for revision is $\frac{1}{10}$. A revision takes exactly two days. During the revision the other machine takes over the production if that machine is available. The production process must be stopped when both machines are in revision. Assuming that there are two repairmen, define an appropriate Markov chain to describe the functioning of the production system and specify the one-step transition probabilities of the Markov chain.

3.3 Containers are temporarily stored at a stockyard with ample capacity. At the beginning of each day precisely one container arrives at the stockyard. Each container stays a certain amount of time at the stockyard before it is removed. The residency times of the containers are independent of each other. Specify for each of the following two cases the state variable(s) and the one-step transition probabilities of a Markov chain that can be used to analyse the number of containers present at the stockyard at the end of each day.

(a) The residency time of a container is exponentially distributed with a mean of $1/\mu$ days.

(b) The residency time of a container has an exponential distribution whose mean is $1/\mu_1$ days with probability p and is $1/\mu_2$ days with probability $1 - p$.

3.4 Two teams, A and B, meet each other in a series of games until either of the teams has won three games in a row. Each game results in a win for either of the teams (no draw is possible). The outcomes of the games are independent of each other. Define an appropriate Markov chain to determine the probability distribution of the length of the match when the two teams are equally strong.

3.5 Consider Exercise 3.4 again, but assume now that team A wins a given game with a probability larger than $\frac{1}{2}$.

(a) Use Markov chain analysis to determine the probability distribution of the length of the match. Explain how to calculate the probability that team A wins the match.

(b) Explain how to modify the Markov chain analysis when a draw between the teams is possible with positive probability?

3.6 You play the following game. A fair coin is flipped until heads appears three times in a row. You get \$12 each time this happens, but you have to pay \$1 for each flip of the coin. Use Markov chain analysis to find out whether this game is fair.

3.7 Consider the following variant of the coupon-collecting problem. A fair die is thrown until each of the six possible outcomes $1, 2, \ldots, 6$ has appeared. Use a Markov chain with seven states to calculate the probability distribution of the number of throws needed.

3.8 The gambler Joe Dalton has \$100 and his goal is to double this amount. Therefore he plays a gambling game in which he loses his stake with probability 0.60, but wins two or

three times his stake with respective probabilities 0.25 and 0.15. His strategy is to bet \$5 each time his payroll is more than \$50 dollars and \$10 otherwise. Define an appropriate Markov chain to compute the probability that Joe reaches his goal. Also calculate the expected number of bets placed by Joe until he has gone broke or reached his goal.

3.9 A training program consists of three parts, each having a length of one month. Fifty percent of the starting students immediately pass the first part after one month, 30% drop out before the end of the first month and 20% take the first part again. Seventy percent of the last group pass the first part after a second trial and the other 30% still drop out. Eighty percent of the students taking the second part pass this second part after the first trial, 10% drop out after the first trial and the other 10% move on after a second trial of the first part. Any student streaming into the third part of the training program will complete it successfully. Calculate the probability that a starting student will be successful.

3.10 Consider a finite-state Markov chain $\{X_n\}$ with no two disjoint closed sets. The matrix of one-step transition probabilities is called *doubly stochastic* when for each column the sum of the column elements equals 1. Verify that the equilibrium distribution of such a Markov chain is a uniform distribution.

3.11 A gambling device is tuned such that a player who wins (loses) on a given play will win on the next play with probability 0.25 (0.50). The player pays \$1 for each play and receives \$2.50 for each play that is won. Use Markov chain analysis to find out whether the game is fair or not.

3.12 A factory has a storage tank with a capacity of 4 m^3 for temporarily storing waste produced by the factory. Each week the factory produces 0, 1, 2 or 3 m^3 waste with respective probabilities $p_0 = \frac{1}{8}$, $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, and $p_3 = \frac{1}{8}$. If the amount of waste produced in one week exceeds the remaining capacity of the tank, the excess is specially removed at a cost of \$30 per cubic metre. At the end of each week there is a regular opportunity to remove waste from the storage tank at a fixed cost of \$25 and a variable cost of \$5 per cubic metre. The following policy is used. If at the end of the week the storage tank contains more than 2 m^3 of waste, the tank is emptied; otherwise no waste is removed. Use Markov chain analysis to find the long-run average cost per week.

3.13 In a series of repeated plays, you can choose each time between games A and B. During each play you win \$1 or you lose \$1. You are also allowed to play when your capital is not positive (a negative capital corresponds to a debt). In game A there is a single coin. This coin lands heads with probability $\frac{1}{2} - \varepsilon$ ($\varepsilon = 0.005$) and tails with probability $\frac{1}{2} + \varepsilon$. In game B there are two coins. One coin lands heads with probability $\frac{1}{10} - \varepsilon$ and the other coin lands heads with probability $\frac{3}{4} - \varepsilon$. If you play game B, then you must take the first coin when your current capital is a multiple of 3 and you must take the other coin otherwise. In each play of either game you win \$1 if the coin lands heads and you lose \$1 otherwise.

(a) Use Markov chain analysis to verify that the long-run fraction of plays you win is 0.4957 when you always play game B (*Hint*: a three-state Markov chain suffices.)

(b) Suppose you alternately play the games $A, A, B, B, A, A, B, B, \ldots$. Use an appropriate Markov chain to verify that the long-run fraction of plays you win is 0.5064.

This problem shows that in special cases with dependencies, a combination of two unfavourable games may result in a favourable game. This paradox is called *Parrondo's paradox* after the Spanish physicist Juan Parrondo.

3.14 At the beginning of each day, a crucial piece of electronic equipment is inspected and then classified as being in one of the working conditions $i = 1, \ldots, N$. Here the working condition i is better than the working condition $i + 1$. If the working condition is $i = N$ the piece must be replaced by a new one and such an enforced replacement takes two days. If the working condition is i with $i < N$ there is a choice between preventively replacing

the piece by a new one and letting the piece operate for the present day. A preventive replacement takes one day. A new piece has working condition $i = 1$. A piece whose present working condition is i has the next day working condition j with known probability q_{ij} where $q_{ij} = 0$ for $j < i$. The following replacement rule is used. The current piece is only replaced by a new one when its working condition is greater than the critical value m, where m is a given integer with $1 \leq m < N$.

(a) Define an appropriate Markov chain and specify its one-step transition probabilities.

(b) Explain how to calculate the long-run fraction of days the equipment is inoperative and the fraction of replacements occurring in the failure state N.

3.15 Consider a stochastically failing piece of equipment with two identical components that operate independently of each other. The lifetime in days of each component has a discrete probability distribution $\{p_j, j = 1, \ldots, M\}$. A component in the failure state at the beginning of a day is replaced instantaneously. It may be economical to preventively replace the other working component at the same time the failed component has to be replaced. The cost of replacing only one component is K_1, while the cost of replacing simultaneously both components equals K_2 with $0 < K_2 < 2K_1$. The control rule is as follows. Replace a component upon failure or upon reaching the age of R days, whichever occurs first. If a component is replaced and the other component is still working, the other component is preventively replaced when it has been in use for r or more days. The parameters r and R are given integers with $1 \leq r < R$.

(a) Define an appropriate Markov chain and specify its one-step transition probabilities.

(b) How can you calculate the long-run average cost per day?

3.16 A transmission channel transmits messages one at a time, and transmission of a message can only start at the beginning of a time slot. The time slots have unit length and the transmission time of a message is one time slot. However, each transmission can fail with some probability f. A failed transmission is tried again at the beginning of the next time slot. The numbers of new messages arriving during the time slots are independent random variables with a common discrete distribution $\{a_k, k = 0, 1, \ldots\}$. Newly arriving messages are temporarily stored in a buffer of ample capacity. It is assumed that the average arrival rate of new messages is smaller than the average number of attempts needed to transmit a message successfully, that is, $\sum_k k a_k < 1/f$. The goal is to find the long-run average throughput per time unit.

(a) Define an appropriate Markov chain with a one-dimensional state space and specify its one-step transition probabilities.

(b) Can you give a recursive algorithm for the computation of the state probabilities? Express the average throughput in terms of the state probabilities.

3.17 Messages arrive at a transmission channel according to a Poisson process with rate λ. The channel can transmit only one message at a time and a new transmission can only start at the beginnings of the time slots $t = 1, 2, \ldots$. The transmission time of a message is one time slot. The following access-control rule is used. The gate is closed for newly arriving messages when the number of messages awaiting transmission has reached the level R and is opened again when the number of messages awaiting transmission has dropped to the level r, where the parameters r and R are given integers with $0 \leq r < R$. The goal is to study the long-run fraction of lost messages as function of r and R.

(a) Define an appropriate Markov chain and specify its one-step transition probabilities.

(b) Show how to calculate the long-run fraction of lost messages.

3.18 In Example 3.5.1 we have determined for the $GI/M/1$ queue the customer-average probability π_j denoting the long-run fraction of customers who find j other customers present upon arrival. Denote by the time-average probability p_j the long-run fraction of time that j customers are present for $j = 0, 1, \ldots$. Use Theorem 3.3.3 and Lemma 1.1.4

to verify that

$$p_j = \sum_{k=j-1}^{\infty} \pi_k \int_0^{\infty} \left(\sum_{\ell=k+1-j}^{\infty} \frac{t}{\ell+1} e^{-\mu t} \frac{(\mu t)^{\ell}}{\ell!} \right) a(t)\, dt, \quad j \geq 1.$$

(*Hint*: fix j and assume that the process incurs a cost at rate 1 whenever j customers are present and a cost at rate 0 otherwise. Imagine that the server continues servicing fictitious customers when the system is empty so that actual or fictitious service completions occur according to a Poisson process with rate μ.)

3.19 In each time unit a job arrives at a conveyor with a single workstation. The workstation can process only one job at a time and has a buffer with ample capacity to store the arriving jobs that find the workstation busy. The processing times of the jobs are independent random variables having a common Erlang (r, μ) distribution. It is assumed that $r/\mu < 1$.

(a) Define an appropriate Markov chain to analyse the number of jobs in the buffer just prior to the arrival epochs of new jobs and specify the one-step transition probabilities.

(b) Explain how to calculate the long-run average delay in the buffer per job.

(c) Prove that the equilibrium distribution of this Markov chain has a geometric tail.

3.20 Consider Exercise 3.19 again but now assume that the buffer has finite capacity. Any arriving job that finds the buffer full is lost. Show how to calculate the long-run fraction of lost jobs and the long-run fraction of time the workstation is busy (*Hint*: use Little's formula for the latter performance measure).

3.21 At the telephone exchange, calls arrive according to a Poisson process with rate λ. The calls are first put in an infinite-capacity buffer before they can be processed further. The buffer is periodically scanned every T time units, and only at those scanning epochs are calls in the buffer allocated to free transmission lines. There are c transmission lines and each transmission line can handle only one call at a time. The transmission times of the calls are independent random variables having a common exponential distribution with mean $1/\mu$.

(a) Use Markov chain analysis to find the equilibrium distribution $\{\pi_j\}$ of the number of calls in the buffer just prior to the scanning epochs.

(b) Argue that the long-run average number of calls in the buffer is given by

$$L_q = \sum_{j=c+1}^{\infty} (j-c)\pi_j + \frac{1}{2}\lambda T.$$

(*Hint*: imagine that each call is marked upon arrival and is unmarked at the next scanning epoch. Argue that the average number of marked calls in the buffer is $\frac{1}{2}\lambda T$.)

(c) What is the long-run average delay in the buffer per call?

3.22 Consider Example 3.4.1 with Poisson arrivals of messages.

(a) Prove the validity of the relation $\lambda = \sum_{j=1}^{c-1} j\pi_j + c\sum_{j=c}^{\infty} \pi_j$ and note that this relation can be used as an accuracy check on the calculated values of the state probabilities π_j, $j = 0, 1, \ldots$.

(b) Use the hint in Exercise 3.21 to prove that the long-run average number of messages in the buffer equals $\sum_{j=c+1}^{\infty}(j-c)\pi_j + \frac{1}{2}\lambda T$.

(c) What is the long-run average delay in the buffer per message?

3.23 Consider Example 3.4.1 again but assume now that the buffer for temporarily storing arriving messages has a finite capacity K. Each arriving message that finds the buffer full is lost.

(a) Modify the one-step transition probabilities of the Markov chain $\{X_n\}$ describing the number of messages in the buffer at the end of the time slots.

(b) Denoting by $\{\pi_j^{(K)}, j = 0, 1, \ldots, K\}$ the equilibrium distribution of the Markov chain, argue that the long-run fraction of messages lost is

$$\pi_{loss}(K) = \frac{1}{\lambda}\left[\lambda - \sum_{j=1}^{c-1} j\pi_j^{(K)} - c\sum_{j=c}^{K}\pi_j^{(K)}\right].$$

(*Hint*: the sum of the average number of messages lost per time unit and the average number of messages transmitted per time unit equals λ.)

(c) Let $K(\alpha)$ be the smallest value of K for which $\pi_{loss}(K) \leq \alpha$ for a given value of α. Letting $\rho = \lambda/c$, compute for $\rho = 0.90, 0.95$ and $c = 1, 5, 10$ the values of $K(\alpha)$ as given in the table below. Note that $K(\alpha)$ increases logarithmically in α as α increases. What does this mean for the asymptotic behaviour of $\pi_{loss}(K)$ as K gets large?

	$\rho = 0.80$			$\rho = 0.95$		
α	$c = 1$	$c = 5$	$c = 10$	$c = 1$	$c = 5$	$c = 10$
10^{-6}	29	32	36	107	110	114
10^{-8}	40	42	46	152	155	159
10^{-10}	50	53	57	197	200	204

3.24 Suppose that a conveyer belt is running at a uniform speed and transporting items on individual carriers equally spaced along the conveyer. There are two workstations $i = 1, 2$ placed in order along the conveyer, where station 1 is the first one. In each time unit an item for processing arrives and is handled by the first workstation that is idle. Any station can process only one item at a time and has no storage capacity. An item that finds both workstations busy is lost. The processing time of an item at station i has an Erlang-r_i distribution with mean m_i, $i = 1, 2$. Give a Markov chain analysis aimed at the computation of the loss probability. Solve these two cases:

(a) The processing times at the stations 1 and 2 are exponentially distributed with respective means $m_1 = 0.75$ and $m_2 = 1.25$ (answer 0.0467).

(b) The processing times at the stations 1 and 2 are Erlang-3 distributed with respective means $m_1 = 0.75$ and $m_2 = 1.25$ (answer 0.0133).

3.25 Leaky bucket control is a control procedure used in telecommunication networks. It controls the average packet input into the network and the maximum number of packets transmitted in succession. To achieve this, a token buffer is used. An arriving packet is admitted to the network only if the token buffer is not empty, otherwise the packet is rejected. If the token buffer is not empty when a packet arrives, the packet immediately removes one token from the token buffer and enters the network. The token buffer is of size M. Tokens are generated periodically every D time units and are stored in the token buffer. Tokens generated when the token buffer is full are lost. Packets arrive at the network according to a Poisson process with rate λ.

(a) Analyse the embedded Markov chain describing the number of tokens in the pool just before a token is generated.

(b) What is the average number of packets admitted in one token generation interval? For several values of M investigate how the average input curve behaves as a function of λD.

BIBLIOGRAPHIC NOTES

Many good textbooks on stochastic processes are available and most of them treat the topic of Markov chains. My favourite books include Cox and Miller (1965), Karlin and Taylor (1975) and Ross (1996), each offering an excellent introduction to Markov chain theory. A very fundamental treatment of denumerable Markov chains can be found in the book of Chung (1967). An excellent book on Markov chains with a general state space is Meyn and Tweedie (1993). The concept of the embedded Markov chain and its application in Example 3.1.3 are due to Kendall (1953). The idea of using the geometric tail behaviour of state probabilities goes back to Feller (1950) and was successfully used in the papers of Everett (1954) and Takahashi and Takami (1976).

REFERENCES

Chung, K.L. (1967) *Markov Chains with Stationary Transition Probabilities*, 2nd edn. Springer-Verlag, Berlin.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes*. Chapman and Hall, London.

Everett, J. (1954) State probabilities in congestion problems characterized by constant holding times. *Operat. Res.*, **1**, 279–285.

Feller, W. (1950) *An Introduction to Probability Models and its Applications*, Vol. I, John Wiley & Sons, Inc., New York.

Fox, B. and Landi, D.M. (1968) An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix. *Commun. ACM*, **11**, 619–621.

Karlin, S. and Taylor, H.M. (1975) *A First Course in Stochastic Processes*, 2nd edn. Academic Press, New York.

Kendall, D.G. (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain. *Ann. Math. Statist.*, **24**, 338–354.

Markov, A.A. (1906) Extension of the law of large numbers to dependent events (in Russian). *Bull. Soc. Phys. Math. Kazan*, **15**, 255–261.

Meyn, S.P. and Tweedie, R. (1993) *Markov Chains and Stochastic Stability.* Springer-Verlag, Berlin.

Ross, S.M. (1996) *Stochastic Processes*, 2nd edn., John Wiley & Sons, Inc., New York.

Stewart, W.J. (1994) *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton NJ.

Takahashi, Y. and Takami, Y. (1976) A numerical method for the steady-state probabilities of a $GI/G/c$ queueing system in a general class. *J. Operat. Res. Soc. Japan*, **19**, 147–157.

CHAPTER 4

Continuous-Time Markov Chains

4.0 INTRODUCTION

In the continuous-time analogue of discrete-time Markov chains the times between successive state transitions are not deterministic, but *exponentially distributed*. However, the state transitions themselves are again governed by a (discrete-time) Markov chain. Equivalently, a continuous-time Markov chain can be represented by so-called *infinitesimal transition rates*. This is in analogy with the 'Δt-representation' of the Poisson process. The representation by infinitesimal transition rates leads naturally to the flow rate equation approach. This approach is easy to visualize and is widely used in practice. The continuous-time Markov chain model is introduced in Section 4.1. In Section 4.2 we discuss the flow rate equation approach. The discussion in Section 4.2 concentrates on giving insights into this powerful approach but no proofs are given. The proofs are given in Section 4.3. Results for discrete-time Markov chains are the basis for the proofs of the ergodic theorems for continuous-time Markov chains.

In Section 4.4 we discuss specialized methods to solve the equilibrium equations for continuous-time Markov chains on a semi-infinite strip in two-dimensional space. Many applications of continuous-time Markov chains have this structure. Section 4.5 deals with transient analysis for continuous-time Markov chains. The basic tools for the computation of the transient state probabilities and first passage time probabilities are Kolmogoroff's method of linear differential equations and the probabilistic method of uniformization. Both methods will be discussed. In Section 4.6 we give algorithms for the computation of the transient probability distribution of the cumulative reward in a continuous-time Markov chain model with a reward structure. A special case of this model is the computation of the transient distribution of the sojourn time of the process in a given set of states.

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

4.1 THE MODEL

In Chapter 3 we considered Markov processes in which the changes of the state only occurred at fixed times $t = 0, 1, \ldots$. However, in numerous practical situations, changes of state may occur at each point of time. One of the most appropriate models for analysing such situations is the continuous-time Markov chain model. In this model the times between successive transitions are exponentially distributed, while the succession of states is described by a discrete-time Markov chain. A wide variety of applied probability problems can be modelled as a continuous-time Markov chain by an appropriate state description.

In analogy with the definition of a discrete-time Markov chain, a continuous-time Markov chain is defined as follows.

Definition 4.1.1 *A continuous-time stochastic process* $\{X(t), t \geq 0\}$ *with discrete state space* I *is said to be a continuous-time Markov chain if*

$$P\{X(t_n) = i_n \mid X(t_0) = i_0, \ldots, X(t_{n-1}) = i_{n-1}\}$$
$$= P\{X(t_n) = i_n \mid X(t_{n-1}) = i_{n-1}\}$$

for all $0 \leq t_0 < \cdots < t_{n-1} < t_n$ *and* $i_0, \ldots, i_{n-1}, i_n \in I$.

Just as in the discrete-time case, the Markov property expresses that the conditional distribution of a future state given the present state and past states depends only on the present state and is independent of the past. In the following we consider time-homogeneous Markov chains for which the transition probability $P\{X(t+u) = j \mid X(u) = i\}$ is independent of u. We write

$$p_{ij}(t) = P\{X(t+u) = j \mid X(u) = i\}.$$

The theory of continuous-time Markov chains is much more intricate than the theory of discrete-time Markov chains. There are very difficult technical problems and some of them are not even solved at present time. Fortunately, the staggering technical problems do not occur in practical applications. In our treatment of continuous-time Markov chains we proceed pragmatically. We impose a regularity condition that is not too strong from a practical point of view but avoids all technical problems.

As an introduction to the modelling by a continuous-time Markov chain, let us construct the following Markov jump process. A stochastic system with a discrete state space I jumps from state to state according to the following rules:

Rule (a) *If the system jumps to state* i, *it then stays in state* i *for an exponentially distributed time with mean* $1/v_i$ *independently of how the system reached state* i *and how long it took to get there.*

Rule (b) *If the system leaves state* i, *it jumps to state* $j (j \neq i)$ *with probability* p_{ij} *independently of the duration of the stay in state* i, *where* $\sum_{j \neq i} p_{ij} = 1$ *for all* $i \in I$.

The convention $p_{ii} = 0$ for all states i is convenient and natural. This convention ensures that the sojourn time in a state is unambiguously defined. If there are no absorbing states, it is no restriction to make this convention (the sum of a geometrically distributed number of independent lifetimes with a common exponential distribution is again exponentially distributed). Throughout this chapter the following assumption is made.

Assumption 4.1.1 *In any finite time interval the number of jumps is finite with probability* 1.

Define now the continuous-time stochastic process $\{X(t), t \geq 0\}$ by

$$X(t) = \text{the state of the system at time } t.$$

The process is taken to be right-continuous; that is, at the transition epochs the state of the system is taken as the state just after the transition. The process $\{X(t)\}$ can be shown to be a continuous-time Markov chain. It will be intuitively clear that the process has the Markov property by the assumption of exponentially distributed sojourn times in the states. Assumption 4.1.1 is needed to exclude pathological cases. For example, suppose the unbounded state space $I = \{1, 2, \dots\}$, take $p_{i,i+1} = 1$ and $\nu_i = i^2$ for all i. Then transitions occur faster and faster so that the process will ultimately face an explosion of jumps. With a finite state space the Assumption 4.1.1 is always satisfied.

Example 4.1.1 Inventory control for an inflammable product

An inflammable product is stored in a special tank at a filling station. Customers asking for the product arrive according to a Poisson process with rate λ. Each customer asks for one unit of the product. Any demand that occurs when the tank is out of stock is lost. Opportunities to replenish the stock in the tank occur according to a Poisson process with rate μ. The two Poisson processes are assumed to be independent of each other. For reasons of security it is only allowed to replenish the stock when the tank is out of stock. At those opportunities the stock is replenished with Q units for a given value of Q.

To work out the long-run average stock in the tank and the long-run fraction of demand that is lost, we need to study the inventory process. For any $t \geq 0$, define

$$X(t) = \text{the amount of stock in the tank at time } t.$$

The stochastic process $\{X(t), t \geq 0\}$ is a continuous-time Markov chain with state space $I = \{0, 1, \dots, Q\}$. The sojourn time in each state is exponentially distributed, since both the times between the demand epochs and the times between the replenishment opportunities are exponentially distributed. Thus the sojourn time in state i has an exponential distribution with parameter

$$\nu_i = \begin{cases} \lambda, & i = 1, \dots, Q, \\ \mu, & i = 0. \end{cases}$$

The state transitions are governed by a discrete-time Markov chain whose one-step transition probabilities have the simple form

$$p_{i,i-1} = 1 \quad \text{for } i = 1, \ldots, Q,$$
$$p_{0Q} = 1 \quad \text{and the other } p_{ij} = 0.$$

Infinitesimal transition rates

Consider the general Markov jump process $\{X(t)\}$ that was constructed above. The sojourn time in any state i has an exponential distribution with mean $1/\nu_i$ and the state transitions are governed by a Markov chain having one-step transition probabilities p_{ij} for $i, j \in I$ with $p_{ii} = 0$ for all i. The Markov process allows for an equivalent representation involving the so-called infinitesimal transition rates. To introduce these rates, let us analyse the behaviour of the process in a very small time interval of length Δt. Recall that the exponential (sojourn-time) distribution has a constant failure rate; see Appendix B. Suppose that the Markov process $\{X(t)\}$ is in state i at the current time t. The probability that the process will leave state i in the next Δt time units with Δt very small equals $\nu_i \Delta t + o(\Delta t)$ by the constant failure rate representation of the exponential distribution. If the process leaves state i, it jumps to state j $(\neq i)$ with probability p_{ij}. Hence, for any $t > 0$,

$$P\{X(t + \Delta t) = j \mid X(t) = i\} = \begin{cases} \nu_i \Delta t \times p_{ij} + o(\Delta t), & j \neq i, \\ 1 - \nu_i \Delta t + o(\Delta t), & j = i, \end{cases}$$

as $\Delta t \to 0$. One might argue that in the next Δt time units state j could be reached from state i by first jumping from state i to some state k and next jumping in the same time interval from state k to state j. However, the probability of two or more state transitions in a very small time interval of length Δt is of the order $(\Delta t)^2$ and is thus $o(\Delta t)$; that is, this probability is negligibly small compared with Δt as $\Delta t \to 0$. Define now

$$q_{ij} = \nu_i p_{ij}, \quad i, j \in I \text{ with } j \neq i.$$

The non-negative numbers q_{ij} are called the *infinitesimal* transition rates of the continuous-time Markov chain $\{X(t)\}$. Note that the q_{ij} uniquely determine the sojourn-time rates ν_i and the one-step transition probabilities p_{ij} by $\nu_i = \sum_{j \neq i} q_{ij}$ and $p_{ij} = q_{ij}/\nu_i$. The q_{ij} themselves are not probabilities but transition rates. However, for Δt very small, $q_{ij} \Delta t$ can be interpreted as the probability of moving from state i to state j within the next Δt time units when the current state is state i.

In applications one usually proceeds in the reverse direction. The infinitesimal transition rates q_{ij} are determined in a direct way. They are typically the result of the interaction of two or more elementary processes of the Poisson type. Contrary to the discrete-time case in which the one-step transition probabilities determine unambiguously a discrete-time Markov chain, it is not generally true that the infinitesimal transition rates determine a unique continuous-time Markov chain.

Here we run into subtleties that are well beyond the scope of this book.* Note that fundamental difficulties may arise when the state space is infinite, but these difficulties are absent in almost all practical applications. To avoid the technical problems, we make the following assumption for the given data q_{ij}.

Assumption 4.1.2 *The rates* $\nu_i = \sum_{j \neq i} q_{ij}$ *are positive and bounded in* $i \in I$.

The boundedness assumption is trivially satisfied when I is finite and holds in most applications with an infinite state space. Using very deep mathematics it can be shown that under Assumption 4.1.2 the infinitesimal transition rates determine a unique continuous-time Markov chain $\{X(t)\}$. This continuous-time Markov chain is precisely the Markov jump process constructed according to the above rules (a) and (b), where the leaving rates are given by $\nu_i = \sum_{j \neq i} q_{ij}$ and the p_{ij} by $p_{ij} = q_{ij}/\nu_i$. The continuous-time Markov chain $\{X(t)\}$ does indeed have the property

$$P\{X(t + \Delta t) = j \mid X(t) = i\} = \begin{cases} q_{ij}\,\Delta t + o(\Delta t), & j \neq i, \\ 1 - \nu_i\,\Delta t + o(\Delta t), & j = i. \end{cases} \tag{4.1.1}$$

It is noted that Assumption 4.1.2 implies that the constructed continuous-time Markov chain $\{X(t)\}$ automatically satisfies Assumption 4.1.1.

In solving specific problems it suffices to specify the infinitesimal transition rates q_{ij}. We now give two examples. In these examples the q_{ij} are determined as the result of the interaction of several elementary processes of the Poisson type. The q_{ij} are found by using the interpretation that $q_{ij}\,\Delta t$ represents the probability of making a transition to state j in the next Δt time units when the current state is i and Δt is very small.

Example 4.1.1 (continued) Inventory control for an inflammable product

The stochastic process $\{X(t), t \geq 0\}$ with $X(t)$ denoting the stock on hand at time t is a continuous-time Markov chain with state space $I = \{0, 1, \ldots, Q\}$. Its infinitesimal transition rates q_{ij} are the result of the interaction of the two independent Poisson processes for the demands and the replenishment opportunities. The q_{ij} are given by

$$q_{i,i-1} = \lambda \quad \text{for } i = 1, \ldots, Q,$$
$$q_{0Q} = \mu \quad \text{and the other } q_{ij} = 0.$$

To see this, note that for any state i with $i \geq 1$,

$$\begin{aligned} P\{X(t + \Delta t) = i - 1 \mid X(t) = i\} &= P\{\text{a demand occurs in } (t, t + \Delta t]\} + o(\Delta t) \\ &= \lambda \Delta t + o(\Delta t) \end{aligned}$$

*Conditions under which the infinitesimal parameters determine a unique continuous-time Markov chain are discussed in depth in Chung (1967).

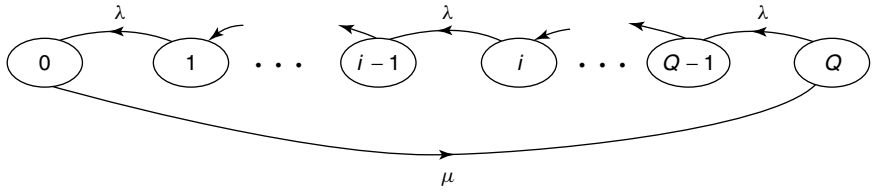


Figure 4.1.1 The transition rate diagram for the inventory process

and

$$\begin{aligned} P\{X(t+\Delta t) = Q \mid X(t) = 0\} &= P\{\text{a replenishment opportunity occurs in } (t, t+\Delta t]\} + o(\Delta t) \\ &= \mu \Delta t + o(\Delta t) \end{aligned}$$

for $\Delta t \to 0$. In the analysis of continuous-time Markov chains, it is very helpful to use a *transition rate diagram*. The nodes of the diagram represent the states and the arrows in the diagram give the possible state transitions. An arrow from node i to node j is only drawn when the transition rate q_{ij} is positive, in which case the arrow is labelled with the value q_{ij}. The transition rate diagram not only visualizes the process, but is particularly useful when writing down its equilibrium equations. Figure 4.1.1 shows the transition rate diagram for the inventory process.

Example 4.1.2 Unloading ships with an unreliable unloader

Ships arrive at a container terminal according to a Poisson process with rate λ. The ships bring loads of containers. There is a single unloader for unloading the ships. The unloader can handle only one ship at a time. The ships are unloaded in order of arrival. It is assumed that the dock has ample capacity for waiting ships. The unloading time of each ship has an exponential distribution with mean $1/\mu$. The unloader, however, is subject to breakdowns. A breakdown can only occur when the unloader is operating. The length of any operating period of the unloader has an exponential distribution with mean $1/\delta$. The time to repair a broken unloader is exponentially distributed with mean $1/\beta$. Any interrupted unloading of a ship is resumed at the point it was interrupted. It is assumed that the unloading times, operating times and repair times are independent of each other and are independent of the arrival process of the ships.

The average number of waiting ships, the fraction of time the unloader is down, and the average waiting time per ship, these and other quantities can be found by using the continuous-time Markov chain model. For any $t \geq 0$, define the random variables

$$X_1(t) = \text{the number of ships present at time } t$$

and

$$X_2(t) = \begin{cases} 1 & \text{if the unloader is available at time } t, \\ 0 & \text{if the unloader is in repair at time } t. \end{cases}$$

Since the underlying distributions are exponential, the process $\{(X_1(t), X_2(t))\}$ is a continuous-time Markov chain. This process has the state space

$$I = \{(i, 0) \mid i = 1, 2, \dots\} \cup \{(i, 1) \mid i = 0, 1, \dots\}.$$

The next step is to determine the infinitesimal transition rates of the process. Putting for abbreviation $X(t) = (X_1(t), X_2(t))$, we have

$$\begin{aligned} &P\{X(t + \Delta t) = (i, 1) \mid X(t) = (i, 0)\} \\ &\quad = P\{\text{the running repair is finished in } (t, t + \Delta t) \text{ and} \\ &\qquad \text{no arrival occurs in } (t, t + \Delta t)\} \\ &\quad = \beta \Delta t\,(1 - \lambda \Delta t) + o\,(\Delta t) = \beta \Delta t + o(\Delta t) \end{aligned}$$

for $\Delta t \to 0$. This gives

$$q_{(i,0)(i,1)} = \beta \quad \text{for } i = 1, 2, \dots\ .$$

Alternatively, $q_{(i,0)(i,1)}$ could have been obtained by noting that the sojourn time in state $(i, 0)$ is exponentially distributed with parameter $\beta + \lambda$ and noting that with probability $\beta/(\beta + \lambda)$ the running repair time is finished before an arrival occurs. Also,

$$\begin{aligned} &P\{X(t + \Delta t) = (i + 1, 0) | X(t) = (i, 0)\} \\ &\quad = P\{\text{an arrival occurs in } (t, t + \Delta t) \text{ and the running repair time} \\ &\qquad \text{is not finished in } (t, t + \Delta t)\} \\ &\quad = \lambda \Delta t(1 - \beta \Delta t) + o(\Delta t) = \lambda \Delta t + o(\Delta t) \end{aligned}$$

for $\Delta t \to 0$. This gives

$$q_{(i,0)(i+1,0)} = \lambda \quad \text{for } i \geq 1.$$

Similarly, we find

$$q_{(i,1)(i,0)} = \delta, \quad q_{(i,1)(i+1,1)} = \lambda \quad \text{and} \quad q_{(i,1)(i-1,1)} = \mu \quad \text{for } i \geq 1.$$

The state transitions and transition rates are summarized in Figure 4.1.2.

4.2 THE FLOW RATE EQUATION METHOD

This section discusses the flow rate equation method for obtaining the equilibrium distribution of a continuous-time Markov chain. The emphasis is to give insight

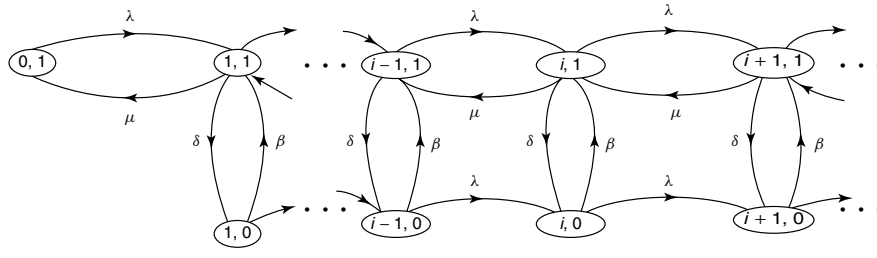


Figure 4.1.2 The transition rate diagram for the unloader

into this powerful method, which is widely used by practitioners. The proofs of the results below are deferred to Section 4.3.

The starting point is a continuous-time Markov chain $\{X(t)\}$ with state space I and infinitesimal transition rates q_{ij} for $i, j \in I$ with $j \neq i$. As before, let

$$\nu_i = \sum_{j \neq i} q_{ij}, \quad i \in I.$$

The quantity ν_i is the parameter of the exponentially distributed sojourn time in state i. It is assumed that the ν_i satisfy Assumption 4.1.2. For any $t \geq 0$, define the probability $p_{ij}(t)$ by

$$p_{ij}(t) = P\{X(t) = j \mid X(0) = i\}, \quad i, j \in I.$$

The computation of the transient probabilities $p_{ij}(t)$ will be discussed in Section 4.5. A deep result from continuous-time Markov chain theory is that $\lim_{t\to\infty} p_{ij}(t)$ always exists for all $i, j \in I$. The issue of possible periodicity in the state transitions is not relevant for continuous-time Markov chains, since the times between state transitions have a continuous distribution. To ensure that the limits of the $p_{ij}(t)$ are independent of the initial state i and constitute a probability distribution, we need the following assumption.

Assumption 4.2.1 *The process $\{X(t), t \geq 0\}$ has a regeneration state r such that*

$$P\{\tau_r < \infty \mid X(0) = i\} = 1 \quad \textit{for all} \quad i \in I \quad \textit{and} \quad E(\tau_r \mid X(0) = r) < \infty,$$

where τ_r is the first epoch beyond epoch 0 at which the process $\{X(t)\}$ makes a transition into state r.

In other words, state r will ultimately be reached with probability 1 from any other state and the mean recurrence time from state r to itself is finite. Under this assumption it can be proved that there is a probability distribution $\{p_j, j \in I\}$ such that

$$\lim_{t\to\infty} p_{ij}(t) = p_j, \quad j \in I,$$

independently of the initial state i. The interested reader is referred to Chung (1967) for a proof. The limiting probability p_j can be interpreted as the probability that an outside observer finds the system in state j when the process has reached statistical equilibrium and the observer has *no knowledge* about the past evolution of the process. The notion of statistical equilibrium relates not only to the length of time the process has been in operation but also to our knowledge of the past evolution of the system. But a more concrete interpretation which better serves our purposes is that

$$\begin{aligned} &\text{the long-run fraction of time the process will be in state } j \\ &\quad = p_j \quad \text{with probability 1,} \end{aligned} \tag{4.2.1}$$

independently of the initial state $X(0) = i$. More precisely, denoting for fixed j the indicator variable $I_j(t)$ by

$$I_j(t) = \begin{cases} 1 & \text{if } X(t) = j, \\ 0 & \text{otherwise,} \end{cases}$$

it holds for any $j \in I$ that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_j(u)\, du = p_j \quad \text{with probability 1,}$$

independently of the initial state $X(0) = i$. A proof of this result will be given in Section 4.3 using the theory of renewal-reward processes. In Section 4.3 we also prove the following important theorem.

Theorem 4.2.1 *Suppose the continuous-time Markov chain* $\{X(t)\}$ *satisfies Assumptions 4.1.2 and 4.2.1. Then the probabilities* p_j, $j \in I$ *are the unique solution to the linear equations*

$$v_j x_j = \sum_{k\neq j} q_{kj} x_k, \quad j \in I \tag{4.2.2}$$

$$\sum_{j\in I} x_j = 1 \tag{4.2.3}$$

in the unknowns x_j, $j \in I$. *Moreover, let* $\{x_j, j \in I\}$ *be any solution to (4.2.2) with* $\sum_j |x_j| < \infty$. *Then, for some constant* c, $x_j = cp_j$ *for all* $j \in I$.

The linear equations (4.2.2) are called the *equilibrium equations* or *balance equations* of the Markov process. The equation (4.2.3) is a *normalizing* equation. The probabilities p_j are called the *equilibrium probabilities* of the continuous-time Markov chain. They can be computed by solving a system of linear equations.

Interpretation of the equilibrium equations

A physical explanation of the equilibrium equations can be given by using the obvious principle that over the long run

$$\text{the average number of transitions out of state } j \text{ per time unit}$$
$$= \text{the average number of transitions into state } j \text{ per time unit.} \quad (4.2.4)$$

Since p_j is the long-run fraction of time the process is in state j and the leaving rate out of state j is ν_j, it is intuitively obvious that

$$\text{the long-run average number of transitions out of state } j$$
$$\text{per time unit} = \nu_j p_j. \quad (4.2.5)$$

Also, the following result will be intuitively obvious:

$$\text{the long-run average number of transitions from state } k \text{ to state } j$$
$$\text{per time unit} = q_{kj} p_k. \quad (4.2.6)$$

For a better understanding of (4.2.6), it is helpful to point out that q_{kj} can be interpreted as the long-run average number of transitions per time unit to state j when averaging over the time the process is in state k. A rigorous proof of the result (4.2.6) is given in Section 4.3. By (4.2.6),

$$\text{the long-run average number of transitions into state } j$$
$$\text{per time unit} = \sum_{k \neq j} q_{kj} p_k. \quad (4.2.7)$$

Together (4.2.4), (4.2.5) and (4.2.7) give the equilibrium equations (4.2.2). These equations may be abbreviated as

$$\text{rate out of state } j = \text{rate into state } j. \quad (4.2.8)$$

This principle is the *flow rate equation method*. To formulate the equilibrium equations in specific applications, it is convenient to use the transition rate diagram that was introduced in the previous section. Putting the transition rate diagram in a physical context, one might think that particles with a total mass of 1 are distributed over the nodes according to the equilibrium distribution $\{p_j\}$. Particles move from one node to another node according to the transition rates q_{ij}. In the equilibrium situation the rate at which particles leave any node must be equal to the rate at which particles enter that node. The 'rate in = rate out' principle (4.2.8) allows for a very useful generalization. More generally, for any set A of states with $A \neq I$,

$$\text{rate out of the set } A = \text{rate into the set } A. \quad (4.2.9)$$

In mathematical terms,

$$\sum_{j\in A} p_j \sum_{k\notin A} q_{jk} = \sum_{k\notin A} p_k \sum_{j\in A} q_{kj}.$$

The balance principle (4.2.9) enables us to write down a recursive equation for the p_j when

$$I = \{0, 1, \ldots, N\} \quad \text{and} \quad q_{ij} = 0 \text{ for } i \geq 1 \text{ and } j \leq i-2,$$

where $N \leq \infty$. Then, by taking $A = \{i, \ldots, N\}$ with $i \neq 0$ and applying the balance principle (4.2.9), we get

$$q_{i,i-1} p_i = \sum_{k=0}^{i-1} p_k \sum_{j=i}^{N} q_{kj}, \quad i = 1, \ldots, N. \tag{4.2.10}$$

This recursive relation is used quite often in queueing applications; see Chapter 5. In queueing applications it is often the case that direct transitions from any state i are either to higher states or to the state $i-1$ directly below state i. A recursive computation of the state probabilities is usually much faster than a computation by any other method. Also the recursion scheme (4.2.10) is numerically stable since it involves no subtractions.

Next we apply the flow rate equation method to the two examples discussed in the previous section.

Example 4.1.1 (continued) Inventory control for an inflammable product

In this example the equilibrium probability p_j represents the long-run fraction of time that the stock in the tank equals j units. Assumptions 4.1.2 and 4.2.1 are trivially satisfied (e.g. take state Q as regeneration state r). Using the transition rate diagram in Figure 4.1.1 and equating the rate at which the process leaves state i to the rate at which the process enters state i, it follows that

$$\begin{aligned} \mu p_0 &= \lambda p_1, \\ \lambda p_j &= \lambda p_{j+1}, \quad j = 1, 2, \ldots, Q-1, \\ \lambda p_Q &= \mu p_0. \end{aligned}$$

These equilibrium equations together with the equation $p_0 + p_1 + \cdots + p_Q = 1$ have a unique solution (in this special case an explicit solution can be given: $p_0 = (1 + Q\mu/\lambda)^{-1}$ and $p_1 = \cdots = p_Q = (\mu/\lambda)p_0$). Next we can answer the questions posed earlier:

$$\text{the long-run average stock on hand} = \sum_{j=0}^{Q} j p_j \tag{4.2.11}$$

$$\text{the long-run fraction of demand that is lost} = p_0. \tag{4.2.12}$$

A few words of explanation are in order. Intuitively, (4.2.11) may be obvious by noting that p_j gives the long-run fraction of time the stock on hand is j. The long-run average stock on hand is defined as $\lim_{t\to\infty}(1/t)\int_0^t X(u)\,du$. This long-run average can be seen as a long-run average cost per time unit by imagining that a cost at rate j is incurred when the stock on hand is j. Using this interpretation, the result (4.2.11) can be seen as a consequence of Theorem 4.2.2, which will be discussed below. The result (4.2.12) uses the PASTA property: in the long run the fraction of customers who find the system out of stock upon arrival equals the fraction of time the system is out of stock. Further, we have

$$\text{the long-run average number of stock replenishments per time unit} = \mu p_0.$$

This result follows from (4.2.6) by noting that the average replenishment frequency equals the average number of transitions from state 0 to state Q per time unit.

Example 4.1.2 (continued) Unloading ships with an unreliable unloader

In this example we need a regularity condition to ensure that Assumption 4.2.1 is satisfied (Assumption 4.1.2 trivially holds). Let γ denote the expected amount of time needed to complete the unloading of a ship. It is not difficult to verify that $\gamma = \mu^{-1}(1+\delta/\beta)$; see (A.5) in Appendix A. In order to satisfy Assumption 4.2.1 it should be required that the arrival rate of ships is less than the reciprocal of the expected completion time γ. That is, the assumption

$$\lambda < \frac{\beta\mu}{\beta+\delta}$$

should be made. The proof is omitted that under this condition the expected cycle length in Assumption 4.2.1 is finite (take state $(0,1)$ for the regeneration state r). Denote the equilibrium probabilities by $p(j,0)$ and $p(j,1)$. The probability $p(j,1)$ gives the long-run fraction of time that j ships are present and the unloader is available and the probability $p(j,0)$ gives the long-run fraction of time that j ships are present and the unloader is in repair. Using the transition rate diagram in Figure 4.1.2 and applying the 'rate in = rate out' principle, we obtain the equilibrium equations:

$$\begin{aligned}
\lambda p(0,1) &= \mu p(1,1),\\
(\lambda+\mu+\delta)p(i,1) &= \lambda p(i-1,1) + \mu p(i+1,1) + \beta p(i,0), \quad i = 1,2,\ldots,\\
(\lambda+\beta)p(1,0) &= \delta p(1,1),\\
(\lambda+\beta)p(i,0) &= \lambda p(i-1,0) + \delta p(i,1), \quad i = 2,3,\ldots .
\end{aligned}$$

This infinite system of linear equations together with the normalizing equation

$$\sum_{i=0}^{\infty} p(i,0) + \sum_{i=1}^{\infty} p(i,1) = 1$$

has a unique solution. A brute-force method for solving the equilibrium equations is to truncate this infinite system through a sufficiently large integer N (to be found by trial and error) such that $\sum_{i=N+1}^{\infty}[p(i,0)+p(i,1)] \leq \varepsilon$ for some prespecified accuracy number ε. In Section 4.4 we discuss a more sophisticated method to solve the infinite system of linear equations. Once the state probabilities have been computed, we find

$$\text{the long-run average number of ships in the harbour} = \sum_{i=1}^{\infty} i[p(i,0)+p(i,1)],$$

$$\text{the fraction of time the unloader is in repair} = \sum_{i=1}^{\infty} p(i,0),$$

$$\text{the long-run average amount of time spent in the harbour per ship} = \frac{1}{\lambda}\sum_{i=1}^{\infty} i[p(i,0)+p(i,1)].$$

The latter result uses Little's formula $L = \lambda W$.

Continuous-time Markov chains with rewards

In many applications a reward structure is imposed on the continuous-time Markov chain model. Let us assume the following reward structure. A reward at a rate of $r(j)$ per time unit is earned whenever the process is in state j, while a lump reward of F_{jk} is earned each time the process jumps from state j to state k ($\neq j$). In addition to Assumption 4.2.1 involving the regeneration state r, we make the following assumption.

Assumption 4.2.2 *(a) The total reward earned between two visits of the process* $\{X(t)\}$ *to state* r *has a finite expectation and*

$$\sum_{j\in I} |r(j)|\, p_j + \sum_{j\in I} p_j \sum_{k\neq j} q_{jk}|F_{jk}| < \infty.$$

(b) For each initial state $X(0) = i$ *with* $i \neq r$, *the total reward earned until the first visit of the process* $\{X(t)\}$ *to state* r *is finite with probability* 1.

This assumption is automatically satisfied when the state space I is finite and Assumption 4.2.1 holds. For each $t > 0$, define the random variable $R(t)$ by

$$R(t) = \text{the total reward earned up to time } t.$$

The following very useful result holds for the long-run average reward.

Theorem 4.2.2 *Suppose the continuous-time Markov chain* $\{X(t)\}$ *satisfies Assumptions 4.1.2, 4.2.1 and 4.2.2. Then, for each initial state* $X(0) = i$,

$$\lim_{t\to\infty} \frac{R(t)}{t} = \sum_{j\in I} r(j)p_j + \sum_{j\in I} p_j \sum_{k\neq j} q_{jk}F_{jk} \quad \textit{with probability } 1.$$

A proof of this ergodic theorem will be given in Section 4.3. Intuitively the theorem can be seen by noting that p_j gives the long-run fraction of time the process is in state j and $p_j q_{jk}$ gives the long-run average number of transitions from state j to state k per time unit.

Example 4.1.1 (continued) Inventory control for an inflammable product

Suppose that the following costs are made in the inventory model. For each unit kept in stock, a holding cost $h > 0$ is incurred for each unit of time the unit is kept in stock. Penalty costs $R > 0$ are incurred for each demand that is lost and fixed costs $K > 0$ are made for each inventory replenishment. Then the long-run average cost per time unit equals

$$h\sum_{j=0}^{Q} jp_j + R\lambda p_0 + K\mu p_0.$$

Strictly speaking, the cost term $R\lambda p_0$ is not covered by Theorem 4.2.2. Alternatively, by using part (a) of Theorem 2.4.1 it can be shown that the long-run average amount of demand that is lost per time unit equals λp_0.

4.3 ERGODIC THEOREMS

In this section we prove Theorems 4.2.1 and 4.2.2. The proofs rely heavily on earlier results for the discrete-time Markov chain model. In our analysis we need the embedded Markov chain $\{X_n, n = 0, 1, \dots\}$, where X_n is defined by

$$X_n = \text{the state of the continuous-time Markov chain just after the } n\text{th state transition}$$

with the convention that $X_0 = X(0)$. The one-step transition probabilities of the discrete-time Markov chain $\{X_n\}$ are given by

$$p_{ij} = \begin{cases} q_{ij}/v_i, & j \neq i, \\ 0, & j = i; \end{cases} \tag{4.3.1}$$

see Section 4.1. It is readily verified that Assumption 4.2.1 implies that the embedded Markov chain $\{X_n\}$ satisfies the corresponding Assumption 3.3.1 and thus state r is a positive recurrent state for the Markov chain $\{X_n\}$.

Definition 4.3.1 *A probability distribution $\{p_j,\ j \in I\}$ is said to be an equilibrium distribution for the continuous-time Markov chain $\{X(t)\}$ if*

$$\nu_j p_j = \sum_{k \neq j} p_k q_{kj}, \quad j \in I.$$

Just as in the discrete-time case, the explanation of the term 'equilibrium distribution' is as follows. If $P\{X(0) = j\} = p_j$ for all $j \in I$, then for any $t > 0$, $P\{X(t) = j\} = p_j$ for all $j \in I$. The proof is non-trivial and will not be given. Next we prove Theorem 4.2.1 in a somewhat more general setting.

Theorem 4.3.1 *Suppose that the continuous-time Markov chain $\{X(t)\}$ satisfies Assumptions 4.1.2 and 4.2.1. Then:*

(a) The continuous-time Markov chain $\{X(t)\}$ has a unique equilibrium distribution $\{p_j,\ j \in I\}$. Moreover

$$p_j = \frac{\pi_j/\nu_j}{\sum_{k \in I} \pi_k/\nu_k}, \quad j \in I, \tag{4.3.2}$$

where $\{\pi_j\}$ is the equilibrium distribution of the embedded Markov chain $\{X_n\}$.

(b) Let $\{x_j\}$ be any solution to $\nu_j x_j = \sum_{k \neq j} x_k q_{kj}$, $j \in I$, with $\sum_j |x_j| < \infty$. Then, for some constant c, $x_j = cp_j$ for all $j \in I$.

Proof We first verify that there is a one-to-one correspondence between the solutions of the two systems of linear equations

$$\nu_j x_j = \sum_{k \neq j} x_k q_{kj}, \quad j \in I$$

and

$$u_j = \sum_{k \in I} u_k p_{kj}, \quad j \in I.$$

If $\{u_j\}$ is a solution to the second system with $\sum |u_j| < \infty$, then $\{x_j = u_j/\nu_j\}$ is a solution to the first system with $\sum |x_j| < \infty$, and conversely. This is an immediate consequence of the definition (4.3.1) of the p_{ij}. The one-to-one correspondence and Theorem 3.5.9 imply the results of Theorem 4.3.1 provided we verify

$$\sum_{j \in I} \frac{\pi_j}{\nu_j} < \infty. \tag{4.3.3}$$

The proof that this condition holds is as follows. By Assumption 4.2.1, the process $\{X(t)\}$ regenerates itself each time the process makes a transition into state r. Let a

cycle be defined as the time elapsed between two consecutive visits of the process to state r. Using Wald's equation, it is readily seen that

$$E(\text{length of one cycle}) = \sum_{j \in I} E(\text{number of visits to state } j \text{ in one cycle}) \times \frac{1}{\nu_j}.$$

Thus, by Lemma 3.5.10,

$$E(\text{length of one cycle}) = \frac{1}{\pi_r} \sum_{j \in I} \frac{\pi_j}{\nu_j}.$$

Since E(length of one cycle) is finite by Assumption 4.2.1, the result now follows. This completes the proof.

Next it is not difficult to prove Theorem 4.2.2

Proof of Theorem 4.2.2 We first prove the result for initial state $X(0) = r$, where r is the regeneration state from Assumptions 4.2.1 and 4.2.2. The process $\{X(t)\}$ regenerates itself each time the process makes a transition into state r. Let a cycle be defined as the time elapsed between two consecutive visits of the process to state r. In the proof of the above theorem we have already shown

$$E(\text{length of one cycle}) = \frac{1}{\pi_r} \sum_{k \in I} \frac{\pi_k}{\nu_k}.$$

The expected length of a cycle is finite. Next apply the renewal-reward theorem from Chapter 2. This gives

$$\lim_{t \to \infty} \frac{R(t)}{t} = \frac{E(\text{reward earned during one cycle})}{E(\text{length of one cycle})} \tag{4.3.4}$$

with probability 1. Using Wald's equation, E(reward earned during one cycle) is

$$\sum_{j \in I} E(\text{number of visits to state } j \text{ during one cycle}) \times \left[\frac{r(j)}{\nu_j} + \sum_{k \neq j} p_{jk} F_{jk} \right].$$

Hence, by Lemma 3.5.10 and relation (4.3.1),

$$\begin{aligned} E(\text{reward earned during one cycle}) &= \sum_{j \in I} \frac{\pi_j}{\pi_r} \left[\frac{r(j)}{\nu_j} + \sum_{k \neq j} p_{jk} F_{jk} \right] \\ &= \frac{1}{\pi_r} \sum_{j \in I} \frac{\pi_j}{\nu_j} \left[r(j) + \sum_{k \neq j} q_{jk} F_{jk} \right]. \end{aligned}$$

Taking the ratio of the expressions for the expected reward earned during one cycle and the expected length of one cycle and using relation (4.3.2), we get the result

of Theorem 4.2.2 for initial state r. It remains to verify that the result also holds for any initial state $X(0) = i$ with $i \neq r$. This verification proceeds along the same lines as the proof of the corresponding result in Theorem 3.5.11.

By choosing an appropriate reward structure, Theorem 4.2.2 provides a rigorous proof of earlier interpretations we gave to the quantities p_j and $q_{jk}\, p_j$.

Corollary 4.3.2 *Suppose that the continuous-time Markov chain* $\{X(t)\}$ *satisfies Assumptions 4.1.2 and 4.2.1. Then*

(a) For each state $k \in I$, *the long-run fraction of time the process is in state* k *equals* p_k *with probability* 1, *independently of the initial state* $X(0) = i$.

(b) For all $j, k \in I$ *with* $j \neq k$, *the long-run average number of transitions from state* k *to state* j *per unit time equals* $p_k q_{kj}$ *with probability* 1, *independently of the initial state* $X(0) = i$.

4.4 MARKOV PROCESSES ON A SEMI-INFINITE STRIP*

Many practical (queueing) problems can be modelled as a continuous-time Markov chain $\{X(t)\}$ on a semi-infinite strip in the plane. That is, the Markov process has the two-dimensional state space

$$I = \{(i, s) \mid i = 0, 1, \ldots ;\ s = 0, 1, \ldots , m\} \tag{4.4.1}$$

for some finite positive integer m. Assuming that the continuous-time Markov chain $\{X(t)\}$ satisfies Assumption 4.2.1, denote its equilibrium probabilities by $p(i, s)$ for $i = 0, 1, \ldots$ and $s = 0, 1, \ldots , m$. These probabilities are determined by an *infinite* system of linear equations. In many cases, however, this infinite system can be reduced to a *finite* system of linear equations of *moderate* size. This can be done by using the *geometric tail approach*, discussed for discrete-time Markov chains in Section 3.4.2. Under rather general conditions the equilibrium probabilities $p(i, s)$ exhibit a geometric tail behaviour as $i \to \infty$, where the decay factor does not depend on s. That is, for constants $\gamma_s > 0$ and a constant η with $0 < \eta < 1$,

$$p(i, s) \sim \gamma_s \eta^i \quad \text{as } i \to \infty, \tag{4.4.2}$$

where the constant η does not depend on s. Then, for a sufficiently large choice of integer M, we have for each s that

$$\frac{p(i+1, s)}{p(i, s)} \approx \eta, \quad i \geq M,$$

or equivalently

$$p(i, s) \approx \eta^{i-M}\, p(M, s), \quad i > M.$$

*This section is more specialized and can be omitted at first reading.

Usually the constant η can be computed beforehand by solving a non-linear equation in a single variable. Once η is known, the infinite system of equilibrium equations is reduced to a finite system of linear equations by replacing any $p(i, s)$ with $i > M$ by $\eta^{i-M} p(M, s)$. It turns out that in practical applications a relatively small value of M usually suffices. As will be seen below, the asymptotic expansion (4.4.2) is valid in the unloader problem of Example 4.1.2.

Markov processes with quasi-birth-death rates

Suppose that the Markov process $\{X(t)\}$ satisfies the following assumption.

Assumption 4.4.1 *In state* (i, s) *the only possible transitions are:*

- *from state* (i, s) *to state* $(i + 1, s)$ *with rate* λ_s $(i = 0, 1, \ldots ; s = 0, 1, \ldots , m)$,
- *from state* (i, s) *to state* $(i - 1, s)$ *with rate* μ_s $(i = 1, 2, \ldots ; s = 0, 1, \ldots , m)$,
- *from state* (i, s) *to state* $(i, s + 1)$ *with rate* β_s $(i = 0, 1, \ldots ; s = 0, 1, \ldots , m - 1)$,
- *from state* (i, s) *to state* $(i, s - 1)$ *with rate* δ_s $(i = 0, 1, \ldots ; s = 1, 2, \ldots , m)$.

It is assumed that the transition rates λ_s, μ_s, β_s and δ_s are such that the Markov chain $\{X(t)\}$ satisfies Assumption 4.2.1 and thus has a unique equilibrium distribution $\{p(i, s)\}$. Under Assumption 4.4.1 the equilibrium equations for the continuous-time Markov chain $\{X(t)\}$ are as follows. Then for $i = 1, 2, \ldots$ and with $0 \leq s \leq m$,

$$(\lambda_s + \mu_s + \beta_s + \delta_s) p(i, s) = \lambda_s p(i - 1, s) + \mu_s p(i + 1, s) + \beta_{s-1} p(i, s - 1) + \delta_{s+1} p(i, s + 1) \quad (4.4.3)$$

provided we put $\beta_{-1} = \beta_m = \delta_0 = \delta_{m+1} = 0$ and define $p(i, -1) = p(i, m + 1) = 0$. For $i = 0$ and $0 \leq s \leq m$,

$$(\lambda_s + \beta_s + \delta_s) p(0, s) = \mu_s p(1, s) + \beta_{s-1} p(0, s - 1) + \delta_{s+1} p(0, s + 1). \quad (4.4.4)$$

Next we use the powerful tool of generating functions. Define for $0 \leq s \leq m$ the generating function $G_s(z)$ by

$$G_s(z) = \sum_{i=0}^{\infty} p(i, s) z^{i}, \quad |z| \leq 1.$$

For notational convenience, define $G_{-1}(z) = G_{m+1}(z) = 0$. Multiplying both sides of (4.4.3) and (4.4.4) by z^i and summing over i, we find for each s that

$$(\lambda_s + \mu_s + \beta_s + \delta_s) \sum_{i=0}^{\infty} p(i, s) z^i - \mu_s p(0, s)$$
$$= \lambda_s \sum_{i=1}^{\infty} p(i-1, s) z^i + \mu_s \sum_{i=0}^{\infty} p(i+1, s) z^i + \beta_{s-1} \sum_{i=0}^{\infty} p(i, s-1) z^i$$
$$+ \delta_{s+1} \sum_{i=0}^{\infty} p(i, s+1) z^i.$$

This gives for $s = 0, 1, \ldots, m$,

$$(\lambda_s + \mu_s + \beta_s + \delta_s) G_s(z) - \mu_s p(0, s) = \lambda_s z G_s(z) + \frac{\mu_s}{z} [G_s(z) - p(0, s)]$$
$$+ \beta_{s-1} G_{s-1}(z) + \delta_{s+1} G_{s+1}(z).$$

We rewrite this as

$$[\lambda_s z^2 + \mu_s - (\lambda_s + \mu_s + \beta_s + \delta_s) z] G_s(z) + \beta_{s-1} z G_{s-1}(z) + \delta_{s+1} z G_{s+1}(z)$$
$$= \mu_s (1 - z) p(0, s), \quad s = 0, 1, \ldots, m. \tag{4.4.5}$$

This is a system of linear equations in the unknowns $G_0(z), \ldots, G_m(z)$. To see what the solution looks like, it is convenient to write (4.4.5) in matrix notation. To do so, define the diagonal matrices $\mathbf{\Lambda}$ and $\mathbf{M}$ by

$$\mathbf{\Lambda} = \operatorname{diag}(\lambda_0, \lambda_1, \ldots, \lambda_m) \quad \text{and} \quad \mathbf{M} = \operatorname{diag}(\mu_0, \mu_1, \ldots, \mu_m).$$

Define the transition rate matrix $\mathbf{T} = (t_{ab})$ with $a, b = 0, 1, \ldots, m$ by

$$t_{s,s-1} = \beta_{s-1}, t_{s,s+1} = \delta_{s+1}, t_{ss} = -(\beta_s + \delta_s) \text{ and } t_{ab} = 0 \text{ otherwise.}$$

Finally, define the matrix $\mathbf{A}(z)$ by

$$\mathbf{A}(z) = \mathbf{\Lambda} z^2 - (\mathbf{\Lambda} - \mathbf{T} + \mathbf{M}) z + \mathbf{M}$$

and the column vectors $\mathbf{p}_0$ and $\mathbf{g}(z)$ by

$$\mathbf{p}_0 = (p(0, 0), \ldots, p(0, m)) \quad \text{and} \quad \mathbf{g}(z) = (G_0(z), \ldots, G_m(z)).$$

Then the linear equations (4.4.5) in matrix notation are

$$\mathbf{A}(z) \mathbf{g}(z) = (1 - z) \mathbf{M} \mathbf{p}_0 \tag{4.4.6}$$

By Cramer's rule for linear equations, the solution of (4.4.6) is given by

$$G_s(z) = \frac{\det \mathbf{A}_s(z)}{\det \mathbf{A}(z)}, \quad s = 0, 1, \ldots, m, \tag{4.4.7}$$

where $\mathbf{A}_s(z)$ is the matrix that results from replacing the $(s+1)$th column vector of $\mathbf{A}(z)$ by the vector $(1-z)\mathbf{M}\mathbf{p}_0$. The functions $\det \mathbf{A}_s(z)$ and $\det \mathbf{A}(z)$ are polynomials in z and are thus defined on the whole complex plane. Assuming that the function $N(z) = \det \mathbf{A}(z)$ satisfies the conditions stated in Theorem C.1 in Appendix C, the representation (4.4.7) implies the following result.

Theorem 4.4.1 *For each $s = 0, 1, \ldots, m$, there is a constant γ_s such that*

$$p(i,s) \sim \gamma_s \eta^i \quad \text{as } i \to \infty, \tag{4.4.8}$$

where η is the reciprocal of the smallest root of the equation

$$\det \mathbf{A}(x) = 0 \tag{4.4.9}$$

on the interval $(1, \infty)$.

How do we solve (4.4.9) in general? A possible way is to use a basic result from linear algebra stating that $\det \mathbf{A}(x)$ equals the product of the eigenvalues of the matrix $\mathbf{A}(x)$. It is a routine matter to determine the eigenvalues of a matrix by using standard software. A search procedure such as bisection can next be used to find the root of (4.4.9). Another approach to compute the roots of $\det \mathbf{A}(z) = 0$ was proposed in Chapter 3 of Daigle (1991). The zeros of $\det \mathbf{A}(z)$ are equivalent to the inverses of the eigenvalues of the $2(m+1)$-dimensional square matrix

$$\mathbf{A}_E = \begin{bmatrix} \mathbf{M}^{-1}(\mathbf{\Lambda} - \boldsymbol{T} + \boldsymbol{M}) & -\mathbf{M}^{-1}\mathbf{\Lambda} \\ \mathbf{I} & \mathbf{O} \end{bmatrix},$$

where $\mathbf{O}$ is the matrix with all entries equal to zero. Note that $\mathbf{M}^{-1}$ exists. To see this, let σ be any zero of $\det \mathbf{A}(z)$ and let $\mathbf{x}_\sigma$ be any non-trivial column vector such that $\mathbf{A}(\sigma)\mathbf{x}_\sigma = 0$. Let $\mathbf{y}_\sigma = \sigma \mathbf{x}_\sigma$. Then, by the definition of $\mathbf{A}(z)$, we have $\sigma^2 \mathbf{\Lambda} \boldsymbol{x}_\sigma - \sigma(\mathbf{\Lambda} - \boldsymbol{T} + \boldsymbol{M})\boldsymbol{x}_\sigma + \mathbf{M}\mathbf{x}_\sigma = 0$. By definition, $\mathbf{y}_\sigma - \sigma \mathbf{x}_\sigma = 0$. Combining these two systems gives

$$\left[\begin{pmatrix} \mathbf{M} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} - \sigma \begin{pmatrix} \mathbf{\Lambda} - \boldsymbol{T} + \boldsymbol{M} & -\mathbf{\Lambda} \\ \mathbf{I} & \mathbf{O} \end{pmatrix} \right] \begin{bmatrix} \mathbf{x}_\sigma \\ \mathbf{y}_\sigma \end{bmatrix} = \mathbf{0}.$$

This system is equivalent to

$$\left[\begin{pmatrix} \mathbf{M}^{-1}(\mathbf{\Lambda} - \boldsymbol{T} + \boldsymbol{M}) & -\mathbf{M}^{-1}\mathbf{\Lambda} \\ \mathbf{I} & \mathbf{O} \end{pmatrix} - \frac{1}{\sigma} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \right] \begin{bmatrix} \mathbf{x}_\sigma \\ \mathbf{y}_\sigma \end{bmatrix} = \mathbf{0}.$$

This proves that the zeros of $\det \mathbf{A}(z)$ are equivalent to the inverses of the eigenvalues of the expanded matrix $\mathbf{A}_E$. The largest of the eigenvalues in (0, 1) gives the decay factor η of the geometric tail of the equilibrium probabilities $p(i,s)$. Daigle (1991) gives a more sophisticated algorithm for the computation of the $p(i,s)$. Using the eigenvalues and the eigenvectors of the matrix $\mathbf{A}_E$, this algorithm computes for each s the probabilities $p(i,s)$ for $i \geq 1$ as a linear combination of a finite number of geometric terms. The interested reader is referred to Daigle's book

for details. The algorithm in Chapter 3 of Daigle (1991) is in fact a special case of the *spectral expansion method* discussed in full generality in Mitrani and Mitra (1992). This is a general method for computing the equilibrium probabilities of a Markov process whose state space is a semi-infinite strip in the two-dimensional plane and whose equilibrium equations can be represented by a vector difference equation with constant coefficients. The solution of that equation is expressed in terms of the eigenvalues and eigenvectors of its characteristic polynomial. Another generally applicable method to compute the equilibrium probabilities for the two-dimensional Markov process with quasi-birth-death rates is the *matrix-geometric method* of Neuts (1981). This method requires solving a matrix quadratic equation. This can be done by a probabilistic and numerically stable algorithm discussed in Latouche and Ramaswami (1993). The computational effort of this algorithm increases logarithmically when the server utilization gets larger. The computational burden of the spectral method, however, is relatively insensitive to the server utilization of the analysed system. Unlike the Latouche–Ramaswami algorithm, the spectral method often becomes numerically unreliable when the server utilization gets very close to 1. For the practitioner, the geometric tail approach is much easier to apply than the other two methods. This approach combines simplicity with effectiveness. The two steps of the geometric tail algorithm are:

(a) Compute the zero of a non-linear equation in a single variable.

(b) Solve a finite system of linear equations.

These steps are simple, and standard software can be used to perform them. The finite system of linear equations is usually relatively small for practical examples. In general it is not possible to use the above computational methods on two-dimensional continuous-time Markov chain problems in which both state variables are unbounded. An example of such a problem is the shortest-queue problem in which arriving customers are assigned to the shortest one of two separate queues each having their own server. Special methods for this type of problem are the so-called compensation method and the power-series algorithm discussed in Adan *et al.* (1993), Blanc (1992) and Hooghiemstra *et al.* (1988).

Example 4.1.2 (continued) Unloading ships with an unreliable unloader

The continuous-time Markov chain in the unloader problem satisfies Assumption 4.4.1 except that the Markov chain cannot take on state $(0, 0)$. The unloader can only break down when it is in operation. However, the assumption made in the foregoing analysis can be released somewhat. Assume that for some integer $N \geq 1$ the state space $I = I_1 \cup I_2$, where $I_1 = \{(i, s) \mid i = N, N+1, \ldots; s = 0, \ldots, m\}$ and I_2 is a non-empty subset of $\{(i, s) \mid i = 0, \ldots, N-1;\ s = 0, \ldots, m\}$. The conditions in Assumption 4.4.1 are only assumed for the states (i, s) with $i \geq N$. Further it must be assumed that the only way to enter the set I_1 from the set I_2 is through the states (N, s). Then a minor modification of the above analysis shows

that Theorem 4.4.1 remains valid with the same matrix $\mathbf{A}(z)$. For the particular case of the unloader problem, we find that (4.4.9) reduces to the polynomial equation

$$(\lambda + \beta - \lambda z)\,(\lambda z^2 + \mu - (\lambda + \mu + \delta)z) + \delta\beta z = 0.$$

Special case of linear birth-death rates

Suppose that the transition rates λ_s, μ_s, β_s and δ_s have the special form

$$\lambda_s = b_1 \times (m - s) + c_1 s, \quad \mu_s = b_{-1} \times (m - s) + c_{-1} s$$
$$\beta_s = a_0 \times (m - s) \quad \text{and } \delta_s = d_0 s \tag{4.4.10}$$

for constants a_0, b_1, b_{-1}, c_1, c_{-1} and d_0. Then the numerical problem of computing the roots of $\det \mathbf{A}(z) = 0$ can be circumvented. The decay factor η in (4.4.2) is then the unique solution of the equation

$$B(x) + C(x) - x[A(1) + B(1) + C(1) + D(1)] + \sqrt{F(x)^2 + 4A(x)D(x)} = 0$$

on the interval (0,1), where

$$A(x) = a_0 x, \;\; B(x) = b_1 + b_{-1}x^2, \;\; C(x) = c_1 + c_{-1}x^2, \;\; D(x) = d_0 x,$$
$$F(x) = [A(1) + B(1) - C(1) - D(1)]x + C(x) - B(x).$$

In a more general context this result has been proved in Adan and Resing (1999). It also follows from this reference that Assumption 4.2.1 holds when $d_0(b_{-1} - b_1) + a_0(c_{-1} - c_1) > 0$. The condition (4.4.10) is satisfied for several interesting queueing models. For example, take a queueing model with m traffic sources which act independently of each other. Each traffic source is alternately on and off, where the on-times and off-times have exponential distributions with respective means $1/\delta$ and $1/\beta$. The successive on- and off-times are assumed to be independent of each other. During on-periods a source generates service requests according to a Poisson process with rate λ. There is a single server to handle the service requests and the server can handle only one request at a time. The service times of the requests are independent random variables that have a common exponential distribution with mean $1/\mu$. This queueing problem can be modelled as a continuous-time Markov chain whose state space is given by (4.4.1) with i denoting the number of service requests in the system and s denoting the number of sources that are on. This Markov chain has the property (4.4.10) with $\lambda_s = \lambda s$, $\mu_s = \mu$, $\beta_s = \beta \times (m - s)$ and $\delta_s = \delta s$.

4.5 TRANSIENT STATE PROBABILITIES

In many practical situations one is not interested in the long-run behaviour of a stochastic system but in its transient behaviour. A typical example concerns airport runway operations. The demand profile for runway operations shows considerable

variation over time with peaks at certain hours of the day. Equilibrium models are of no use in this kind of situation. The computation of transient solutions for Markov systems is a very important issue that arises in numerous problems in queueing, inventory and reliability. In this section we discuss two basic methods for the computation of the transient state probabilities of a continuous-time Markov chain. The next section deals with the computation of the transient distribution of the cumulative reward in a continuous-time Markov chain with a reward structure.

The transient probabilities of a continuous-time Markov chain $\{X(t), t \geq 0\}$ are defined by

$$p_{ij}(t) = P\{X(t) = j \mid X(0) = i\}, \quad i, j \in I \text{ and } t > 0.$$

In Section 4.5.1 we discuss the method of linear differential equations. The probabilistic method of uniformization will be discussed in Section 4.5.2. In Section 4.5.3 we show that the computation of first passage time probabilities can be reduced to the computation of transient state probabilities by introducing an absorbing state.

4.5.1 The Method of Linear Differential Equations

This basic approach has a solid backing by tools from numerical analysis. We first prove the following theorem.

Theorem 4.5.1 (Kolmogoroff's forward differential equations) *Suppose that the continuous-time Markov chain* $\{X(t), t \geq 0\}$ *satisfies Assumption 4.1.2. Then for any* $i \in I$,

$$p'_{ij}(t) = \sum_{k \neq j} q_{kj}\, p_{ik}(t) - v_j\, p_{ij}(t), \quad j \in I \text{ and } t > 0. \tag{4.5.1}$$

Proof We sketch the proof only for the case of a finite state space I. The proof of the validity of the forward equations for the case of an infinite state space is very intricate. Fix $i \in I$ and $t > 0$. Let us consider what may happen in $(t, t + \Delta t]$ with Δt very small. The number of transitions in any finite time interval is finite with probability 1, so we can condition on the state that will occur at time t:

$$\begin{aligned}
p_{ij}(t + \Delta t) &= P\{X(t + \Delta t) = j \mid X(0) = i\} \\
&= \sum_{k \in I} P\{X(t + \Delta t) = j \mid X(0) = i,\ X(t) = k\} \\
&\quad \times P\{X(t) = k \mid X(0) = i\} \\
&= \sum_{k \in I} P\{X(t + \Delta t) = j \mid X(t) = k\} p_{ik}(t) \\
&= \sum_{k \neq j} q_{kj}\, \Delta t\, p_{ik}(t) + (1 - v_j \Delta t)\, p_{ij}(t) + o(\Delta t),
\end{aligned}$$

since a finite sum of $o(\Delta t)$ terms is again $o(\Delta t)$. Hence

$$\frac{p_{ij}(t+\Delta t)-p_{ij}(t)}{\Delta t}=\sum_{k\neq j} q_{kj}\,p_{ik}(t)-\nu_j\,p_{ij}(t)+\frac{o(\Delta t)}{\Delta t}.$$

Letting $\Delta t \to 0$ yields the desired result.

The linear differential equations (4.5.1) can be explicitly solved only in very special cases.

Example 4.5.1 An on-off source

A source submitting messages is alternately on and off. The on-times are independent random variables having a common exponential distribution with mean $1/\alpha$ and the off-times are independent random variables having a common exponential distribution with mean $1/\beta$. Also the on-times and the off-times are independent of each other. The source is on at time 0. What is the probability that the source will be off at time t?

This system can be modelled as a continuous-time Markov chain with two states. Let the random variable $X(t)$ be equal to 0 when the source is off at time t and equal to 1 otherwise. Then $\{X(t)\}$ is a continuous-time Markov chain with state space $I=\{0,1\}$. The transient probabilities $p_{10}(t)$ and $p_{11}(t)$ satisfy

$$p'_{10}(t)=-\beta p_{10}(t)+\alpha p_{11}(t), \quad t\geq 0,$$
$$p'_{11}(t)=\beta p_{10}(t)-\alpha p_{11}(t), \quad t\geq 0.$$

A standard result from the theory of linear differential equations states that the general solution of this system is given by

$$(p_{10}(t),\,p_{11}(t))=c_1e^{\lambda_1 t}\mathbf{x}_1+c_2e^{\lambda_2 t}\mathbf{x}_2, \quad t\geq 0$$

provided that the coefficient matrix

$$\mathbf{A}=\begin{pmatrix}-\beta & \alpha\\ \beta & -\alpha\end{pmatrix}$$

has distinct eigenvalues λ_1 and λ_2. The vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are the corresponding eigenvectors. The constants c_1 and c_2 are determined by the boundary conditions $p_{10}(0)=0$ and $p_{11}(0)=1$. The eigenvalues of the matrix $\mathbf{A}$ are $\lambda_1=0$ and $\lambda_2=-(\alpha+\beta)$ with corresponding eigenvectors $\mathbf{x}_1=(\beta^{-1},\alpha^{-1})$ and $\mathbf{x}_2=(1,-1)$. Next it follows from $c_1/\beta+c_2=0$ and $c_1/\alpha-c_2=1$ that $c_1=\alpha\beta/(\alpha+\beta)$ and $c_2=-\alpha/(\alpha+\beta)$. This yields

$$p_{10}(t)=\frac{\alpha}{\alpha+\beta}-\frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t}, \quad t\geq 0.$$

Example 4.5.2 Transient analysis for the $M/M/1$ queue

In the $M/M/1$ queue customers arrive according to a Poisson process with rate λ and the service times of the customers have an exponential distribution with mean $1/\mu$. Letting $X(t)$ denote the number of customers present at time t, the process $\{X(t)\}$ is a continuous-time Markov chain. Kolmogoroff's forward differential equations are as follows for the $M/M/1$ queue:

$$p'_{ij}(t) = \mu p_{i,j+1}(t) + \lambda p_{i,j-1}(t) - (\lambda + \mu) p_{ij}(t), \quad i, j = 0, 1, \ldots \text{ and } t > 0$$

with $p_{i,-1}(t) = 0$. An explicit solution of these equations is given by

$$p_{ij}(t) = \frac{2}{\pi}\rho^{(j-i)/2} \int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma(y)} a_i(y) a_j(y)\, dy + \begin{cases} (1-\rho)\rho^j, & \rho < 1, \\ 0, & \rho \geq 1, \end{cases}$$

for $i, j = 0, 1, \ldots$, where $\rho = \lambda/\mu$ and the functions $\gamma(y)$ and $a_k(y)$ are defined by

$$\gamma(y) = 1 + \rho - 2\sqrt{\rho}\cos(y) \quad \text{and} \quad a_k(y) = \sin(ky) - \sqrt{\rho}\sin[(k+1)y].$$

A proof of this explicit solution is not given here; see Morse (1955) and Takács (1962). The trigonometric integral representation for $p_{ij}(t)$ is very convenient for numerical computations. A recommended numerical integration method is Gauss–Legendre integration. Integral representations can also be given for the first two moments of the number of customers in the system. The formulas will only be given for the case of $\rho < 1$. Denoting by $L(i, t)$ the number of customers in the system at time t when initially there are i customers present, we have

$$E[L(i,t)] = \frac{2}{\pi}\rho^{(1-i)/2} \int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma^2(y)} a_i(y) \sin(y)\, dy + \frac{\rho}{1-\rho}$$

and

$$E[L^2(i,t)] = \frac{4(1-\rho)}{\pi}\rho^{(1-i)/2} \int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma^3(y)} a_i(y) \sin(y)\, dy + 2\rho(1-\rho)^{-2} - E[L(i,t)],$$

assuming that $\rho < 1$. If $\rho < 1$, the transient probabilities $p_{ij}(t)$ converge to the equilibrium probabilities $p_j = (1-\rho)\rho^j$ as $t \to \infty$ and, similarly, $E[L(i,t)]$ converges to $\rho/(1-\rho)$ as $t \to \infty$. A natural question is how fast the convergence occurs. A heuristic answer to this question has been given by Odoni and Roth (1983) in the context of the $M/G/1$ queue. Letting c_S^2 denote the squared coefficient of variation of the service time, the $M/G/1$ queue will 'forget' its initial state after a time comparable to

$$t_{relax} = \frac{(1 + c_S^2)E(S)}{2.8(1 - \sqrt{\rho})^2}$$

provided that the system is empty at epoch 0.

In general the linear differential equations (4.5.1) have to be solved numerically. Let us assume in the remainder of this section that the state space I of the Markov chain is finite. There are several possibilities to numerically solve the homogeneous system (4.5.1) of linear differential equations with constant coefficients. In most applications the matrix of coefficients has distinct eigenvalues and is thus diagonalizable. In those situations one might compute the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix and the corresponding eigenvectors. The transient probabilities $p_{ij}(t)$ are then a linear combination of pure exponential functions $e^{\lambda_1 t}, \ldots, e^{\lambda_n t}$ (zero is always among the eigenvalues and the corresponding eigenvector gives the equilibrium distribution of the Markov process up to a multiplicative constant). In general, however, one uses a so-called Runge–Kutta method to solve the linear differential equations numerically. Standard codes for this method are widely available. From a numerical point of view, the Runge–Kutta method is in general superior to the eigenvalue approach. The Runge–Kutta method has the additional advantage that it can also be applied to continuous-time Markov processes with *time-dependent* transition rates. Another possible way to compute the $p_{ij}(t)$ is to use the discrete FFT method when an explicit expression for the generating function $P(t, z) = \sum_{j \in I} p_{ij}(t) z^j$, $|z| \leq 1$ is available.

4.5.2 The Uniformization Method

This method falls back on an earlier construction of a continuous-time Markov chain in Section 4.1. In this construction the process leaves state i after an exponentially distributed time with mean $1/\nu_i$ and then jumps to another state $j (j \neq i)$ with probability p_{ij}. Letting X_n denote the state of the process just after the nth state transition, the discrete-time stochastic process $\{X_n\}$ is an embedded Markov chain with one-step transition probabilities p_{ij}. The jump probabilities p_{ij} and the infinitesimal transition rates q_{ij} are related to each other by

$$q_{ij} = \nu_i p_{ij}, \quad i, j \in I \text{ with } j \neq i. \tag{4.5.2}$$

To introduce the uniformization method, consider first the special case in which the leaving rates ν_i of the states are identical, say $\nu_i = \nu$ for all i. Then the transition epochs are generated by a Poisson process with rate ν. In this situation an expression for $p_{ij}(t)$ is directly obtained by conditioning on the number of Poisson events up to time t and using the n-step transition probabilities of the embedded Markov chain $\{X_n\}$. However, the leaving rates ν_i are in general not identical. Fortunately, there is a simple trick for reducing the case of non-identical leaving rates to the case of identical leaving rates. The uniformization method transforms the original continuous-time Markov chain with non-identical leaving rates into an equivalent stochastic process in which the transition epochs are generated by a Poisson process at a *uniform* rate. However, to achieve this, the discrete-time Markov chain describing the state transitions in the transformed process has to allow for self-transitions leaving the state of the process unchanged.

To formulate the uniformization method, choose a finite number ν with

$$\nu \geq \nu_i, \quad i \in I.$$

Define now $\{\overline{X}_n\}$ as the discrete-time Markov chain whose one-step transition probabilities $\overline{p}_{ij}$ are given by

$$\overline{p}_{ij} = \begin{cases} (\nu_i/\nu)p_{ij}, & j \neq i, \\ 1 - \nu_i/\nu, & j = i, \end{cases}$$

for any $i \in I$. Let $\{N(t), t \geq 0\}$ be a Poisson process with rate ν such that the process is independent of the discrete-time Markov chain $\{\overline{X}_n\}$. Define now the continuous-time stochastic process $\{\overline{X}(t), t \geq 0\}$ by

$$\overline{X}(t) = \overline{X}_{N(t)}, \quad t \geq 0. \tag{4.5.3}$$

In other words, the process $\{\overline{X}(t)\}$ makes state transitions at epochs generated by a Poisson process with rate ν and the state transitions are governed by the discrete-time Markov chain $\{\overline{X}_n\}$ with one-step transition probabilities $\overline{p}_{ij}$. Each time the Markov chain $\{\overline{X}_n\}$ is in state i, the next transition is the same as in the Markov chain $\{X_n\}$ with probability ν_i/ν and is a self-transition with probability $1 - \nu_i/\nu$. The transitions out of state i are in fact delayed by a time factor of ν/ν_i, while the time itself until a state transition from state i is condensed by a factor of ν_i/ν. This heuristically explains why the continuous-time process $\{\overline{X}(t)\}$ is probabilistically identical to the original continuous-time Markov chain $\{X(t)\}$. Another heuristic way to see that the two processes are identical is as follows. For any $i, j \in I$ with $j \neq i$

$$\begin{aligned} P\{\overline{X}(t + \Delta t) = j \mid \overline{X}(t) = i\} &= \nu \Delta t \times \overline{p}_{ij} + o(\Delta t) \\ &= \nu_i \Delta t \times p_{ij} + o(\Delta t) = q_{ij} \Delta t + o(\Delta t) \\ &= P\{X(t + \Delta t) = j \mid X(t) = i\} \quad \text{for } \Delta t \to 0. \end{aligned}$$

In the next theorem we give a formal proof that the two processes $\{X(t)\}$ and $\{\overline{X}(t)\}$ are probabilistically equivalent.

Theorem 4.5.2 *Suppose that the continuous-time Markov chain $\{X(t)\}$ satisfies Assumption 4.1.2. Then*

$$p_{ij}(t) = P\{\overline{X}(t) = j \mid \overline{X}(0) = i\}, \quad i, j \in I \text{ and } t > 0.$$

Proof For any $t > 0$, define the matrix $\mathbf{P}(t)$ by $\mathbf{P}(t) = (p_{ij}(t))$, $i, j \in I$. Denote by $\mathbf{Q}$ the matrix $\mathbf{Q} = (q_{ij})$, $i, j \in I$, where the diagonal elements q_{ii} are defined by

$$q_{ii} = -\nu_i.$$

Then Kolmogoroff's forward differential equations can be written as $\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}$ for any $t > 0$. It is left to the reader to verify that the solution of this system of differential equations is given by

$$\mathbf{P}(t) = e^{t\mathbf{Q}} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbf{Q}^n, \quad t \geq 0. \tag{4.5.4}$$

The matrix $\overline{\mathbf{P}} = (\overline{p}_{ij})$, $i, j \in I$, can be written as $\overline{\mathbf{P}} = \mathbf{Q}/\nu + \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus

$$\mathbf{P}(t) = e^{t\mathbf{Q}} = e^{\nu t(\overline{\mathbf{P}}-\mathbf{I})} = e^{\nu t\overline{\mathbf{P}}}e^{-\nu t\mathbf{I}} = e^{-\nu t}e^{\nu t\overline{\mathbf{P}}} = \sum_{n=0}^{\infty} e^{-\nu t}\frac{(\nu t)^n}{n!}\overline{\mathbf{P}}^n.$$

On the other hand, by conditioning on the number of Poisson events up to time t in the $\{\overline{X}(t)\}$ process, we have

$$P\{\overline{X}(t) = j \mid \overline{X}(0) = i\} = \sum_{n=0}^{\infty} e^{-\nu t}\frac{(\nu t)^n}{n!}\overline{p}_{ij}^{(n)},$$

where $\overline{p}_{ij}^{(n)}$ is the n-step transition probability of the discrete-time Markov chain $\{\overline{X}_n\}$. Together the latter two equations yield the desired result.

Corollary 4.5.3 *The probabilities $p_{ij}(t)$ are given by*

$$p_{ij}(t) = \sum_{n=0}^{\infty} e^{-\nu t}\frac{(\nu t)^n}{n!}\overline{p}_{ij}^{(n)}, \quad i, j \in I \text{ and } t > 0, \tag{4.5.5}$$

where the probabilities $\overline{p}_{ij}^{(n)}$ can be recursively computed from

$$\overline{p}_{ij}^{(n)} = \sum_{k\in I} \overline{p}_{ik}^{(n-1)}\overline{p}_{kj}, \quad n = 1, 2, \ldots \tag{4.5.6}$$

starting with $\overline{p}_{ii}^{(0)} = 1$ and $\overline{p}_{ij}^{(0)} = 0$ for $j \neq i$.

This probabilistic result is extremely useful for computational purposes. The series in (4.5.5) converges much faster than the series expansion (4.5.4). The computations required by (4.5.5) are simple and transparent. For fixed $t > 0$ the infinite series can be truncated beforehand, since

$$\sum_{n=M}^{\infty} e^{-\nu t}\frac{(\nu t)^n}{n!}\overline{p}_{ij}^{(n)} \leq \sum_{n=M}^{\infty} e^{-\nu t}\frac{(\nu t)^n}{n!}.$$

For a prespecified accuracy number $\varepsilon > 0$, we choose M such that the right-hand side of this inequality is smaller than ε. By the normal approximation to the Poisson

distribution, the truncation integer M can be chosen as

$$M = \nu t + c\sqrt{\nu t}$$

for some constant c with $0 < c \leq c_0(\varepsilon)$, where $c_0(\varepsilon)$ depends only on the tolerance number ε. As a consequence the computational complexity of the uniformization method is $O(\nu t N^2)$ where N is the number of states of the Markov chain. Hence the uniformization method should be applied with the choice

$$\nu = \max_{i \in I} \nu_i .$$

The number νt is a crucial factor for the computational complexity of the uniformization method, as it is for the Runge–Kutta method, and is called the index of stiffness. Also, the following remark may be helpful. For fixed initial state i, the recursion scheme (4.5.6) boils down to the multiplication of a vector with the matrix $\overline{\mathbf{P}}$. In many applications the matrix $\overline{\mathbf{P}}$ is sparse. Then the computational effort can be considerably reduced by using a data structure for sparse matrix multiplications. The uniformization results (4.5.5) and (4.5.6) need only a minor modification when the initial state $X(0)$ has a given probability distribution $\{\pi_0(i), i \in I\}$. The probability $\overline{p}_{ij}^{(n)}$ should then be replaced by $\overline{p}_j^{(n)} = \sum_{i \in I} \pi_0(i) \overline{p}_{ij}^{(n)}$ and this probability can recursively be computed from $\overline{p}_j^{(n)} = \sum_{k \in I} \overline{p}_k^{(n-1)} p_{kj}$ starting with $\overline{p}_j^{(0)} = \pi_0(j)$ for $j \in I$. For example, this modification may be used to compute the transient state probabilities in finite-capacity queueing systems with a nonstationary Poisson arrival process and exponential services, where the arrival rate function $\lambda(t)$ is (approximately) a piecewise-constant function. One then computes the transient state probabilities for each interval separately on which $\lambda(t)$ is constant and uses the probability distribution of the state at the beginning of the interval as the distribution of the initial state.

Expected transient rewards

Assume that a reward at rate $r(j)$ is earned whenever the continuous-time Markov chain $\{X(t)\}$ is in state j, while a lump reward of F_{jk} is earned each time the process makes a transition from state j to state k $(\neq j)$. Let

$$R(t) = \text{the total reward earned up to time } t, \quad t \geq 0.$$

The following lemma shows that it is a simple matter to compute the expected value of the reward variable $R(t)$. The computation of the probability distribution of $R(t)$ is much more complex and will be addressed in Section 4.6.

Lemma 4.5.4 *Suppose that the continuous-time Markov chain* $\{X(t)\}$ *satisfies Assumption 4.1.2. Then*

$$E[R(t) \mid X(0) = i] = \sum_{j \in I} r(j) E_{ij}(t) + \sum_{j \in I} E_{ij}(t) \sum_{k \neq j} q_{jk} F_{jk}, \quad t > 0, \qquad (4.5.7)$$

where $E_{ij}(t)$ is the expected amount of time that the process $\{X(t)\}$ is in state j up to time t when the process starts in state i. For any $i, j \in I$,

$$E_{ij}(t) = \frac{1}{\nu} \sum_{k=1}^{\infty} e^{-\nu t} \frac{(\nu t)^k}{k!} \sum_{n=0}^{k-1} \overline{p}_{ij}^{(n)}, \quad t > 0. \tag{4.5.8}$$

Proof The first term on the right-hand side of the relation for the expected reward is obvious. To explain the second term, we use the PASTA property. Fix $j, k \in I$ with $k \neq j$. Observe that the transitions out of state j occur according to a Poisson process with rate ν_j whenever the process $\{X(t)\}$ is in state j. Hence, using part (b) of Theorem 1.1.3, transitions from state j to state $k(\neq j)$ occur according to a Poisson process with rate q_{jk} $(= p_{jk}\nu_j)$ whenever the process $\{X(t)\}$ is in state j. Next, by applying part (a) of Theorem 2.4.1, it is readily seen that the expected number of transitions from state j to state k up to time t equals q_{jk} times the expected amount of time the process $\{X(t)\}$ is in state j up to time t. This proves (4.5.7). To prove (4.5.8), note that the representation (4.5.5) implies

$$\begin{aligned} E_{ij}(t) &= \int_0^t p_{ij}(u)\, du = \int_0^t \left[\sum_{n=0}^{\infty} e^{-\nu u} \frac{(\nu u)^n}{n!} \overline{p}_{ij}^{(n)} \right] du \\ &= \sum_{n=0}^{\infty} \overline{p}_{ij}^{(n)} \int_0^t e^{-\nu u} \frac{(\nu u)^n}{n!}\, du. \end{aligned}$$

Except for a factor ν we have an integral over an Erlang $(n+1, \nu)$ density. Thus

$$E_{ij}(t) = \frac{1}{\nu} \sum_{n=0}^{\infty} \overline{p}_{ij}^{(n)} \sum_{k=n+1}^{\infty} e^{-\nu t} \frac{(\nu t)^k}{k!}.$$

By interchanging the order of summation, we next get the desired result.

4.5.3 First Passage Time Probabilities

In this section it is assumed that the state space I of the continuous-time Markov chain $\{X(t)\}$ is finite. For a given set C of states with $C \neq I$, define

$$\tau_C = \text{the first epoch at which the continuous-time Markov chain } \{X(t)\} \text{ makes a transition into a state of the set } C.$$

Also, define the first passage time probability $Q_{iC}(t)$ by

$$Q_{iC}(t) = P\{\tau_C > t \mid X(0) = i\}, \quad i \notin C \text{ and } t > 0.$$

The computation of the first passage time probabilities $Q_{iC}(t)$ can be reduced to the computation of the transient state probabilities in a modified continuous-time Markov chain with an *absorbing state*. The most convenient way to model an

absorbing state is to take its leaving rate equal to zero. In the modified continuous-time Markov chain the set C is replaced by a single absorbing state to be denoted by a. The state space I^* and the leaving rates v_i^* in the modified continuous-time Markov chain are taken as

$$I^* = (I \backslash C) \cup \{a\} \quad \text{and} \quad v_i^* = \begin{cases} v_i, & i \in I \backslash C, \\ 0, & i = a. \end{cases}$$

The infinitesimal transition rates q_{ij}^* are taken as

$$q_{ij}^* = \begin{cases} q_{ij}, & i, j \in I \backslash C \text{ with } j \neq i, \\ \sum_{k \in C} q_{ik}, & i \in I \backslash C, \ j = a, \\ 0, & i = a, \ j \in I \backslash C. \end{cases}$$

Denoting by $p_{ij}^*(t)$ the transient state probabilities in the modified continuous-time Markov chain, it is readily seen that

$$Q_{iC}(t) = 1 - p_{ia}^*(t), \quad i \notin C \text{ and } t \geq 0.$$

The $p_{ij}^*(t)$ can be computed by using the uniformization algorithm in the previous subsection (note that $\overline{p}_{aa}^* = 1$ in the uniformization algorithm).

Example 4.5.3 The Hubble space telescope

The Hubble space telescope is an astronomical observatory in space. It carries a variety of instruments, including six gyroscopes to ensure stability of the telescope. The six gyroscopes are arranged in such a way that any three gyroscopes can keep the telescope operating with full accuracy. The operating times of the gyroscopes are independent of each other and have an exponential distribution with failure rate λ. Upon a fourth gyroscope failure, the telescope goes into sleep mode. In sleep mode, further observations by the telescope are suspended. It requires an exponential time with mean $1/\mu$ to put the telescope into sleep mode. Once the telescope is in sleep mode, the base station on earth receives a sleep signal. A shuttle mission to the telescope is next prepared. It takes an exponential time with mean $1/\eta$ before the repair crew arrives at the telescope and has repaired the stabilizing unit with the gyroscopes. In the meantime the other two gyroscopes may fail. If the last gyroscope fails, a crash destroying the telescope will be inevitable. What is the probability that the telescope will crash in the next T years?

This problem can be analysed by a continuous-time Markov chain with an absorbing state. The transition diagram is given in Figure 4.5.1. The state labelled as the crash state is the absorbing state. As explained above, this convention enables us to apply the uniformization method for the state probabilities to compute the first passage time probability

$$Q(T) = P\{\text{no crash will occur in the next } T \text{ years when currently all six gyroscopes are working}\}.$$

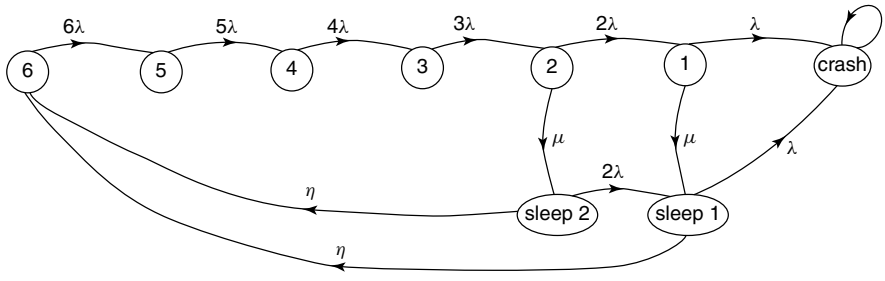


Figure 4.5.1 The transition rate diagram for the telescope

Taking one year as time unit, consider the numerical example with the data

$$\lambda = 0.1,\ \mu = 100 \text{ and } \eta = 5.$$

The uniformization method is applied with the choice $\nu = 100$ for the uniformized leaving rate ν (the value 0 is taken for the leaving rate from the state crash). The calculations yield the value 0.000504 for the probability $1 - Q(T)$ that a crash will occur in the next $T = 10$ years. Similarly, one can calculate that with probability 0.3901 the sleep mode will not be reached in the next 10 years. In other words, the probability of no shuttle mission in the next 10 years equals 0.3901. However, if one wishes to calculate the probability distribution of the number of required shuttle missions in the next 10 years, one must use the Markov reward model with lump rewards (assume a lump reward of 1 each time the process jumps from either state 2 or state 1 to the sleep mode). This Markov reward model is much more difficult to solve and will be discussed in the next section.

4.6 TRANSIENT DISTRIBUTION OF CUMULATIVE REWARDS

A basic and practically important problem is the calculation of the transient probability distribution of the cumulative reward in a continuous-time Markov chain. For example, a practical application is an oil-production platform which has to meet a contractually agreed production level over a specified time period. The production rate is not constant but depends on the stochastically varying state of the platform. In this example a continuous-time Markov chain with reward rates may be appropriate, where the reward rate $r(j)$ represents the production rate in state j.

In Section 4.6.1 we first consider the special case of a Markov reward model with reward rates that are either 0 or 1. The cumulative reward in this model corresponds to the cumulative sojourn time of the process in a certain set of (good) states. A nice and simple extension of the uniformization method can be given to compute the transient distribution of the cumulative sojourn time in some given set of states. The general Markov reward model with both reward rates and lump rewards is dealt with in Section 4.6.2. A discretization algorithm will be discussed

for this model. Throughout this section it is assumed that the continuous-time Markov chain $\{X(t)\}$ has a finite state space I.

4.6.1 Transient Distribution of Cumulative Sojourn Times

Consider a continuous-time Markov chain $\{X(t)\}$ whose finite state space I is divided into two disjoint subsets I_0 and I_f with

$$I_0 = \text{the set of operational states} \quad \text{and} \quad I_f = \text{the set of failed states.}$$

Define for any $t > 0$ the indicator random variable $I(t)$ by

$$I(t) = \begin{cases} 1 & \text{if } X(t) \in I_0, \\ 0 & \text{otherwise.} \end{cases}$$

Then the random variable

$$O(t) = \int_0^t I(u)\, du$$

represents the cumulative operational time of the system during $(0, t)$. The transient probability distribution of $O(t)$ can be calculated by a nice probabilistic algorithm that is based on the uniformization method.

To find $P\{O(t) \leq x\}$, we first uniformize the continuous-time Markov chain $\{X(t)\}$ according to (4.5.3). Denoting by $\overline{O}(t)$ the cumulative operational time in the uniformized process $\{\overline{X}(t)\}$, we have

$$P\{O(t) \leq x\} = P\{\overline{O}(t) \leq x\},$$

since the uniformized process is probabilistically equivalent with the original process. By conditioning on the number of state transitions of the uniformized process during $(0, t)$, we have

$$P\{O(t) \leq x\} = \sum_{n=0}^{\infty} P\{\overline{O}(t) \leq x \mid \text{ the uniformized process makes } n \text{ state transitions in } (0, t)\}\, e^{-\nu t} \frac{(\nu t)^n}{n!}.$$

The next key step in the analysis is the relation between the Poisson process and the uniform distribution. In the uniformized process the epochs of the state transitions are determined by a Poisson process that is independent of the discrete-time Markov chain governing the state transitions. Under the condition that the uniformized process has made n state transitions during $(0, t)$, the joint distribution of the epochs of these state transitions is the same as the joint distribution of the order statistics $U^{(1)}, \ldots, U^{(n)}$ of n independent random variables $U_1, \ldots, U_n$ that are uniformly distributed on $(0, t)$; see Theorem 1.1.5. Note that $U^{(k)}$ is the smallest

kth of the U_i. The n state transitions in the interval $(0, t)$ divide this interval into $n + 1$ intervals whose lengths are given by

$$Y_1 = U^{(1)},\ Y_2 = U^{(2)} - U^{(1)}, \ldots, Y_n = U^{(n)} - U^{(n-1)} \text{ and } Y_{n+1} = t - U^{(n)}.$$

The random variables $Y_1, \ldots, Y_{n+1}$ are obviously dependent variables, but they are exchangeable. That is, for any permutation $i_1, \ldots, i_{n+1}$ of $1, \ldots, n + 1$,

$$P\{Y_{i_1} \le x_1, Y_{i_2} \le x_2, \ldots, Y_{i_{n+1}} \le x_{n+1}\} = P\{Y_1 \le x_1, Y_2 \le x_2, \ldots, Y_{n+1} \le x_{n+1}\}.$$

As a consequence

$$P\{Y_{i_1} + \cdots + Y_{i_k} \le x\} = P\{Y_1 + \cdots + Y_k \le x\}$$

for any sequence $(Y_{i_1}, \ldots, Y_{i_k})$ of k interval lengths. The probability distribution of $Y_1 + \cdots + Y_k$ is easily given. Let $k \le n$. Then $Y_1 + \cdots + Y_k = U^{(k)}$ and so

$$\begin{aligned} P\{Y_1 + \cdots + Y_k \le x\} &= P\{U^{(k)} \le x\} = P\{\text{at least } k \text{ of the } U_i \text{ are } \le x\} \\ &= \sum_{j=k}^{n} \binom{n}{j} \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{n-j}. \end{aligned}$$

The next step of the analysis is to condition on the number of times the uniformized process visits operational states during $(0, t)$ given that the process makes n state transitions in $(0, t)$. If this number of visits equals k ($k \le n+1$), then the cumulative operational time during $(0, t)$ is distributed as $Y_1 + \cdots + Y_k$. For any given $n \ge 0$, define

$$\begin{aligned} \alpha(n, k) = P\{&\text{the uniformized process visits } k \text{ times an operational} \\ &\text{state in } (0, t) \mid \text{the uniformized process makes } n \\ &\text{state transitions in } (0, t)\} \end{aligned}$$

for $k = 0, 1, \ldots, n + 1$. Before showing how to calculate the $\alpha(n, k)$, we give the final expression for $P\{O(t) \le x\}$. Note that $O(t)$ has a positive mass at $x = t$. Choose $x < t$. Using the definition of $\alpha(n, k)$ and noting that $\overline{O}(t) \le x$ only if the uniformized process visits at least one non-operational state in $(0, t)$, it follows that

$$\begin{aligned} &P\{\overline{O}(t) \le x \mid \text{the uniformized processes makes } n \text{ state transitions in } (0, t)\} \\ &\quad = \sum_{k=0}^{n} P\{\overline{O}(t) \le x \mid \text{the uniformized process makes } n \text{ state transitions} \\ &\qquad \text{in } (0, t) \text{ and visits } k \text{ times an operational state in } (0, t)\}\ \alpha(n, k) \end{aligned}$$

$$= \sum_{k=0}^{n} P\{Y_1 + \cdots + Y_k \le x\}\, \alpha(n,k)$$

$$= \sum_{k=0}^{n} \alpha(n,k) \sum_{j=k}^{n} \binom{n}{j} \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{n-j}, \quad 0 \le x < t.$$

This gives the desired expression

$$P\{O(t) \le x\} = \sum_{n=0}^{\infty} e^{-\nu t} \frac{(\nu t)^n}{n!} \sum_{k=0}^{n} \alpha(n,k) \sum_{j=k}^{n} \binom{n}{j} \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{n-j} \tag{4.6.1}$$

for $0 \le x < t$. The random variable $\overline{O}(t)$ assumes the value t if the uniformized process visits only operational states in $(0, t)$. Thus

$$P\{O(t) = t\} = \sum_{n=0}^{\infty} \alpha(n, n+1) e^{-\nu t} \frac{(\nu t)^n}{n!}.$$

The above expression for $P\{O(t) \le x\}$ is well suited for numerical computations, since the summations involve only positive terms. As before, the infinite sum can be truncated to M terms, where the error associated with the truncation is bounded by $\sum_{n=M}^{\infty} e^{-\nu t}(\nu t)^n/n!$ so that M can be determined beforehand for a given error tolerance.

Computation of the $\alpha(n,k)$

The probabilities $\alpha(n,k)$ are determined by the discrete-time Markov chain $\{\overline{X}_n\}$ that governs the state transitions in the uniformized process. The one-step transition probabilities of this discrete-time Markov chain are given by $\overline{p}_{ij} = (\nu_i/\nu)p_{ij}$ for $j \ne i$ and $\overline{p}_{ii} = 1 - \nu_i/\nu$, where $p_{ij} = q_{ij}/\nu_i$. To calculate the $\alpha(n,k)$, let $\alpha(n,k,j)$ be the joint probability that the discrete-time Markov chain $\{\overline{X}_t\}$ visits k times an operational state over the first n state transitions and is in state j after the nth transition. Then

$$\alpha(n,k) = \sum_{j \in I} \alpha(n,k,j), \quad k = 0, 1, \ldots, n+1 \text{ and } n = 0, 1, \ldots .$$

The probabilities $\alpha(n,k,j)$ can be recursively computed. In the recursion we have to distinguish between states $j \in I_0$ and states $j \in I_f$. Obviously,

$$\alpha(n,k,j) = \sum_{i \in I} \alpha(n-1, k-1, i)\overline{p}_{ij}, \quad j \in I_0$$

and

$$\alpha(n,k,j) = \sum_{i \in I} \alpha(n-1, k, i)\overline{p}_{ij}, \quad j \in I_f.$$

Denoting by $\{\alpha_i\}$ the probability distribution of the initial state of the original process $\{X(t)\}$, we have the boundary conditions

$$\alpha(0, 1, j) = \alpha_j, \ \alpha(0, 0, j) = 0, \quad j \in I_0$$

and

$$\alpha(0, 0, j) = \alpha_j, \ \alpha(0, 1, j) = 0, \quad j \in I_f.$$

Example 4.5.3 (continued) The Hubble telescope problem

Assume that the telescope is needed to make observations of important astronomical events during a period of half a year two years from now. What is the probability that during this period of half a year the telescope will be available for at least 95% of the time when currently all six gyroscopes are in perfect condition? The telescope is only working properly when three or more gyroscopes are working. In states 1 and 2 the telescope produces blurred observations and in states sleep 2, sleep 1 and crash the telescope produces no observations at all. Let us number the states sleep 2, sleep 1 and crash as the states 7, 8 and 9. To answer the question posed, we split the state space $I = \{1, 2, \ldots, 9\}$ into the set I_0 of operational states and the set I_f of failed states with

$$I_0 = \{6, 5, 4, 3\} \text{ and } I_f = \{2, 1, 7, 8, 9\}.$$

Before applying the algorithm (4.6.1) with $t = \frac{1}{2}$ and $x = 0.95t$, we first use the standard uniformization method from Section 4.5 to compute the probability distribution of the state of the telescope two years from now. Writing $\alpha_i = p_{6i}(2)$, we obtain the values

$$\alpha_1 = 3.83 \times 10^{-7}, \ \alpha_2 = 0.0001938, \ \alpha_3 = 0.0654032, \ \alpha_4 = 0.2216998,$$
$$\alpha_5 = 0.4016008, \ \alpha_6 = 0.3079701, \ \alpha_7 = 0.0030271, \ \alpha_8 = 0.0000998,$$
$$\alpha_9 = 0.0000050$$

for the data $\lambda = 0.1$, $\mu = 100$ and $\eta = 5$. Next the algorithm (4.6.1) leads to the value 0.9065 for the probability that the telescope will be properly working for at least 95% of the time in the half-year that comes two years from now.

4.6.2 Transient Reward Distribution for the General Case

In the general case the continuous-time Markov chain $\{X(t)\}$ earns a reward at rate $r(j)$ for each unit of time the process is in state j and earns a lump reward of F_{jk} each time the process makes a state transition from state j to another state k. It is assumed that the $r(j)$and the F_{jk} are both non-negative. It is possible to extend the algorithm from Section 4.6.1 to the general case. However, the generalized algorithm is very complicated and, worse, it is not numerically stable. For this

reason we prefer to present a simple-minded discretization approach for the general reward case. For fixed $t > 0$, let

$$R(t) = \text{the cumulative reward earned up to time } t.$$

Assume that for each state $j \in I$ the joint probability distribution function $P\{R(t) \leq x, X(t) = j\}$ has a density with respect to the reward variable x (a sufficient condition is that $r(j) > 0$ for all $j \in I$). Then we can represent $P\{R(t) \leq x\}$ as

$$P\{R(t) \leq x\} = \sum_{j \in I} \int_0^x f_j(t, y)\, dy, \quad x \geq 0,$$

where $f_j(t, x)$ is the joint probability density of the cumulative reward up to time t and the state of the process at time t. The idea is to discretize the reward variable x and the time variable t in multiples of Δ, where $\Delta > 0$ is chosen sufficiently small (the probability of more than one state transition in a time period of length Δ should be negligibly small). The discretized reward variable x can be restricted to multiples of Δ when the following assumptions are made:

(a) the reward rates $r(j)$ are non-negative integers,

(b) the non-negative lump rates F_{jk} are multiples of Δ.

For practical applications it is no restriction to make these assumptions. How do we compute $P\{R(t) \leq x\}$ for fixed t and x? It is convenient to assume a probability distribution

$$\alpha_i = P\{X(0) = i\}, \quad i \in I$$

for the initial state of the process. In view of the probabilistic interpretation

$$f_j(t, x)\Delta x \approx P\{x \leq R(t) < x + \Delta x,\ X(t) = j\} \quad \text{for } \Delta x \text{ small},$$

we approximate for fixed $\Delta > 0$ the density $f_j(u, y)$ by a discretized function $f_j^{\Delta}(\tau, r)$. The discretized variables τ and r run through multiples of Δ. For fixed $\Delta > 0$ the discretized functions $f_j^{\Delta}(\tau, r)$ are defined by the recursion scheme

$$\begin{aligned} f_j^{\Delta}(\tau, r) = {} & f_j^{\Delta}(\tau - \Delta, r - r(j)\Delta)(1 - \nu_j \Delta) \\ & + \sum_{k \neq j} f_k^{\Delta}(\tau - \Delta, r - r(k)\Delta - F_{kj}) q_{kj} \Delta \end{aligned}$$

for $\tau = 0, \Delta, \ldots, (t/\Delta)\,\Delta$ and $r = 0, \Delta, \ldots, (x/\Delta)\,\Delta$ (for ease assume that x and t are multiples of Δ). For any $j \in I$, the boundary conditions are

$$f_j^{\Delta}(0, r) = \begin{cases} \alpha_j/\Delta, & r = 0, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$f_j^{\Delta}(\tau, r) = 0 \quad \text{for any } \tau \geq 0 \text{ when } r < 0.$$

Using the simple-minded approximation

$$\int_0^x f_j(t, y)\, dy \approx \sum_{\ell=0}^{x/\Delta-1} f_j^{\Delta}(t, \ell\Delta)\Delta,$$

the desired probability $P\{R(t) \leq x\}$ is approximated by

$$P\{R(t) \leq x\} \approx \sum_{j \in I} \sum_{\ell=0}^{x/\Delta-1} f_j^{\Delta}(t, \ell\Delta)\Delta. \tag{4.6.2}$$

For fixed x and t, the computational effort of the algorithm is proportional to $1/\Delta^2$ and so it quadruples when Δ is halved. Hence the computation time of the algorithm will become very large when the probability $P\{R(t) \leq x\}$ is desired at high accuracy and there are many states. Another drawback of the discretization algorithm is that no estimate is available for the discretization error. Fortunately, both difficulties can be partially overcome. Let

$$P(\Delta) = \sum_{j \in I} \sum_{\ell=0}^{x/\Delta-1} f_j^{\Delta}(t, \ell\Delta)\Delta$$

be the first-order estimate for $P\{R(t) \leq x\}$ and let the error term

$$e(\Delta) = P(\Delta) - P\{R(t) \leq x\}.$$

The following remarkable result was empirically found:

$$e(\Delta) \approx P(2\Delta) - P(\Delta)$$

when Δ is not too large. Thus the first-order approximation $P(\Delta)$ to $P\{R(t) \leq x\}$ is much improved when it is replaced by

$$\widetilde{P}(\Delta) = P(\Delta) - [P(2\Delta) - P(\Delta)]. \tag{4.6.3}$$

Example 4.5.3 (continued) The Hubble telescope problem

What is the probability distribution of the number of repair missions that will be prepared in the next 10 years when currently all six gyroscopes are in perfect condition? To consider this question we impose the following reward structure on the continuous-time Markov chain that is described in Figure 4.5.1 (with the states sleep 2 and sleep 1 numbered as the states 7 and 8). The reward rates $r(j)$ and the lump rewards F_{jk} are taken as

$$r(j) = 0 \text{ for all } j, \quad F_{27} = F_{18} = 1 \quad \text{and the other } F_{jk} = 0.$$

Then the cumulative reward variable $R(t)$ represents the number of repair missions that will be prepared up to time t. Note that in this particular case the stochastic variable $R(t)$ has a discrete distribution rather than a continuous distribution. However, the discretization algorithm also applies to the case of a reward variable $R(t)$ with a non-continuous distribution. For the numerical example with $\lambda = 0.1$, $\mu = 100$ and $\eta = 5$ we found that $P\{R(t) > k\}$ has the respective values 0.6099, 0.0636 and 0.0012 for $k = 0$, 1 and 2 (accurate to four decimal places with $\Delta = 1/256$).

EXERCISES

4.1 A familiar sight in Middle East street scenes are the so-called sheroots. A sheroot is a seven-seat cab that drives from a fixed stand in a town to another town. A sheroot leaves as soon as all seven seats are occupied by passengers. Consider a sheroot stand which has room for only one sheroot. Potential passengers arrive at the stand according to a Poisson process at rate λ. If upon arrival a potential customer finds no sheroot present and seven other customers already waiting, the customer goes elsewhere for transport; otherwise, the customer waits until a sheroot departs. After a sheroot leaves the stand, it takes an exponential time with mean $1/\mu$ until a new sheroot becomes available.

Formulate a continuous-time Markov chain model for the situation at the sheroot stand. Specify the state variable(s) and the transition rate diagram.

4.2 In a certain city there are two emergency units, 1 and 2, that cooperate in responding to accident alarms. The alarms come into a central dispatcher who sends one emergency unit to each alarm. The city is divided in two districts, 1 and 2. The emergency unit i is the first-due unit for response area i for $i = 1, 2$. An alarm coming in when only one of the emergency units is available is handled by the idle unit. If both units are not available, the alarm is settled by some unit from outside the city. Alarms from the districts 1 and 2 arrive at the central dispatcher according to independent Poisson processes with respective rates λ_1 and λ_2. The amount of time needed to serve an alarm from district j by unit i has an exponential distribution with mean $1/\mu_{ij}$. The service times include travel times.

Formulate a continuous-time Markov chain model to analyse the availability of the emergency units. Specify the state variable(s) and the transition rate diagram.

4.3 An assembly line for a certain product has two stations in series. Each station has only room for a single unit of the product. If the assembly of a unit is completed at station 1, it is forwarded immediately to station 2 provided station 2 is idle; otherwise the unit remains in station 1 until station 2 becomes free. Units for assembly arrive at station 1 according to a Poisson process with rate λ, but a newly arriving unit is only accepted by station 1 when no other unit is present in station 1. Each unit rejected is handled elsewhere. The assembly times at stations 1 and 2 are exponentially distributed with respective means $1/\mu_1$ and $1/\mu_2$.

Formulate a continuous-time Markov chain to analyse the situation at both stations. Specify the state variable(s) and the transition rate diagram.

4.4 Cars arrive at a gasoline station according to a Poisson process with an average of 10 customers per hour. A car enters the station only if less than four other cars are present. The gasoline station has only one pump. The amount of time required to serve a car has an exponential distribution with a mean of four minutes.

(a) Formulate a continuous-time Markov chain to analyse the situation of the gasoline station. Specify the state diagram.

(b) Solve the equilibrium equations.

(c) What is the long-run average number of cars in the station?
(d) What is the long-run fraction of potential customers that are lost?

4.5 A production hall contains a fast machine and a slow machine to process incoming orders. Orders arrive according to a Poisson process with rate λ. An arriving order that finds both machines occupied is rejected. Unless both machines are occupied, an arriving order is assigned to the fast machine if available; otherwise, the order is assigned to the slow machine. The processing time of an order is exponentially distributed with mean $1/\mu_1$ at the fast machine and mean $1/\mu_2$ at the slow machine. It is not possible to transfer an order from the slow machine to the fast machine.

(a) Formulate a continuous-time Markov chain to analyse the situation in the production hall. Specify the state variable(s) and the transition rate diagram

(b) Specify the equilibrium equations for the state probabilities. What is the long-run fraction of time that the fast (slow) machine is used? What is the long-run fraction of incoming orders that are lost?

4.6 In Gotham City there is a one-man taxi company. The taxi company has a stand at the railway station. Potential customers arrive according to a Poisson process with an average of four customers per hour. The taxi leaves the station immediately a customer arrives. A potential customer finding no taxi present waits until the taxi arrives only if there are less than three other customers waiting; otherwise, the customer goes elsewhere for alternative transport. If the taxi returns to the stand and finds waiting customers, it picks up all waiting customers and leaves. The amount of time needed to return to the stand has an exponential distribution with mean $1/\mu_i$ when the taxi leaves the stand with i customers, $i = 1, 2, 3$.

(a) Formulate a continuous-time Markov chain to analyse the situation at the taxi stand. Specify the state variable(s) and the transition rate diagram.

(b) What is the long-run fraction of time the taxi waits idle at the taxi stand? What is the long-run fraction of potential customers who go elsewhere for transport?

4.7 A container terminal has a single unloader to unload trailers which bring loads of containers. The unloader can serve only one trailer at a time and the unloading time has an exponential distribution with mean $1/\mu_1$. After a trailer has been unloaded, the trailer leaves but the unloader needs an extra finishing time for the unloaded containers before the unloader is available to unload another trailer. The finishing time has an exponential distribution with mean $1/\mu_2$. A leaving trailer returns with the next load of containers after an exponentially distributed time with mean $1/\lambda$. There are a finite number of N unloaders active at the terminal.

(a) Formulate a continuous-time Markov chain to analyse the situation at the container terminal. Specify the state variable(s) and the transition rate diagram.

(b) What is the long-run fraction of time the unloader is idle? What is the long-run average number of trailers unloaded per time unit?

(c) What is the long-run average number of trailers waiting to be unloaded? What is the long-run average waiting time per trailer?

(d) Write a computer program to compute the performance measures in (b) and (c) for the numerical data $N = 10$, $\mu_1 = 1/3$, $\mu_2 = 2$ and $\lambda = 1/50$.

4.8 Messages for transmission arrive at a communication channel according to a Poisson process with rate λ. The channel can transmit only one message at a time. The transmission time is exponentially distributed with mean $1/\mu$. The following access control rule is used. A newly arriving message is accepted as long as less than R other messages are present at the communication channel (including any message in transmission). As soon as the number of messages in the system has dropped to r, newly arriving messages are again admitted to the transmission channel. The control parameters r and R are given integers with $0 \leq r < R$.

(a) Formulate a continuous-time Markov chain to analyse the situation at the communication channel. Specify the state variable(s) and the transition rate diagram.

(b) What is the long-run fraction of time the channel is idle? What is the long-run fraction of messages that are rejected?

(c) What is the long-run average number of messages waiting to be transmitted? What is the long-run average delay in queue per accepted message?

4.9 An information centre has one attendant; people with questions arrive according to a Poisson process with rate λ. A person who finds n other customers present upon arrival joins the queue with probability $1/(n+1)$ for $n = 0, 1, \ldots$ and goes elsewhere otherwise. The service times of the persons are independent random variables having an exponential distribution with mean $1/\mu$.

(a) Verify that the equilibrium distribution of the number of persons present has a Poisson distribution with mean λ/μ.

(b) What is the long-run fraction of persons with requests who actually join the queue? What is the long-run average number of persons served per time unit?

4.10 (a) Consider Exercise 4.1 again. Specify the equilibrium equations for the state probabilities. What is the long-run average waiting time of a carried passenger? What is the long-run fraction of potential customers who are lost?

(b) Answer the questions in (a) again for the modified situation in which a potential customer only waits when, upon his arrival, a sheroot is present.

4.11 Consider Exercise 4.2 again and denote by S_{ij} the time needed to serve an alarm for district j by unit i. Assume that S_{ij} has a Coxian-2 distribution for all i, j. Show how to calculate the following performance measures: π_L = the fraction of alarms that is lost and P_i = the fraction of time that unit i is busy for $i = 1, 2$. Letting m_{ij} and c_{ij}^2 denote the mean and the squared coefficient of variation of S_{ij}, assume the numerical data $\lambda_1 = 0.25$, $\lambda_2 = 0.25$, $m_{11} = 0.75$, $m_{12} = 1.25$, $m_{21} = 1.25$ and $m_{22} = 1$. Write a computer program to verify the following numerical results:

(i) $\pi_L = 0.0704$, $P_1 = 0.2006$, $P_2 = 0.2326$ when $c_{ij}^2 = \frac{1}{2}$ for all i, j;

(ii) $\pi_L = 0.0708$, $P_1 = 0.2004$, $P_2 = 0.2324$ when $c_{ij}^2 = 1$ for all i, j;

(iii) $\pi_L = 0.0718$, $P_1 = 0.2001$, $P_2 = 0.2321$ when $c_{ij}^2 = 4$ for all i, j.

Here the values $c_{ij}^2 = \frac{1}{2}$, 1 and 4 correspond to the E_2 distribution, the exponential distribution and the H_2 distribution with balanced means.

4.12 In an inventory system for a single product the depletion of stock is due to demand and deterioration. The demand process for the product is a Poisson process with rate λ. The lifetime of each unit product is exponentially distributed with mean $1/\mu$. The stock control is exercised as follows. Each time the stock drops to zero an order for Q units is placed. The lead time of each order is negligible. Determine the average stock and the average number of orders placed per time unit.

4.13 Messages arrive at a communication channel according to a Poisson process with rate λ. The message length is exponentially distributed with mean $1/\mu$. An arriving message finding the line idle is provided with service immediately; otherwise the message waits until access to the line can be given. The communication line is only able to submit one message at a time, but has available two possible transmission rates σ_1 and σ_2 with $0 < \sigma_1 < \sigma_2$. Thus the transmission time of a message is exponentially distributed with mean $1/(\sigma_i \mu)$ when the transmission rate σ_i is used. It is assumed that $\lambda/(\sigma_2 \mu) < 1$. At any time the transmission line may switch from one rate to the other. The transmission rate is controlled by a rule that uses a single critical number. The transmission rate σ_1 is used whenever less than R messages are present, otherwise the faster transmission rate σ_2 is used. The following costs are involved. There is a holding cost at rate hj whenever there are j messages in the system. An operating cost at rate $r_i > 0$ is incurred when the line is transmitting a message using rate σ_i, while an operating cost at rate $r_0 \geq 0$ is incurred when the line is idle.

(a) Derive a recursion scheme for computing the limiting distribution of the number of messages present and give an expression for the long-run average cost per time unit.

(b) Write a computer program for calculating the value of R which minimizes the average cost and solve for the numerical data $\lambda = 0.8$, $\mu = 1$, $\sigma_1 = 1$, $\sigma_2 = 1.5$, $h = 1$, $r_0 = 0$, $r_1 = 5$ and $r_2 = 25$.

4.14 Customers asking for a certain product arrive according to a Poisson process with rate λ. The demand sizes of the customers are independent random variables and have a common discrete probability distribution $\{p_k, k = 1, 2, \dots\}$. Any demand that cannot be directly satisfied from stock on hand is back ordered. The control rule is based on the inventory position, which is defined as the stock on hand minus the amount back ordered plus the amount on order. Each time the inventory position reaches the reorder level s or drops below it, the smallest multiple of the basic quantity Q is ordered to bring the inventory position level above s. The lead time of any replenishment order is a fixed constant $L > 0$.

(a) Prove that the limiting distribution of the inventory position is a discrete uniform distribution. (*Hint*: use relation (4.3.2) and verify that the one-step transition matrix of the embedded Markov chain is doubly stochastic.)

(b) Derive the limiting distribution of the stock on hand.

(c) What is the average replenishment frequency and what is the average stock on hand?

(d) What is the fraction of customers whose demands are (partially) back ordered? What is the fraction of demand that is not satisfied directly from stock on hand?

4.15 Consider the transient probabilities $p_{ij}(t)$ in a continuous-time Markov chain with finite space $I = \{1, \dots, n\}$. Let the $n \times n$ matrix $\mathbf{Q}$ be defined as in the proof of Theorem 4.5.2. Assume that the matrix $\mathbf{Q}$ has n different eigenvalues $\lambda_1, \dots, \lambda_n$. Let $\mathbf{a}_k$ be an eigenvector corresponding to the eigenvalue λ_k for $k = 1, \dots, n$ and let $\mathbf{S}$ be the $n \times n$ matrix whose kth column vector is $\mathbf{a}_k$. For each initial state i, denote by $\mathbf{p}_i(t)$ the vector whose jth element equals $p_{ij}(t)$. Use results from Section 1.4 to verify the representation

$$\mathbf{p}_i(t) = \sum_{k=1}^{n} c_{ik} e^{\lambda_k t} \mathbf{a}_k, \quad t \geq 0,$$

for constants $c_{i1}, \dots, c_{in}$, where the vector $\mathbf{c}_i = (c_{i1}, \dots, c_{in})$ is given by $\mathbf{c}_i = \mathbf{S}^{-1}\mathbf{e}_i$ with $\mathbf{e}_i$ denoting the ith unit vector $(0, \dots, 1, \dots, 0)$.

4.16 An operating system has $r + s$ identical units where r units must be operating and s units are in preoperation (warm standby). A unit in operation has a constant failure rate of λ, while a unit in preoperation has a constant failure rate of β with $\beta < \lambda$. Failed units enter a repair facility that is able to repair at most c units simultaneously. The repair of a failed unit has an exponential distribution with mean $1/\mu$. An operating unit that fails is replaced immediately by a unit from the warm standby if one is available. The operating system goes down when less than r units are in operation. Show how to calculate the probability distribution function of the time until the system goes down for the first time when all of the $r + s$ units are in good condition at time 0.

4.17 An electronic system uses one operating unit but has built-in redundancy in the form of R standby units. The standby units are not switched on (cold standby). The operating unit has an exponentially distributed lifetime with mean $1/\lambda$. If the operating unit fails, it is immediately replaced by a standby unit if available. Each failed unit enters repair immediately and is again available after an exponentially distributed repair time with mean $1/\mu$. It is assumed that the mean repair time is much smaller than the mean lifetime. There are ample repair facilities. The system is down when all $R + 1$ units are in repair. Assuming that all $R + 1$ units are in perfect condition at time 0, let the random variable τ be the time until the first system failure.

(a) Use the uniformization method to compute $E(\tau)$, $\sigma(\tau)$ and $P\{\tau > t\}$ for $t = 2, 5$ and 10 when $\lambda = 1$, $\mu = 10$ and the number of standby units is varied as $R = 1, 2$ and 3.

(b) Extend the analysis in (a) for the case that the repair time has a Coxian-2 distribution and investigate how sensitive the results in (a) are to the second moment of the repair-time distribution.

4.18 Messages arrive at a node in a communication network according to a Poisson process with rate λ. Each arriving message is temporarily stored in an infinite-capacity buffer until it can be transmitted. The messages have to be routed over one of two communication lines each with a different transmission time. The transmission time over the communication line is i exponentially distributed with mean $1/\mu_i (i = 1, 2)$, where $1/\mu_1 < 1/\mu_2$ and $\mu_1 + \mu_2 > \lambda$. The faster communication line is always available for service, but the slower line will be used only when the number of messages in the buffer exceeds some critical level. Each line is only able to handle one message at a time and provides non-pre-emptive service. With the goal of minimizing the average sojourn time (including transmission time) of a message in the system, the following control rule with switching level L is used. The slower line is turned on for transmitting a message when the number of messages in the system exceeds the level L and is turned off again when it completes a transmission and the number of messages left behind is at or below L. Show how to calculate the average sojourn time of a message in the system. This problem is taken from Lin and Kumar (1984).

4.19 Two communication lines in a packet switching network share a finite storage space for incoming messages. Messages of the types 1 and 2 arrive at the storage area according to two independent Poisson processes with respective rates λ_1 and λ_2. A message of type j is destined for communication line j and its transmission time is exponentially distributed with mean $1/\mu_j$, $j = 1, 2$. A communication line is only able to transmit one message at a time. The storage space consists of M buffer places. Each message requires exactly one buffer place and occupies the buffer place until its transmission time has been completed. A number N_j of buffer places are reserved for messages of type j and a number N_0 of buffer places are to be used by messages of both types, where $N_0 + N_1 + N_2 = M$. That is, an arriving message of type j is accepted only when the buffer is not full and less than $N_0 + N_1$ other messages of the same type j are present; otherwise, the message is rejected. Discuss how to calculate the optimal values of N_0, N_1 and N_2 when the goal is to minimize the total rejection rate of both types of message. Write a computer program and solve for the numerical data $M = 15$, $\lambda_1 = \lambda_2 = 1$ and $\mu_1 = \mu_2 = 1$. This problem is based on Kamoun and Kleinrock (1980).

4.20 A traffic source is alternately on and off, where the on- and off-times are exponentially distributed with respective means $1/\delta$ and $1/\beta$. During on-periods the traffic source generates messages for a transmission channel according to a Poisson process with rate λ. The transmission channel can handle only one message at a time and the transmission time of a message has an exponential distribution with mean $1/\mu$. The on-times, off-times and transmission times are independent of each other. Further, it is assumed that $\lambda\beta/[\mu(\delta + \beta)] < 1$. Let the states $(i, 0)$ and $(i, 1)$ correspond to the situation that there are i messages at the transmission channel and the traffic source is off or on respectively.

(a) Verify for the numerical values $\lambda = 1$, $\mu = 1$, $\beta = 2$, $\delta = 0.5$ that the system of linear equations (4.4.6) is given by

$$\begin{pmatrix} 1 - 3z & 0.5z \\ 2z & z^2 - 2.5z + 1 \end{pmatrix} \begin{pmatrix} G_0(z) \\ G_1(z) \end{pmatrix} = \begin{pmatrix} (1 - z)p_{00} \\ (1 - z)p_{01} \end{pmatrix}.$$

Verify the roots of $\det \mathbf{A}(z) = 0$ are $z_0 = 1$, $z_1 = 0.2712865$ and $z_2 = 1.2287136$.

(b) Use the roots z_0 and z_1 and the fact that $G_i(z)$ is analytic for $|z| \leq 1$ to find p_{00} and p_{01}.

(c) Use partial-fraction expansion to show that $p(i,s) = \gamma_s (z_2)^{-i}$ for $i = 1, 2, \ldots$ and $s = 0, 1$. Specify the values of γ_0 and γ_1.

4.21 Consider a multi-server queueing system with c unreliable servers. Jobs arrive according to a Poisson process with rate λ. The required service times of the jobs are independent random variables having a common exponential distribution with mean $1/\mu$. The service of a job may be interrupted by a server breakdown. The server operates uninterruptedly during an exponentially distributed time with mean $1/\delta$. It takes an exponentially distributed time with mean $1/\beta$ to bring a broken-down server to the operative state. Any interrupted service is resumed at the point it was interrupted. It is assumed that an interrupted service is taken over by the first available server.

Denote by $p(i,s)$ the limiting probability of having i jobs present and s operative servers for $i \geq 0$ and $0 \leq s \leq c$. Prove that the probabilities $p(i,s)$ can be computed by using the geometric tail approach. In particular, verify that

$$p(i,s) \sim \gamma_s \eta^i \quad \text{as } i \to \infty$$

for a constant γ_s, where η is the reciprocal of the smallest root of $\det[\mathbf{M}(z)] = 0$ on the interval $(1, \infty)$. Here $\mathbf{M}(z) = (m_{st}(z)), s, t = 0, 1, \ldots, c$ is a tridiagonal $(c+1) \times (c+1)$ matrix with $m_{ss}(z) = \lambda z - [\lambda + s(\mu + \delta) + (c-s)\beta] + s\mu/z$, $m_{s,s-1}(z) = (c-s+1)\beta$ and $m_{s,s+1}(z) = (s+1)\delta$. This problem is based on Mitrani and Avi-Itzhak (1968).

4.22 Consider the unloader problem from Example 4.1.2 again. Assume now that the unloading time of a ship has an Erlang (L, μ) distribution and the repair time of the unloader has an Erlang (R, β) distribution. Letting $\rho = (\lambda L/\mu)(1 + \delta R/\beta)$, it is assumed that the server utilization ρ is less than 1. Interpret the unloading time of a ship as a sequence of L independent unloading phases each having an exponential distribution with mean $1/\mu$. Also, interpret the repair time of the unloader as a sequence of R independent repair phases each having an exponential distribution with mean $1/\beta$. Let state $(i, 0)$ correspond to the situation the unloader is available and i uncompleted unloading phases are present $(i \geq 0)$. Let state (i, r) correspond to the situation that there are i uncompleted unloading phases $(i \geq 1)$ and the unloader is in repair with r remaining repair phases $(1 \leq r \leq R)$. Denote by $p(i,s)$ the equilibrium probability of state (i,s) and define the generating functions $G_s(z)$ by $G_0(z) = \sum_{i=0}^{\infty} p(i,0) z^i$ and $G_r(z) = \sum_{i=1}^{\infty} p(i,r) z^i$ for $|z| \leq 1$.

(a) Verify that

$$G_s(z) = \frac{\det \mathbf{A}_s(z)}{\det \mathbf{A}(z)}, \quad s = 0, 1, \ldots, R.$$

Here $\mathbf{A}(z)$ is the $(R+1) \times (R+1)$ matrix $\mathbf{A}(z) = (1-z)\mathbf{M} - \lambda z(1 - z^L)\mathbf{I} + z\mathbf{Q}^T$, where $\mathbf{M} = \text{diag}(\mu, 0, \ldots, 0)$ and $\mathbf{Q}^T$ is the transpose of the transition matrix $\mathbf{Q} = (q_{ij})$ with $q_{0R} = -q_{00} = \delta$, $q_{i,i-1} = -q_{ii} = \beta$ for $1 \leq i \leq R$ and the other $q_{ij} = 0$. The matrix $\mathbf{A}_s(z)$ results from replacing the $(s+1)$th column vector of $\mathbf{A}(z)$ by the vector $\mathbf{b}(z)$ with $\mathbf{b}^T(z) = ((\mu(1-z) - \delta z)p(0,0), 0, \ldots, 0)$.

(b) Conclude that for any $s = 0, 1, \ldots, R$,

$$p(i,s) \sim \gamma_s \eta^i \quad \text{as } i \to \infty$$

for a constant γ_s, where η is the reciprocal of the smallest root of $\det \mathbf{A}(x) = 0$ on the interval $(1, \infty)$. Note that for Erlangian service the polynomial equation

$$\det \mathbf{A}(z) = (-1)^{R+1}[\{\lambda z(1 - z^L) - \mu(1-z) + \delta z\}\{\lambda z(1 - z^L) + \beta z\}^R - \delta z(\beta z)^R] = 0$$

is obtained by expanding $\det \mathbf{A}(z)$ in the cofactors of its first row.

4.23 Repeat the analysis in Exercise 4.22 when the repair time is H_2 distributed with parameters (p_1, ν_1, p_2, ν_2) rather than Erlang (R, λ) distributed. Verify that the results remain the same when we take $R = 2$ and replace the matrix $\mathbf{Q}$ by

$$\mathbf{Q} = \begin{pmatrix} -\delta & \delta p_1 & \delta p_2 \\ \nu_1 & -\nu_1 & 0 \\ \nu_2 & 0 & -\nu_2 \end{pmatrix}$$

4.24 At a facility for train maintenance, work is done on a number of separate parallel tracks. On each of these tracks there is room for two trains on a front part and a back part. Trains can leave the tracks only on the same side they enter the tracks. That is, upon completion of its maintenance a train may be locked in by another train that arrived later on the same track but has not yet completed its maintenance. For each of the tracks there are two maintenance crews, one for the train at the front part of the track and one for the train at the back. Trains requesting maintenance arrive at the maintenance facility according to a Poisson process with rate λ. A train immediately receives maintenance when it finds a free place at one of the tracks upon arrival; otherwise, the train waits until a maintenance place becomes free. A newly arriving train is directed to a front part if both a front part and a back part are free. The amount of time needed to serve a train has an exponential distribution with mean $1/\mu$. It is assumed that $\lambda < \frac{3}{2}c\mu$.

(a) Formulate a continuous-time Markov time chain for the performance evaluation of the maintenance track.

(b) Argue that the geometric tail approach can be used to reduce the infinite system of equilibrium equations to a finite system of linear equations. This problem is based on Adan *et al*. (1999).

BIBLIOGRAPHIC NOTES

The theory of continuous-time Markov chains is more delicate than the theory of discrete-time Markov chains. Basic references are Anderson (1991) and Chung (1967). The continuous-time Markov chain model is the most versatile model in applied probability. The powerful technique of equating the flow out of a state to the flow into that state has a long history and goes back to the pioneering work of Erlang on stochastic processes in the early 1900s; see also Kosten (1973). The uniformization technique for the transient analysis of continuous-time Markov chains goes back to Jensen (1953) and is quite useful for both analytical and computational purposes. The extension of the uniformization method to compute the transient probability distribution of the sojourn time in a given set of states is due to De Soua e Silva and Gail (1986). The material in Section 4.6.2 for the computation of the transient reward distribution is based on Goyal and Tantawi (1988) and Tijms and Veldman (2000); see also Sericola (2000) for an alternative method. The Hubble telescope problem from Example 4.5.3 is taken from Hermanns (2001).

REFERENCES

Adan, I.J.B.F. and Resing, J.A.C. (1999) A class of Markov processes on a semi-infinite strip. In *Proc. 3rd International Meeting on the Numerical Solution of Markov Chains*, edited by B. Plateau, W.J. Stewart and M. Silva, pp. 41–57. Zaragoza University Press.

Adan, I.J.B.F., Wessels, J. and Zijm, W.H.M. (1993) A compensation approach for two-dimensional Markov processes. *Adv. Appl. Prob.*, **25**, 783–817.

Adan, I.J.B.F., De Kok, A.G. and Resing, J.A.C. (1999) A multi-server queueing model with locking. *Euro. J. Operat. Res.*, **116**, 249–258.

Anderson, W.J. (1991) *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer-Verlag, Berlin.

Blanc, J.P.C. (1992) The power-series algorithm application to the shortest queue model. *Operat. Res.*, **40**, 157–167.

Chung, K.L. (1967) *Markov Chains with Stationary Transition Probabilities*, 2nd edn. Springer-Verlag, Berlin.

Daigle, J.N. (1991) *Queueing Theory for Telecommunications*. Addison-Wesley, Reading MA.

De Soua e Silva, E. and Gail, H.R. (1986) Calculating cumulative operational time distributions of repairable computer systems. *IEEE Trans. Comput.*, **35**, 322–332.

Goyal, A. and Tantawi, A.N. (1988) A measure of guaranteed availability and its numerical evaluation. *IEEE Trans. Comput.*, **37**, 25–32.

Hermanns, H. (2001) Construction and verification of performance and reliability models. *Bull. EACTS*, **74**, 135–154.

Hooghiemstra, G., Keane, M. and Van de Ree, S. (1988) Power series for stationary distribution of coupled processor models. *SIAM J. Math. Appl.*, **48**, 1159–1166.

Jensen, A. (1953) Markov chains as an aid in the study of Markoff process. *Skand. Aktuarietidskr.*, **36**, 87–91.

Kamoun, F. and Kleinrock, L. (1980) Analysis of a shared finite storage in a computer network node environment under general traffic conditions. *IEEE Trans. Commun.*, **28**, 992–1003.

Kosten, L. (1973) *Stochastic Theory of Service Systems*. Pergamon Press, London.

Latouche, G. and Ramaswami, V. (1993) A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Prob.*, **30**, 650–674.

Lin, W. and Kumar, P. (1984) Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automat. Contr.*, **29**, 696–703.

Mitrani, I. and Avi-Itzhak, B. (1968) A many-server queue with service interruptions. *Operat. Res.*, **16**, 628–638.

Mitrani, I. and Mitra, D. (1992) A spectral expansion method for random walks on semi-infinite strips. In *Iterative Methods in Linear Algebra*, edited by R. Beauwens and P. Groen. North-Holland, Amsterdam.

Morse, P.M. (1955) Stochastic properties of waiting lines. *Operat. Res.*, **3**, 255–261.

Neuts, M. (1981) *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore MD.

Odoni, A.R. and Roth, E. (1983) An empirical investigation of the transient behaviour of stationary queueing systems. *Operat. Res.*, **31**, 432–455.

Sericola, B. (2000) Occupation times in Markov processes. *Stochastic Models*, **16**, 339–351.

Takács, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, New York.

Tijms, H.C. and Veldman, R. (2000) A fast algorithm for the transient reward distribution in continuous-time Markov chains. *Operat. Res. Lett.*, **26**, 155–158.

CHAPTER 5

Markov Chains and Queues

5.0 INTRODUCTION

Markov chain theory has numerous applications to queueing systems. This chapter gives a first introduction to the analysis of queues and stochastic networks. In Section 5.1 we consider the Erlang delay model with Poisson arrivals and exponential services. We first analyse the single-server $M/M/1$ queue and next the multi-server $M/M/c$ queue. Section 5.2 deals with both the Erlang loss model with Poisson input and the Engset loss model with finite-source input. The Erlang delay model and Erlang's loss formula will be used in Section 5.3 to obtain a square-root staffing rule for the design of stochastic service systems. The Erlang loss model and the Engset loss model have the so-called insensitivity property stating that the equilibrium distribution of the number of customers present is insensitive to the form of the service-time distribution and requires only the mean service time. This insensitivity property, being of utmost importance in practice, will be discussed in a more general framework in Section 5.4. The so-called phase method is the subject of Section 5.5. This powerful method uses the idea that any probability distribution function of a non-negative random variable can be arbitrarily closely approximated by a mixture of Erlangian distributions with the same scale parameters. This fundamental result greatly enhances the applicability of the continuous-time Markov chain model. In Section 5.6 the theory of continuous-time Markov chains will be used to analyse open and closed queueing networks. In particular, a product-form formula will be established for the joint distribution of the number of customers present at the various nodes of the network.

5.1 THE ERLANG DELAY MODEL

Consider a multi-server station at which customers arrive according to a Poisson process with rate λ. There are c servers with a shared infinite-capacity waiting line. If an arriving customer finds a free server, the customer immediately enters service; otherwise, the customer joins the queue. The service times of the customers are

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

independent random variables having a common exponential distribution with mean $1/\mu$. The service times and the arrival process are independent of each other. Let

$$\rho = \frac{\lambda}{c\mu}. \tag{5.1.1}$$

It is assumed that $\rho < 1$. Rewriting this condition as $\lambda/\mu < c$, the condition states that the average amount of work offered to the servers per time unit is less than the total service capacity. The factor ρ is called the *server utilization*. This model is a basic model in queueing theory and is often called the *Erlang delay model*. It is usually abbreviated as the $M/M/c$ queue. Using continuous-time Markov chain theory we will derive the distributions of the queue size and the delay in queue of a customer. Let

$$X(t) = \text{the number of customers present at time } t$$

(including any customer in service). Then the stochastic process $\{X(t)\}$ is a continuous-time Markov chain with infinite state space $I = \{0, 1, \dots\}$. The assumption $\rho < 1$ implies that the Markov chain satisfies Assumption 4.2.1 with regeneration state 0 and thus has a unique equilibrium distribution $\{p_j\}$ (a formal proof is omitted). The probability p_j gives the long-run fraction of time that j customers are present.

5.1.1 The $M/M/1$ Queue

For ease of presentation, we first analyse the single-server case with $c = 1$. The transition rate diagram of the process $\{X(t)\}$ is given in Figure 5.1.1

Note that for each state i the transition rate $q_{ij} = 0$ for $j \leq i-2$. This implies that the equilibrium probabilities p_j can be recursively computed; see formula (4.2.10). By equating the rate at which the process leaves the set $\{i, i+1, \dots\}$ to the rate at which the process enters this set, it follows that

$$\mu p_i = \lambda p_{i-1}, \quad i = 1, 2, \dots .$$

The recurrence equation allows for an explicit solution. Iterating the equation yields $p_i = (\lambda/\mu)^i p_0$ for all $i \geq 1$. Noting that this relation also holds for $i = 0$ and substituting it into the normalizing equation $\sum_{i=0}^{\infty} p_i = 1$, we find $p_0(1-\lambda/\mu)^{-1} =$

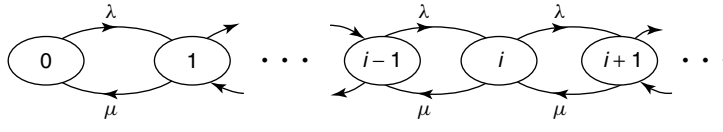


Figure 5.1.1 The transition rate diagram for the $M/M/1$ queue

1 and so $p_0 = 1 - \lambda/\mu$. Hence we find the explicit solution

$$p_i = (1 - \rho)\rho^i, \quad i = 0, 1, \ldots \tag{5.1.2}$$

with $\rho = \lambda/\mu$. In particular, $1 - p_0 = \rho$ and so ρ can be interpreted as the long-run fraction of time the server is busy. This explains why ρ is called the server utilization. Let

$$L_q = \text{the long-run average number of customers in queue}$$

(excluding any customer in service). The constant L_q is given by

$$L_q = \sum_{j=1}^{\infty} (j - 1) p_j,$$

as can be rigorously proved by assuming a cost at rate k whenever k customers are waiting in queue and applying Theorem 4.2.2. Substituting (5.1.2) into the formula for L_q, we obtain

$$L_q = \frac{\rho^2}{1 - \rho},$$

in agreement with the Pollaczek–Khintchine formula for the general $M/G/1$ queue.

To determine the waiting-time probabilities we need the so-called *customer-average* probabilities

$$\pi_j = \text{the long-run fraction of customers who find } j \text{ other customers present upon arrival, } j = 0, 1, \ldots .$$

In the $M/M/1$ case the customer-average probabilities π_j are identical to the time-average probabilities p_j, that is,

$$\pi_j = p_j, \quad j = 0, 1, \ldots . \tag{5.1.3}$$

This identity can be seen from the PASTA property. Alternatively, the identity can be proved by noting that in a continuous-time Markov chain, $p_j q_{jk}$ represents the long-run average number of transitions from state j to state k ($\neq j$) per time unit. Thus in the $M/M/1$ case the long-run average number of transitions from state j to state $j + 1$ per time unit equals λp_j. In other words, the long-run average number of arrivals per time unit finding j other customers present equals λp_j. Dividing λp_j by the average arrival rate λ yields the customer-average probability π_j. The probability distribution $\{\pi_j\}$ is the equilibrium distribution of the embedded Markov chain describing the number of customers present just before the arrival epochs of customers. This probability distribution enables us to find the steady-state waiting-time probabilities under the assumption of service in order of arrival. Let

$$W_q(x) = \lim_{n \to \infty} P\{D_n \leq x\}, \quad x \geq 0, \tag{5.1.4}$$

with D_n denoting the delay in queue of the nth customer. The existence of the limit will be shown below. It holds that

$$W_q(x) = 1 - \rho e^{-\mu(1-\rho)x}, \quad x \geq 0. \tag{5.1.5}$$

A key step in the proof is the observation that the conditional delay in queue of a customer finding j other customers present upon arrival has an Erlang (j, μ) distribution for $j \geq 1$. This follows by noting that the delay in queue of this customer is the sum of j independent exponential random variables with the same mean $1/\mu$ (the remaining service time of the customer found in service also has an exponential distribution with mean $1/\mu$). The probability distribution function of the Erlang (j, μ) distribution is given by $1 - \sum_{k=0}^{j-1} e^{-\mu x}(\mu x)^k/k!$. Denoting by $\pi_j^{(n)}$ the probability that the nth arriving customer finds j other customers present upon arrival, it follows that

$$P\{D_n > x\} = \sum_{j=1}^{\infty} \pi_j^{(n)} \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!}, \quad x \geq 0. \tag{5.1.6}$$

The embedded Markov chain describing the number of customers present just before the arrival epoch is irreducible and has the property that all states are *aperiodic* and positive recurrent. Thus $\lim_{n\to\infty} \pi_j^{(n)}$ exists and equals π_j for all j; see also relation (3.5.11). Using the bounded convergence theorem from Appendix A, it now follows that $\lim_{n\to\infty} P\{D_n > x\}$ exists and is given by

$$\lim_{n\to\infty} P\{D_n > x\} = \sum_{j=1}^{\infty} \pi_j \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!}, \quad x \geq 0. \tag{5.1.7}$$

To obtain (5.1.5) from (5.1.7), we use (5.1.2) and (5.1.3). This gives

$$\begin{aligned} 1 - W_q(x) &= \sum_{j=1}^{\infty} \pi_j \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!} = \sum_{k=0}^{\infty} e^{-\mu x} \frac{(\mu x)^k}{k!} \sum_{j=k+1}^{\infty} \pi_j \\ &= \sum_{k=0}^{\infty} e^{-\mu x} \frac{(\mu x)^k}{k!} \rho^{k+1} = \rho e^{-\mu x} \sum_{k=0}^{\infty} \frac{(\mu \rho x)^k}{k!} = \rho e^{-\mu x} e^{\mu \rho x}, \end{aligned}$$

which verifies (5.1.5). It is noted that the probability $W_q(x)$ can also be interpreted as the long-run fraction of customers whose delay in queue is no more than x.

5.1.2 The $M/M/c$ Queue

The analysis of the multi-server $M/M/c$ queue is a rather straightforward extension of the analysis of the $M/M/1$ queue. The transition rate diagram for the $\{X(t)\}$ process is given in Figure 5.1.2.

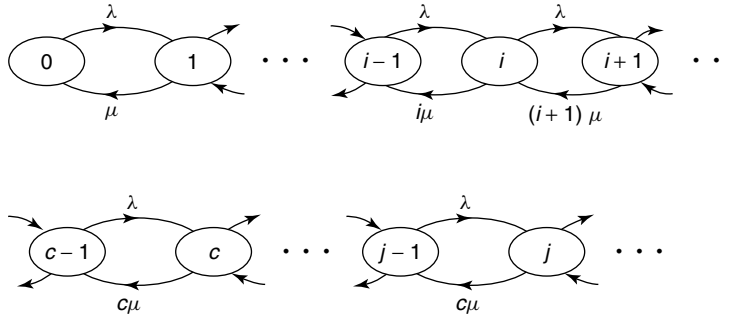


Figure 5.1.2 The transition rate diagram for the $M/M/c$ queue

Using the technique of equating the rate at which the process leaves the set of states $\{j, j+1, \dots\}$ to the rate at which the process enters this set, we obtain

$$\min(j, c)\mu p_j = \lambda p_{j-1}, \quad j = 1, 2, \dots . \tag{5.1.8}$$

An explicit solution for the p_j is easily given, but this explicit solution is of little use for computational purposes. A simple computational scheme can be based on the recursion relation (5.1.8). To do so, note that $p_j = \rho p_{j-1}$ for $j \geq c$. This implies $p_j = \rho^{j-c+1} p_{c-1}$ for $j \geq c$ and so

$$\sum_{j=c}^{\infty} p_j = \frac{\rho p_{c-1}}{1-\rho}. \tag{5.1.9}$$

A simple algorithm now follows.

Algorithm

Step 0. Initialize $\overline{p}_0 := 1$.
Step 1. For $j = 1, \dots, c-1$, let $\overline{p}_j := \lambda \overline{p}_{j-1}/(j\mu)$.
Step 2. Calculate the normalizing constant γ from

$$\gamma = \left[\sum_{j=0}^{c-1} \overline{p}_j + \frac{\rho \overline{p}_{c-1}}{1-\rho} \right]^{-1}.$$

Normalize the $\overline{p}_j$ according to $p_j := \gamma \overline{p}_j$ for $j = 0, 1, \dots, c-1$.
Step 3. For any $j \geq c$, $p_j := \rho^{j-c+1} p_{c-1}$.

As before, define the customer-average probability π_j as the long-run fraction of customers who see j other customers present upon arrival. By the same arguments as used for the $M/M/1$ queue, we have $\pi_j = p_j$ for $j = 0, 1, \dots$. Denote by $P_{delay} = \sum_{j=c}^{\infty} \pi_j$ the long-run fraction of customers who are delayed. By $\pi_j = p_j$

for all j and (5.1.9),

$$P_{delay} = \frac{\rho}{1-\rho} p_{c-1}. \tag{5.1.10}$$

It is also possible to give an explicit expression for the delay probability:

$$P_{delay} = \frac{(c\rho)^c/c!}{[(c\rho)^c/c! + (1-\rho)\sum_{k=0}^{c-1}(c\rho)^k/k!]}. \tag{5.1.11}$$

The delay probability for the $M/M/c$ queue is often called *Erlang's delay probability*. Given the representation $L_q = \sum_{j=c}^{\infty}(j-c)p_j$ for the long-run average queue size, it follows from $p_j = \rho^{j-c+1}p_{c-1}$ for $j \geq c$ that

$$L_q = \frac{\rho^2}{(1-\rho)^2} p_{c-1}. \tag{5.1.12}$$

Under the assumption that customers are served in order of arrival, define the steady-state waiting-time probability $W_q(x)$ in the same way as for the $M/M/1$ queue. The formula (5.1.5) generalizes to

$$W_q(x) = 1 - \frac{\rho}{1-\rho} p_{c-1} e^{-c\mu(1-\rho)x}, \quad x \geq 0. \tag{5.1.13}$$

This result is obtained by a slight modification of the derivation of (5.1.5). Since the service times are exponentially distributed and the minimum of c (remaining) service times has an exponential distribution with mean $1/(c\mu)$, service completions occur according to a Poisson process with rate $c\mu$ as long as c or more customers are present. Thus the conditional delay in queue of a customer finding $j \geq c$ other customers present upon arrival has an Erlang $(j-c+1, c\mu)$ distribution. This gives

$$1 - W_q(x) = \sum_{j=c}^{\infty} \pi_j \sum_{k=0}^{j-c} e^{-c\mu x} \frac{(c\mu x)^k}{k!}, \quad x \geq 0,$$

which leads to (5.1.13) after some algebra. In particular, the average delay in queue of a customer equals

$$W_q = \frac{\rho}{c\mu(1-\rho)^2} p_{c-1} \tag{5.1.14}$$

in agreement with (5.1.12) and Little's formula $L_q = \lambda W_q$. Also, by Little's formula, the long-run average number of busy servers equals $c\rho$; see Section 2.3 Thus the long-run fraction of time that a given server is busy equals ρ.

5.1.3 The Output Process and Time Reversibility

Define for the $M/M/c$ queue

T_n = the epoch at which the nth service completion occurs.

Then the following important result holds for the output process.

Burke's output theorem *For any* $k \geq 1$,

$$\lim_{n\to\infty} P\{T_{n+1} - T_n \leq x_1, \ldots, T_{n+k} - T_{n+k-1} \leq x_k\}$$
$$= (1 - e^{-\lambda x_1}) \cdots (1 - e^{-\lambda x_k}) \quad \textit{for all } x_1, \ldots, x_k \geq 0.$$

In other words, in statistical equilibrium the process describing the departures of served customers is a Poisson process with rate λ.

We first give a heuristic argument for this result. If at a given time t there are i customers present, then the probability that in $(t, t + \Delta t)$ a service is completed equals $\min(i, c)\mu\Delta t + o(\Delta t)$ for $\Delta t \to 0$. The equilibrium probability of being in state i at an arbitrary point in time is given by p_i. Assuming that the process is in statistical equilibrium, it follows that the probability of a customer leaving in $(t, t + \Delta t)$ is given by

$$\sum_{i=0}^{c-1} i\mu\Delta t p_i + \sum_{i=c}^{\infty} c\mu\Delta t p_i + o(\Delta t) = \left[\sum_{i=0}^{c-1} i p_i + c\sum_{i=c}^{\infty} p_i\right]\mu\Delta t + o(\Delta t)$$

as $\Delta t \to 0$. The expression between brackets gives the long-run average number of busy servers and is thus equal to $c\rho$ by Little's formula. Since $\rho = \lambda/(c\mu)$ it follows that the probability of a customer leaving in $(t, t + \Delta t)$ equals

$$c\rho\mu\Delta t + o(\Delta t) = \lambda\Delta t + o(\Delta t)$$

as $\Delta t \to 0$. This indicates that the departure process of customers is indeed a Poisson process with rate λ when the $M/M/c$ system has reached statistical equilibrium. This result is of utmost importance for tandem queues when the first station in the tandem queue is described by an $M/M/c$ system.

Time reversibility

The practically useful result that the output process of an $M/M/c$ queue is a Poisson process can be given a firm basis by the important concept of time reversibility. Consider a continuous-time Markov chain $\{X(t)\}$ that satisfies Assumption 4.2.1 and has the property that all states communicate with each other. The continuous-time Markov chain $\{X(t)\}$ is said to satisfy *detailed balance* if its unique equilibrium distribution $\{p_j\}$ has the property that

$$p_k q_{kj} = p_j q_{jk} \quad \text{for all } j, k \in I \text{ with } j \neq k. \tag{5.1.15}$$

In other words, the long-run average number of transitions from state k to state j per time unit is equal to the long-run average number of transitions from state j to state k per time unit for all $j \neq k$. Detailed balance is intimately related to time reversibility. A convenient way to characterize time reversibility is to consider the stationary version of the Markov chain $\{X(t)\}$. In the stationary version the initial state at time $t = 0$ is chosen according to the equilibrium distribution $\{p_j\}$. For the

stationary process $\{X(t)\}$ it holds that $P\{X(t) = j\} = p_j$, $j \in I$, for all $t \geq 0$. It can be shown that the condition (5.1.15) is satisfied if and only if the stationary version of the Markov process $\{X(t)\}$ has the property that for all $n \geq 1$ and all $u > 0$,

$$(X(u_1), \ldots, X(u_n)) \text{ is distributed as } (X(u - u_1), \ldots, X(u - u_n)) \qquad (5.1.16)$$

for all $0 \leq u_1 < \cdots < u_n \leq u$. A Markov process with this property is said to be *time reversible*. In other words, the process reversed in time has the same probabilistic structure as the original process when the process has reached statistical equilibrium. It is as if you would see the same film shown in reverse. Let us return to the $M/M/c$ system. In the $M/M/c$ system the rate at which the process goes directly from state i to state $i+1$ is then equal to the rate at which the process goes directly from state $i+1$ to state i; see relation (5.1.8). Hence the $M/M/c$ system has the property (5.1.16). Going forward in time, the time points at which the number in the system increases by 1 are exactly the arrival epochs of customers and thus constitute a Poisson process. Going backwards in time, the time points at which the number in the system increases by 1 are exactly the time points at which customers depart. Hence, by time reversibility, the departure process of customers must be a Poisson process when the $M/M/c$ system has reached statistical equilibrium.

5.2 LOSS MODELS

In a delay system each customer finding no free server upon arrival waits until a server becomes available. Opposite to delay systems are loss systems in which customers finding no free server upon arrival are lost and have no further influence on the system. In this section we consider two basic loss models. The famous Erlang loss model with Poisson input is dealt with in Section 5.2.1. Section 5.2.2 considers the Engset loss model with finite-source input.

5.2.1 The Erlang Loss Model

Consider a communication system with c transmission channels at which messages are offered according to a Poisson process with rate λ. The system has no buffer to temporarily store messages that arrive when all channels are occupied. An arriving message that finds all c channels busy is lost and has no further influence on the system; otherwise, the message is assigned to a free channel and its transmission immediately starts. The transmission times of the messages are independent and identically distributed random variables. Also, the arrival process and the transmission times are independent of each other. The goal is to find an expression for the long-run fraction of messages that are lost. This model is called *Erlang's loss model* after the Danish telephone engineer A.K. Erlang. It is often abbreviated as the $M/G/c/c$ queue. In the early 1900s Erlang studied this model in the framework of a telephone switch which can handle only c calls. Though the theory of stochastic processes was not yet developed in Erlang's time, Erlang (1917) was able

to find a formula for the fraction of calls that are lost. He established this formula first for the particular case of exponentially distributed holding times. Also, Erlang conjectured that the formula for the loss probability remains valid for generally distributed holding times. His conjecture was that the loss probability is insensitive to the form of the holding time distribution but depends only on the first moment of the holding time. A proof of this insensitivity result was only given many years after Erlang made his conjecture; see for example Cohen (1976) and Takács (1969). The proof of Takács (1969) is rather technical and involves Kolmogoroff's forward equations for Markov processes with a general state space. The more insightful proof in Cohen (1976) is based on the concept of reversible Markov processes.

In Section 5.4 we will discuss the issue of insensitivity for loss systems in a more general context. It is the insensitivity property that makes the Erlang loss model such a useful model. Still nowadays the model is often used in the analysis of telecommunication systems. The Erlang loss model also has applications in a variety of other fields, including inventory and reliability; see Exercises 5.9 to 5.14. A nice application is the $(S-1, S)$ inventory system in which the demand process is a Poisson process and demands occurring when the system is out of stock are lost (the back ordering case was analysed in Section 1.1.3 through the $M/G/\infty$ queueing model).

In view of the above discussion, we now assume that the transmission times have an exponential distribution with mean $1/\mu$. For any $t \geq 0$, let

$$X(t) = \text{the number of busy channels at time } t.$$

The stochastic process $\{X(t), t \geq 0\}$ is a continuous-time Markov chain with state space $I = \{0, 1, \ldots, c\}$. Its transition rate diagram is given in Figure 5.2.1. The time-average probability p_i gives the long-run fraction of time that i channels are occupied. Since for each state i the transition rate $q_{ij} = 0$ for $j \leq i-2$, the equilibrium probabilities p_i can be recursively computed. Equating the rate out of the set of states $\{i, i+1, \ldots, c\}$ to the rate into this set, we obtain

$$i\mu p_i = \lambda p_{i-1}, \quad i = 1, \ldots, c.$$

This equation can be solved explicitly. Iterating the equation gives $p_i = (\lambda/\mu)^i p_0/i!$ for $i = 1, \ldots, c$. Using the normalizing equation $\sum_{i=0}^{c} p_i = 1$, we obtain

$$p_i = \frac{(\lambda/\mu)^i/i!}{\sum_{k=0}^{c}(\lambda/\mu)^k/k!}, \quad i = 0, 1, \ldots, c. \tag{5.2.1}$$

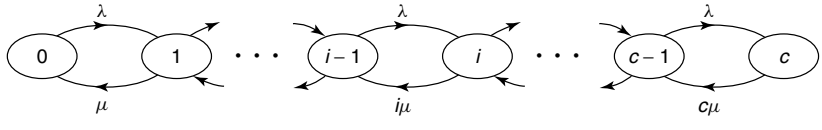


Figure 5.2.1 The transition rate diagram for the Erlang loss model

Note that the distribution in (5.2.1) is a *truncated Poisson distribution* (multiply both the numerator and the denominator by $e^{-\lambda/\mu}$). Denote by the customer-average probability π_i the long-run fraction of messages that find i other messages present upon arrival. Then, by the PASTA property,

$$\pi_i = p_i, \quad i = 0, 1, \ldots, c.$$

In particular, denoting by P_{loss} the long-run fraction of messages that are lost,

$$P_{loss} = \frac{(\lambda/\mu)^c/c!}{\sum_{k=0}^{c}(\lambda/\mu)^k/k!}. \tag{5.2.2}$$

This formula is called the *Erlang loss formula.* As said before, the formula (5.2.1) for the time-average probabilities p_j and the formula (5.2.2) for the loss probability remain valid when the transmission time has a general distribution with mean $1/\mu$. The state probabilities p_j are *insensitive* to the form of the probability distribution of the transmission time and require only the mean transmission time. Letting $c \to \infty$ in (5.2.1), we get the Poisson distribution with mean λ/μ in accordance with earlier results for the $M/G/\infty$ queue. The insensitivity property of this infinite-server queue was proved in Section 1.1.3.

5.2.2 The Engset Model

The Erlang loss model assumes Poisson arrivals and thus has an infinite source of potential customers. The Engset model differs from the Erlang loss model only by assuming a finite source of customers. There are M sources which generate service requests for c service channels. It is assumed that $M > c$. A service request that is generated when all c channels are occupied is lost. Each source is alternately on and off. A source is off when it has a service request being served, otherwise the source is on. A source in the on-state generates a new service request after an exponentially distributed time (the think time) with mean $1/\alpha$. The sources act independently of each other. The service time of a service request has an exponential distribution with mean $1/\mu$ and is independent of the think time. This model is called the *Engset model* after Engset (1918).

We now let

$$X(t) = \text{the number of occupied channels at time } t.$$

The process $\{X(t), t \geq 0\}$ is a continuous-time Markov chain with state space $I = \{0, 1, \ldots, c\}$. Its transition rate diagram is given in Figure 5.2.2. By equating

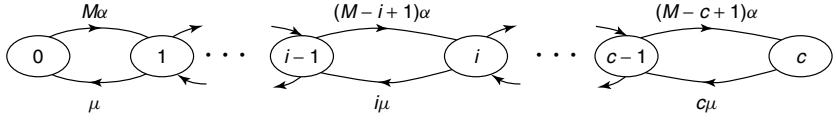


Figure 5.2.2 The transition rate diagram for the Engset loss model

the rate at which the process leaves the set of states $\{i, i+1, \ldots, c\}$ to the rate at which the process enters this set, we obtain the recursive equation

$$i\mu p_i = (M - i + 1)\alpha p_{i-1}, \quad i = 1, \ldots, c.$$

This recursive equation allows for the explicit solution (verify):

$$p_i = \frac{\binom{M}{i} p^i (1-p)^{M-i}}{\sum_{k=0}^{c} \binom{M}{k} p^k (1-p)^{M-k}}, \quad i = 0, 1, \ldots, c, \tag{5.2.3}$$

where p is given by

$$p = \frac{1/\mu}{1/\mu + 1/\alpha}.$$

The distribution (5.2.3) is a *truncated binomial distribution*. To compute the fraction of service requests that are lost, we need the customer-average probabilities

$$\pi_i = \text{the long-run fraction of service requests that find } i \text{ busy channels upon arrival}, \quad i = 0, 1, \ldots, c.$$

The π_i are found by noting that

$$\pi_i = (\text{the long-run average number of service requests that are generated per time unit and find } i \text{ busy channels upon arrival}) \big/ (\text{the long-run average number of service requests that are generated per time unit}).$$

In state i, service requests are generated at a rate $(M - i)\alpha$. Thus the arrival rate of service requests that see i busy channels equals $(M - i)\alpha p_i$. Hence

$$\pi_i = \frac{(M-i)\alpha p_i}{\sum_{k=0}^{c} (M-k)\alpha p_k}, \quad i = 0, 1, \ldots, c.$$

It next follows from (5.2.3) that

$$\pi_i = \frac{\binom{M-1}{i} p^i (1-p)^{M-1-i}}{\sum_{k=0}^{c} \binom{M-1}{k} p^k (1-p)^{M-1-k}}, \quad i = 0, 1, \ldots, c. \tag{5.2.4}$$

It is a remarkable finding that the distribution $\{\pi_i\}$ is the same as the distribution $\{p_i\}$ except that M is replaced by $M-1$. In other words, the equilibrium distribution of the state *just prior* to the arrival epochs of new service requests is the same as

the equilibrium distribution of the state at an *arbitrary* epoch in the system with *one source less*. In particular, we find

$$\text{the long-run fraction of lost service requests} = \frac{\binom{M-1}{c} p^c (1-p)^{M-1-c}}{\sum_{k=0}^{c} \binom{M-1}{k} p^k (1-p)^{M-1-k}}. \tag{5.2.5}$$

The formulas (5.2.3) to (5.2.5) have been derived under the assumption of exponentially distributed think times and exponentially distributed service times. This assumption is not needed. The Engset model has the insensitivity property that the formulas (5.2.3) to (5.2.5) remain valid when the think time has a general probability distribution with mean $1/\alpha$ and the service time has a general distribution with mean $1/\mu$. This insensitivity result requires the technical condition that either of these two distributions has a positive density on some interval. We come back to this insensitivity result in the next section. By letting $M \to \infty$ and $\alpha \to 0$ such that $M\alpha$ remains equal to the constant λ, it follows from the Poisson approximation to the binomial probability that the right-hand side of (5.2.3) converges to

$$\frac{e^{-\lambda/\mu}(\lambda/\mu)^i / i!}{\sum_{k=0}^{c} e^{-\lambda/\mu}(\lambda/\mu)^k / k!}, \quad i = 0, 1, \ldots, c$$

in agreement with (5.2.1). In other words, the Erlang loss model is a limiting case of the Engset model. This is not surprising, since the arrival process of service requests becomes a Poisson process with rate λ when we let $M \to \infty$ and $\alpha \to 0$ such that $M\alpha = \lambda$.

5.3 SERVICE-SYSTEM DESIGN

The Erlang delay model has many practical applications. In particular, it can be used to analyse capacity and staffing problems such as those arising in the area of telemarketing and call centre design and in the area of healthcare facilities planning. In this section it will be shown that a normal approximation to Erlang's delay formula is very helpful in analysing such problems. The normal approximation enables us to derive an insightful square-root staffing rule.

The mathematical analysis of the $M/M/c$ queue was given in Section 5.1.2. In the $M/M/c$ queue customers arrive according to a Poisson process with rate λ, the service times of the customers are exponentially distributed with mean $1/\mu$ and there are c identical servers. It is convenient to denote the offered load to the system by

$$R = \frac{\lambda}{\mu}.$$

Note that R is a dimensionless quantity that gives the average amount of work offered per time unit to the c servers. The offered load R is often expressed as R erlangs of work. In order to ensure the existence of a steady-state regime for the queue, it should be assumed that the service capacity c is larger than the offered load R. Hence the assumption is made that the server utilization

$$\rho = \frac{R}{c}$$

is less than 1. Note that ρ represents the long-run fraction of time a given server is busy. In the single-server case the server utilization ρ should not be too close to 1 in order to avoid excessive waiting of the customers. A rule of thumb for practical applications of the $M/M/1$ model is that the server utilization should not be much above 0.8. A natural question is how this rule of thumb should be adjusted for the multi-server case. It is instructive to have a look at Table 5.3.1. This table gives for several values of c and R the delay probability P_W, the average waiting T_W over the delayed customers and the 95% percentile $\eta_{0.95}$ of the steady-state waiting-time distribution of the delayed customers. In Table 5.3.1 we have normalized the mean service time $1/\mu$ as 1. The delay probability $P_W (= P_{delay})$ is given by formula (5.1.11). Since $T_W = W_q / P_W$, it follows from (5.1.10) and (5.1.14) that

$$T_W = \frac{1}{c\mu(1-\rho)}.$$

By (5.1.10) and (5.1.13), the steady-state probability that a delayed customer has to wait longer than x time units is given by $e^{-c\mu(1-\rho)x}$ for $x \geq 0$. Thus the pth percentile η_p of the steady-state waiting-time distribution of the delayed customers is found from $e^{-c\mu(1-\rho)x} = 1 - p$. This gives

$$\eta_p = \frac{-1}{c\mu(1-\rho)} \ln(1-p), \quad 0 < p < 1.$$

The following conclusion can be drawn from Table 5.3.1: *high values of the server utilization ρ do not conflict with acceptable service to the customers when*

Table 5.3.1 Service measures as function of c and R

	$\rho = R/c = 0.8$			$\rho = R/c = 0.95$			$\rho = R/c = 0.99$		
	P_W	T_W	$\eta_{0.95}$	P_W	T_W	$\eta_{0.95}$	P_W	T_W	$\eta_{0.95}$
$c = 1$	0.8	5	14.98	0.95	20	59.91	0.99	100	299.6
$c = 2$	0.711	2.5	7.49	0.926	10	29.96	0.985	50	149.8
$c = 5$	0.554	1	3.0	0.878	4	11.98	0.975	20	59.91
$c = 10$	0.409	0.5	1.5	0.826	2	5.99	0.964	10	29.96
$c = 25$	0.209	0.2	0.6	0.728	0.8	2.40	0.942	4	11.98
$c = 50$	0.087	0.1	0.3	0.629	0.4	1.20	0.917	2	5.99
$c = 100$	0.020	0.05	0.15	0.506	0.2	0.60	0.883	1	3.0
$c = 250$	3.9E-4	0.02	0.06	0.318	0.08	0.24	0.818	0.4	1.2
$c = 500$	8.4E-7	0.01	0.03	0.177	0.04	0.12	0.749	0.2	0.6

there are sufficiently many servers. The larger the number of servers, the higher the server utilization before the service to the customers seriously degrades. A relatively large value of P_W does not necessarily imply bad service to the customers. For example, take $c = 100$ and $\rho = 0.95$. Then on average 50.6% of the customers must wait, but the average wait of a delayed customer is only $\frac{1}{5}$ of its mean service time. Moreover, on average, only 5% of the delayed customers have to wait more than $\frac{3}{5}$ of the mean service time. The situation of many servers is encountered particularly in the telephone call centre industry. Service level is a key performance metric of a call centre. In practice it is often defined as '80% of the calls answered in 20 seconds'.

Square-root staffing rule

In the remainder of this section we take the delay probability as service measure. What is the least number c^* of servers such that the delay probability P_W is below a prespecified level α, e.g. $\alpha = 0.20$? From a numerical point of view it is of course no problem at all to find the exact value of c^* by searching over c in formula (5.1.11) for a given value of R $(= c\rho)$. However, for practitioners it is helpful to have an *insightful* approximation formula. Such a formula can be given by using the normal distribution. The formula is called the *square-root staffing rule*. This simple rule of thumb for staffing large call centres provides very useful information to the management. In its simplest form the square-root formula is obtained by approximating the $M/M/c$ queue with many servers by the $M/M/\infty$ queue. This approach was used in Example 1.1.3. However, this first-order approximation can considerably be improved by using a relation between Erlang's delay probability in the $M/M/c$ delay system and Erlang's loss probability in the $M/M/c/c$ loss system. The improved approximation to the least number c^* of servers such that $P_W \leq \alpha$ is given by the square-root formula

$$c^* \approx R + k_\alpha \sqrt{R}, \tag{5.3.1}$$

where the safety factor k_α is the solution of the equation

$$\frac{k\Phi(k)}{\varphi(k)} = \frac{1-\alpha}{\alpha} \tag{5.3.2}$$

with $\Phi(x)$ denoting the standard normal probability distribution function and $\varphi(x) = (1/\sqrt{2\pi})e^{-\frac{1}{2}x^2}$ denoting its density. It is important to note that the safety factor k_α does not depend on R. Also, it is interesting to point out the similarity of the square-root staffing rule with the famous rule for the reorder point s in the (s, Q)-inventory model with a service-level constraint. The factor k_α can be found by solving (5.3.2) by bisection. For example, for $\alpha = 0.8, 0.5, 0.2$ and 0.1 the safety factor k_α has the respective values 0.1728, 0.5061, 1.062 and 1.420. The approximation (5.3.1) clarifies the interplay of the process parameters and

Table 5.3.2 The exact and approximate values of c^*

	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$	
	exa	app	exa	app	exa	app
$R = 1$	2	2	3	3	3	3
$R = 5$	7	7	8	8	9	9
$R = 10$	12	12	14	14	16	15
$R = 50$	54	54	58	58	61	61
$R = 100$	106	106	111	111	115	115
$R = 250$	259	259	268	267	274	273
$R = 500$	512	512	525	524	533	532
$R = 1000$	1017	1017	1034	1034	1046	1045

increases the manager's intuitive understanding of the system. In particular, the square-root staffing rule quantifies the economies of scale in staffing levels that can be achieved by combining several call centres into a single call centre. To illustrate this, consider two identical call centres each having an offered load of R erlangs of work and each having the same service requirement $P_W \leq \alpha$. For two separate call centres a total of $2(R + k_\alpha\sqrt{R})$ agents is needed, whereas for one combined call centre $2R + k_\alpha\sqrt{2R}$ agents are needed. A reduction of $(2 - \sqrt{2})k_\alpha\sqrt{R}$ agents.

The quality of the approximation (5.3.1) is excellent. Rounding *up* the approximation for c^* to the nearest integer, numerical investigations indicate that the approximate value is equal to the exact value in most cases and is never off by more than 1. Table 5.3.2 gives the exact and approximate values of c^* for several values of R and α.

Derivation of the square-root formula

The following relation holds between the delay probability P_{delay} in the $M/M/c$ delay system and the loss probability P_{loss} in the $M/M/c/c$ loss system:

$$P_{loss} = \frac{(1 - \rho)P_{delay}}{1 - \rho P_{delay}}. \tag{5.3.3}$$

This relation can be directly verified from the explicit formulas (5.1.11) and (5.2.2) for P_{delay} and P_{loss}. In Section 9.8 we establish the relation (5.3.3) in a more general framework by showing that the state probabilities in a finite-capacity queue with Poisson arrivals are often proportional to the state probabilities in the corresponding infinite-capacity model. By formula (5.2.2),

$$P_{loss} = \frac{e^{-R}R^c/c!}{\sum_{k=0}^{c} e^{-R}R^k/k!}.$$

For fixed R, let the random variable X_R be Poisson distributed with mean R. Then the above formula for P_{loss} can be written as

$$P_{loss} = \frac{P\{X_R = c\}}{P\{X_R \leq c\}}.$$

A Poisson distribution with mean R can be approximated by the normal distribution with mean R and standard deviation R when R is large. Now take

$$c = R + k\sqrt{R}$$

for some constant k. Then $P\{X_R \leq c\} = P\{(X_R - R)/\sqrt{R} \leq k\}$ and so, by the normal approximation to the Poisson distribution,

$$P\{X_R \leq c\} \approx \Phi(k).$$

Writing $P\{X_R = c\} = P\{c - 1 < X_R \leq c\}$, we also have that

$$P\{X_R = c\} = P\left\{k - \frac{1}{\sqrt{R}} < \frac{X_R - R}{\sqrt{R}} \leq k\right\} \approx \Phi(k) - \Phi\left(k - \frac{1}{\sqrt{R}}\right) \approx \frac{1}{\sqrt{R}}\varphi(k).$$

This gives

$$P_{loss} \approx \frac{1}{\sqrt{R}} \frac{\varphi(k)}{\Phi(k)}. \tag{5.3.4}$$

By (5.3.3) and $\rho = R/c$, we have $P_{delay} = cP_{loss}/(c - R + RP_{loss})$. Substituting $c = R + k\sqrt{R}$ in this formula, noting that $k\sqrt{R} << R$ for R large and using (5.3.4), we find with the abbreviation $z = \varphi(k)/\Phi(k)$ that

$$P_{delay} \approx \frac{(R + k\sqrt{R})z}{kR + Rz} \approx \frac{Rz}{kR + Rz} = \left(1 + \frac{k}{z}\right)^{-1} = \left[1 + \frac{k\Phi(k)}{\varphi(k)}\right]^{-1}. \tag{5.3.5}$$

Equating the last term to α gives the relation (5.3.2). This completes the derivation of the square-root formula (5.3.1).

5.4 INSENSITIVITY

In many stochastic service systems in which arriving customers *never queue*, it turns out that the performance measures are insensitive to the form of the service-time distribution and require only the mean of the service time. The most noteworthy examples of such service systems are infinite-server systems and loss systems. In the $M/G/\infty$ queue with Poisson arrivals and infinitely many servers, rather simple arguments enable us to prove that the limiting distribution of the number of busy servers is insensitive to the form of the service-time distribution; see Section 1.1.3. The Erlang loss model with Poisson input and the Engset model with finite-source input provide other examples of stochastic service systems possessing the insensitivity property. Other examples of stochastic service systems having the insensitivity property will be given in this section and in the exercises. Nowadays

a well-developed theory for insensitivity is available; see Schassberger (1986) and Whittle (1985). This theory will not be discussed here. In this section the insensitivity property for the Erlang loss model and the Engset loss model is made plausible through a closed two-node network model. This model is also used to argue insensitivity in a controlled loss model with several customer classes. Also, the $M/G/1$ queue with the processor-sharing discipline is discussed as an example of a stochastic service system with no queueing and possessing the insensitivity property.

5.4.1 A Closed Two-node Network with Blocking

Consider a closed network model with two nodes in cyclic order. A fixed number of M jobs move around in the network. If a job has completed service at one of the nodes, it places a request for service at the other node. Node 1 is an *infinite-server* node, that is, there is an ample number of servers at node 1. Node 2 is the only node at which blocking can occur. A job that is accepted at node 2 is immediately provided with a free server. Further, it is assumed that there are r different job types $h = 1, \ldots, r$ with M_h jobs of type h, where $M_1 + \cdots + M_r = M$. The blocking protocol is as follows: if a job of type h arrives at node 2 when n_2 jobs are already present at node 2, including $n_2^{(h)}$ jobs of type h, then the arriving job of type h is accepted at node 2 with probability

$$A(n_2)A_h(n_2^{(h)}), \quad h = 1, \ldots, r \tag{5.4.1}$$

for given functions $A(.), A_1(.), \ldots, A_r(.)$. An accepted job is *immediately* provided a free server and receives *uninterrupted* service at a *constant* rate. If a job is rejected at node 2, it returns to node 1 and undergoes a complete *new service* at node 1. The service time of a job of type h at node i has a general probability distribution function with mean $1/\mu_{ih}$ for $i = 1, 2$ and $h = 1, \ldots, r$. For each type of job it is assumed that the service-time distribution for at least one of the nodes has a positive density on some interval. The service requirements at the nodes are assumed to be independent of each other.

The system is said to be in state $\mathbf{n} = (n_i^{(h)})$ when there are $n_i^{(h)}$ jobs present at node i for $i = 1, 2$ and $h = 1, \ldots, r$ with $n_1^{(h)} + n_2^{(h)} = M_h$ for $h = 1, \ldots, r$. Let $p(\mathbf{n})$ denote the limiting probability that the process is in state $\mathbf{n}$ at an arbitrary point in time. Also, for fixed job type ℓ, let $\pi_i^{(\ell)}(\tilde{\mathbf{n}})$ denote the limiting probability that a job of type ℓ arriving at node i finds the other jobs in state $\tilde{\mathbf{n}}$ with $\tilde{n}_1^{(\ell)} + \tilde{n}_2^{(\ell)} = M_\ell - 1$ and $\tilde{n}_1^{(h)} + \tilde{n}_2^{(h)} = M_h$ for $h \neq \ell$. Assuming that each of the service-time distributions is a mixture of Erlangian distributions with the same scale parameters, Van Dijk and Tijms (1986) used rather elementary arguments to prove that the probabilities $p(\mathbf{n})$ and $\pi_i^{(h)}(\tilde{\mathbf{n}})$ depend on the service-time distributions only through their means and are thus insensitive to the form of the service-time distributions.* Next, by deep mathematics, the insensitivity property for general

*Also the so-called product-form solution applies to these probabilities. The product-form solution will be discussed in detail in Section 5.6.

service-time distributions can be concluded by a continuity argument. This argument is based on the fact that the class of mixtures of Erlangian distributions with the same scale parameters is dense in the class of all probability distributions on the non-negative axis; see Hordijk and Schassberger (1982) and Whitt (1980). In Section 5.5 we give an elementary proof that any service-time distribution can be arbitrarily closely approximated by a mixture of Erlangian distributions with the same scale parameters. Taking for granted the insensitivity property of the closed two-node network model, we give two applications of loss systems with the insensitivity property.

Example 5.4.1 Insensitivity for a finite-source model with grading

Let us consider a finite-source model with grading. Such a model is an extension of the Engset model discussed in Section 5.2.2. In the Engset model a newly generated message is only blocked when all c servers are occupied. In the grading model a newly generated message hunts for a free server among K servers that are randomly chosen from the c servers, with K fixed. The message is blocked when no free server is found among the K chosen servers. The closed two-node model with a single job type applies ($r = 1$). The blocking protocol indeed allows for the representation (5.4.1). This follows by taking

$$A(n_2) = 1 - \binom{n_2}{K} \Big/ \binom{c}{K} \quad \text{and } A_h(n_2^{(h)}) = 1, \quad h = 1, \ldots, r$$

with the convention $\binom{n}{m} = 0$ for $n < m$. Thus we can conclude that the time-average and customer-average probabilities in the grading model are insensitive to both the form of the think-time distribution and the form of the service-time distribution. The Engset model is a special case of the grading model with $K = c$. Thus we also have insensitivity for the Engset model. By letting the number of sources tend to infinity and the thinking rate to zero, the input process becomes a Poisson process. It will now intuitively be clear that the Erlang loss model has the insensitivity property. However, a rigorous proof of this fact requires deep mathematics.

Example 5.4.2 A loss model with competing customers

Messages of types 1 and 2 arrive at a communication system according to two independent Poisson processes with the respective rates λ_1 and λ_2. The communication system has c identical service channels for handling the messages but there is no buffer to temporarily store messages which find all channels occupied. Each channel can handle only one message at a time. The transmission times of the messages are independent of each other and the transmission times of messages of the same type j have a general probability distribution with mean $1/\mu_j$ for $j = 1, 2$. The following admission rule for arriving messages is used. Messages of type 1 are always accepted whenever a free service channel is available. However, for a

given control parameter L, messages of type 2 are only accepted when less than L messages of type 2 are present and not all of the channels are occupied. Such a control rule is used to increase the throughput of accepted messages. What is the optimal value of L?

To compute the average throughput for a given L-policy, it is no restriction to assume exponentially distributed transmission times. The reason is that the long-run average throughput is insensitive to the form of the transmission time distributions. The average throughput is the difference between the average arrival rate $\lambda_1 + \lambda_2$ and the average number of messages lost per time unit. To argue that the loss probabilities for both types of messages are insensitive to the form of the transmission-time distribution, consider the finite-source variant of the model with Poisson input. Messages of type j are generated by M_j identical sources for $j = 1, 2$, where the think time of the sources has a probability density. A source can only start a think time when it has no message in transmission at the communication system. The sources act independently of each other. This finite-source model can be seen as a cyclic closed two-node network model, where a fixed number of type 1 jobs, M_1, and a fixed number of type 2 jobs, M_2, move around in the network. Node 1 is an infinite-server node, while node 2 is a blocking node with c servers. In the two-node closed network, take the blocking protocol (5.4.1) with

$$A(n_2) = \begin{cases} 1, & n_2 < c, \\ 0, & n_2 = c, \end{cases}$$

and

$$A_1(n_2^{(1)}) = 1, \quad A_2(n_2^{(2)}) = \begin{cases} 1 & \text{for } n_2^{(2)} < L, \\ 0 & \text{otherwise.} \end{cases}$$

The closed two-node network with this blocking protocol behaves identically to the finite-source model. Thus the finite-source model has the insensitivity property. This result provides a simple but heuristic argument that the controlled loss model with Poisson input also has the insensitivity property. In general, insensitivity holds for a wide class of loss networks; see Kelly (1991) and Ross (1995).

Let us now assume exponentially distributed transmission times for the loss model controlled by an L-policy. Define

$$X_j(t) = \text{the number of channels occupied by type } j \text{ messages at time } t$$

for $j = 1, 2$. The stochastic process $\{(X_1(t), X_2(t))\}$ is a continuous-time Markov chain with state space

$$I = \{(i_1, i_2) \mid 0 \le i_1 + i_2 \le c, i_1 \ge 0, 0 \le i_2 \le L\}.$$

Its transition rate diagram is given in Figure 5.4.1. By equating the rate out of state (i_1, i_2) to the rate into state (i_1, i_2), we obtain the equilibrium equations for the

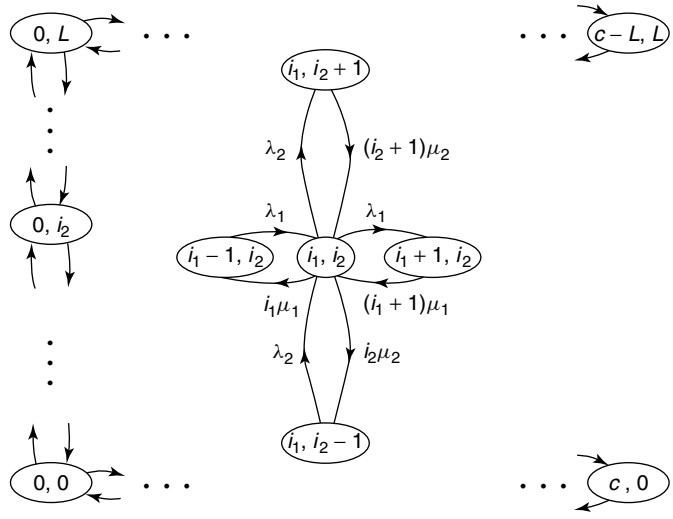


Figure 5.4.1 The transition rate diagram for the L-rule

state probabilities $p(i_1, i_2)$. For the states (i_1, i_2) with $i_1 + i_2 < c$ and $i_2 < L$,

$$\begin{aligned}(i_1\mu_1 + i_2\mu_2 + \lambda_1 + \lambda_2)p(i_1, i_2) = \lambda_1 p(i_1 - 1, i_2) + \lambda_2 p(i_1, i_2 - 1)\\ + (i_1 + 1)\mu_1 p(i_1 + 1, i_2)\\ + (i_2 + 1)\mu_2 p(i_1, i_2 + 1).\end{aligned}$$

For the states (i_1, i_2) with $i_1 + i_2 < c$ and $i_2 = L$,

$$\begin{aligned}(i_1\mu_1 + i_2\mu_2 + \lambda_1)p(i_1, i_2) = \lambda_1 p(i_1 - 1, i_2) + \lambda_2 p(i_1, i_2 - 1)\\ + (i_1 + 1)\,\mu_1 p(i_1 + 1, i_2).\end{aligned}$$

For the states (i_1, i_2) with $i_1 + i_2 = c$ and $i_2 \le L$,

$$(i_1\mu_1 + i_2\mu_2)p(i_1, i_2) = \lambda_1 p(i_1 - 1, i_2) + \lambda_2 p(i_1, i_2 - 1).$$

The state probabilities $p(i_1, i_2)$ exhibit the so-called product form

$$p(i_1, i_2) = C\frac{(\lambda_1/\mu_1)^{i_1}}{i_1!}\frac{(\lambda_2/\mu_2)^{i_2}}{i_2!}, \quad i_1, i_2 \in I$$

for some constant $C > 0$. The reader may verify this result by direct substitution into the equilibrium equations. Since service completions occur in state (i_1, i_2) at a rate of $i_1\mu_1 + i_2\mu_2$, the average throughput is given by

$$T(L) = \sum_{(i_1, i_2)} (i_1\mu_1 + i_2\mu_2)p(i_1, i_2).$$

Denote by $\Pi_j(L)$ the long-run fraction of type j messages that are lost. Using the PASTA property, it follows that

$$\Pi_1(L) = \sum_{\substack{(i_1,i_2):\\ i_1+i_2=c}} p(i_1,i_2) \quad \text{and} \quad \Pi_2(L) = \sum_{i_1=0}^{c-L-1} p(i_1,L) + \sum_{\substack{(i_1,i_2):\\ i_1+i_2=c}} p(i_1,i_2)$$

Since the sum of the average number of messages lost per time unit and the average number of messages transmitted per time unit equals the arrival rate $\lambda_1 + \lambda_2$, we have the identity $\lambda_1\Pi_1(L) + \lambda_2\Pi_2(L) + T(L) = \lambda_1 + \lambda_2$. This relation is useful as an accuracy check for the calculated values of the $p(i_1, i_2)$. As an illustration, we consider the following numerical data:

$$c = 10,\ \lambda_1 = 10,\ \lambda_2 = 7,\ \mu_1 = 10,\ \mu_2 = 1.$$

Table 5.4.1 gives the values of $T(L)$, $\Pi_1(L)$ and $\Pi_2(L)$ for $L = 7, 8$ and 9.

The L-policy with $L = 8$ maximizes the long-run average throughput among the class of L-policies. The above analysis restricted itself to the easily implementable L-policies, but other control rules are conceivable. The question of how to compute the overall optimal control rule among the class of all conceivable control rules will be addressed in the Chapters 6 and 7, which deal with Markov decision processes. The best L-policy is in general not optimal among the class of all possible control rules. However, numerical investigations indicate that using the best L-policy rather than the overall optimal policy often leads to only a small deviation from the theoretically maximal average throughput. For example, for the above numerical data the average throughput of 15.209 for the best L-policy is only 0.16% below the theoretically optimal value of 15.233. This optimal value is achieved by the following control rule. Each arriving message of type 1 is accepted as long as not all channels are occupied. A message of type 2 finding i messages of type 1 present upon arrival is accepted only when less than L_i other messages of the same type 2 are present and not all of the channels are occupied. The optimal values of the L_i are $L_0 = L_1 = 8$, $L_2 = L_3 = 7$, $L_4 = 6$, $L_5 = 5$, $L_6 = 4$, $L_7 = 3$, $L_8 = 2$ and $L_9 = 1$. The insensitivity property is no longer exactly true for the L_i-policy, but numerical investigations indicate that the dependency on the distributional form of the transmission times is quite weak. The above L_i-policy was simulated for lognormally distributed transmission times. Denoting by c_i^2 the squared coefficient of variation of the transmission time for type i messages, we varied (c_1^2, c_2^2) as (1, 1), (2, 0.5) and (0.5, 2). For these three examples the average

Table 5.4.1 Numerical values

L	$T(L)$	$\Pi_1(L)$	$\Pi_2(L)$
7	15.050	0.0199	0.2501
8	15.209	0.0501	0.1843
9	15.095	0.0926	0.1399

throughputs of the given L_i-policy have the respective values 15.236 ($\pm$0.004), 15.244 ($\pm$0.004) and 15.231 ($\pm$0.005), where the numbers in parentheses indicate the 95% confidence intervals.

5.4.2 The $M/G/1$ Queue with Processor Sharing

Another queueing system in which the limiting distribution of the number of customers in the system is insensitive to the service-time distribution is the $M/G/1$ queue with the processor-sharing service discipline. Under this service discipline a customer never has to wait in queue and the processing rate of the server is equally divided among all customers present. The $M/G/1$ processor-sharing system can be used to approximate time-shared computer systems among others. To formulate the model, assume that customers arrive according to a Poisson process with rate λ and that the service requirements of the customers are independent random variables which are distributed according to the random variable S. It is assumed that S has a general probability distribution. A *generalized processor-sharing rule* is used: if i customers are present, each of the i customers is provided with service at a rate of $f(i)$ per time unit. That is, the attained service time of each of the i customers grows by an amount $f(i)\Delta x$ in a time Δx with Δx small. Here $f(i)$ is a given positive function. Let $\rho = \lambda E(S)$ denote the offered load and let

$$\phi(j) = \left\{ \prod_{k=1}^{j} f(k) \right\}^{-1}, \quad j = 0, 1, \ldots$$

with $\phi(0) = 1$ by convention. Assuming that $\sum_{k=0}^{\infty} \rho^k \phi(k)/k!$ is finite, it holds that the limiting distribution $\{p_j,\ j = 0, 1, \ldots\}$ of the number of customers present is insensitive to the form of the service-requirement distribution and is given by

$$p_j = \frac{(\rho^j/j!)\phi(j)}{\sum_{k=0}^{\infty}(\rho^k/k!)\phi(k)}, \quad j = 0, 1, \ldots .$$

A proof of this result can be found in Cohen (1979). Denoting by $E(W \mid s)$ the expected amount of time spent in the system by a customer who arrives when the system has reached statistical equilibrium and whose required service time is s, it was also shown in Cohen (1979) that

$$E(W \mid s) = \frac{s \displaystyle\sum_{k=0}^{\infty}(\rho^k/k!)\phi(k+1)}{\displaystyle\sum_{k=0}^{\infty}(\rho^k/k!)\phi(k)}, \quad s > 0,$$

This remarkable result shows that the processor-sharing rule discriminates between customers in a fair way. A customer requiring a service time twice as long as some other will spend on average twice as long in the system. The standard $M/G/1$

processor-sharing queue corresponds to the case of

$$f(i) = \frac{1}{i}, \quad i = 1, 2, \ldots .$$

In this case $\phi(i) = i!$ for $i = 0, 1, \ldots$ and the above formulas reduce to

$$p_j = (1 - \rho)\rho^j, \quad j = 0, 1, \ldots \quad \text{and} \quad E(W \mid s) = \frac{s}{1 - \rho}, \quad s > 0.$$

In other words, in the standard $M/G/1$ processor-sharing queue with general service times, the equilibrium distribution of the number of customers present is the same as in the $M/M/1$ queue with the first-come first-served discipline. This finding also applies to the $M/G/1$ queue with the *pre-emptive resume, last-in first-out* discipline. Under this service discipline each customer begins service upon arrival, pre-empting anyone in service, and at each time, the most recently arrived customer receives service.

5.5 A PHASE METHOD

The phase method makes it possible to use the continuous-time Markov chain approach for a wide variety of practical probability problems in which the underlying probability distributions are not necessarily exponential. The method essentially goes back to A.K. Erlang, who did pioneering work on stochastic processes at the beginning of the twentieth century. In his analysis of telephone problems, Erlang devised the trick of considering the duration of a call as the sum of a number of sequential phases whose lengths are exponentially distributed. There are several versions of the phase method (or method of stages). A very useful version is the one that approximates a positive random variable by a random sum of exponentials with the same means. In other words, the probability distribution of the positive random variable is approximated by a mixture of Erlangian distributions with the *same* scale parameters. The theoretical basis for the use of such mixtures of Erlangian distributions is provided by the following theorem.

Theorem 5.5.1 *Let $F(t)$ be the probability distribution function of a positive random variable. For fixed $\Delta > 0$ define the probability distribution function $F_\Delta(x)$ by*

$$F_\Delta(x) = \sum_{j=1}^{\infty} p_j(\Delta) \left\{ 1 - \sum_{k=0}^{j-1} e^{-x/\Delta} \frac{(x/\Delta)^k}{k!} \right\}, \quad x \geq 0, \tag{5.5.1}$$

where $p_j(\Delta) = F(j\Delta) - F((j-1)\Delta)$, $j = 1, 2, \ldots$. Then

$$\lim_{\Delta \to 0} F_\Delta(x) = F(x)$$

for each continuity point x of $F(t)$.

Proof For fixed $\Delta, x > 0$, let $U_{\Delta,x}$ be a Poisson distributed random variable with

$$P\{U_{\Delta,x} = k\Delta\} = e^{-x/\Delta}\frac{(x/\Delta)^k}{k!}, \quad k = 0, 1, \ldots .$$

It is immediately verified that $E(U_{\Delta,x}) = x$ and $\sigma^2(U_{\Delta,x}) = x\Delta$. Let $g(t)$ be any bounded function. We now prove that

$$\lim_{\Delta\to 0} E[g(U_{\Delta,x})] = g(x) \tag{5.5.2}$$

for each continuity point x of $g(t)$. To see this, fix $\varepsilon > 0$ and a continuity point x of $g(t)$. Then there exists a number $\delta > 0$ such that $|g(t) - g(x)| \le \varepsilon/2$ for all t with $|t - x| \le \delta$. Also, let $M > 0$ be such that $|g(t)| \le M/2$ for all t. Then

$$\begin{aligned} |E[g(U_{\Delta,x})] - g(x)| &\le \sum_{k=0}^{\infty} |g(k\Delta) - g(x)|\, P\{U_{\Delta,x} = k\Delta\} \\ &\le \frac{\varepsilon}{2} + M \sum_{k:|k\Delta - x| > \delta} P\{U_{\Delta,x} = k\Delta\} \\ &= \frac{\varepsilon}{2} + MP\{|U_{\Delta,x} - E(U_{\Delta,x})| > \delta\}. \end{aligned}$$

By Chebyshev's inequality, $P\{|U_{\Delta,x} - E(U_{\Delta,x})| > \delta\} \le x\Delta/\delta^2$. For Δ small enough, we have $Mx\Delta/\delta^2 \le \frac{1}{2}\varepsilon$. This proves the relation (5.5.2). Next, we apply (5.5.2) with $g(t) = F(t)$. Hence, for any continuity point x of $F(t)$,

$$\begin{aligned} F(x) &= \lim_{\Delta\to 0} E[F(U_{\Delta,x})] = \lim_{\Delta\to 0} \sum_{k=0}^{\infty} F(k\Delta) e^{-x/\Delta}\frac{(x/\Delta)^k}{k!} \\ &= \lim_{\Delta\to 0} \sum_{k=0}^{\infty} e^{-x/\Delta}\frac{(x/\Delta)^k}{k!} \sum_{j=1}^{k} p_j(\Delta), \end{aligned}$$

where the latter equality uses that $F(0) = 0$. Interchanging the order of summation, we next obtain

$$F(x) = \lim_{\Delta\to 0} \sum_{j=1}^{\infty} p_j(\Delta) \sum_{k=j}^{\infty} e^{-x/\Delta}\frac{(x/\Delta)^k}{k!},$$

yielding the desired result.

The proof of Theorem 5.5.1 shows that the result also holds when $F(t)$ has a positive mass at $t = 0$. We should then add the term $F(0)$ to the right-hand side of (5.5.1). Roughly stated, Theorem 5.5.1 tells us that the probability distribution of any positive random variable can be arbitrarily closely approximated by a mixture of Erlangian distributions with the *same* scale parameters. The fact that the Erlangian distributions have identical scale parameters simplifies the construction

of an appropriate continuous-time Markov chain in specific applications. In practice it is not always obvious how to choose a mixture that is sufficiently close to the distribution considered. One often confines oneself to a mixture of two Erlangian distributions by matching only the first two moments of the distribution considered; see Appendix B.

The phase method is very useful both for theoretical purposes and practical purposes. We give two examples to illustrate its power.

Example 5.5.1 The $M/G/1$ queue and the phase method

Customers arrive at a single-server station according to a Poisson process with rate λ. The service times of the customers are independent and identically distributed random variables and are also independent of the arrival process. The single server can handle only one customer at a time and customers are served in order of arrival. The phase method will be applied to obtain a computationally useful representation of the waiting-time distribution of a customer when the probability distribution of the service time of a customer is given by

$$P\{S \le x\} = \sum_{j=1}^{\infty} \beta_j \left(1 - \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!} \right), \quad x \ge 0, \tag{5.5.3}$$

where $\beta_j \ge 0$ and $\sum_{j=1}^{\infty} \beta_j = 1$. The random variable S denotes the service time. It is assumed that $\lambda E(S) < 1$. In view of (5.5.3) we can think of the service time of a customer as follows. With probability β_j the customer has to go through j sequential service phases before its service is completed. The phases are processed one at a time and their durations are independent and exponentially distributed random variables with mean $1/\mu$. This interpretation enables us to define a continuous-time Markov chain. For any $t \ge 0$, let

$$X(t) = \text{the number of uncompleted service phases present at time } t.$$

The process $\{X(t)\}$ is a continuous-time Markov chain with infinite state space $I = \{0, 1, \dots\}$. Its transition rate diagram is displayed in Figure 5.5.1.

Denote the equilibrium distribution of the process $\{X(t)\}$ by $\{f_j, j = 0, 1, \dots\}$. The time-average probability f_j denotes the long-run fraction of time there are j

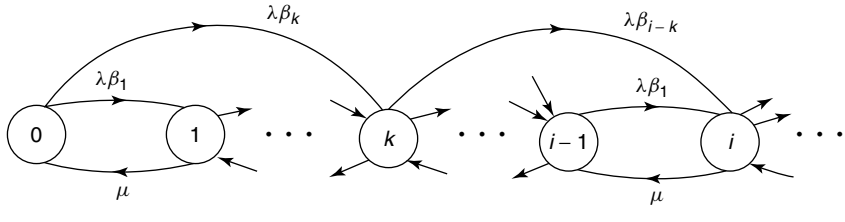


Figure 5.5.1 The transition diagram of the phase process

uncompleted service phases present. To find the waiting-time distribution, we need the customer-average probabilities

$$\pi_j = \text{the long-run fraction of customers who find } j \text{ uncompleted service phases present upon arrival, } j = 0, 1, \ldots .$$

Under the assumption of service in order of arrival, let

$$W_q(x) = \lim_{n\to\infty} P\{D_n \le x\}, \quad x \ge 0$$

with D_n denoting the delay in queue of the nth arriving customer. In the same way as (5.1.7) was derived in Section 5.1, it can be shown that this limit exists and is given by

$$W_q(x) = 1 - \sum_{j=1}^{\infty} \pi_j \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!}, \quad x \ge 0. \tag{5.5.4}$$

By the PASTA property, we have

$$\pi_j = f_j, \quad j = 1, 2, \ldots .$$

The probabilities f_j allow for a recursive computation, since the transition rates of the continuous-time Markov chain $\{X(t)\}$ have the property that $q_{ij} = 0$ for $j \le i - 2$. By equating the rate at which the process leaves the set of states $\{i, i+1, \ldots\}$ to the rate at which the process enters this set, we obtain

$$\mu f_i = \sum_{k=0}^{i-1} f_k \left(1 - \sum_{j=0}^{i-k-1} \beta_j\right), \quad i = 1, 2, \ldots . \tag{5.5.5}$$

This recursion provides an effective method for computing the f_j. Note that the recursion can be initialized with $f_0 = 1 - \lambda E(S)$, since by Little's formula the long-run fraction of time the server is busy equals $\lambda E(S)$. Note that $E(S) = (1/\mu)\sum_{j=1}^{\infty} j\beta_j$. Once the probabilities π_j $(= f_j)$ have been computed by applying (5.5.5), the waiting-time probability $W_q(x)$ can be calculated from (5.5.4).

The expression (5.5.4) for $W_q(x)$ is very useful for computational purposes. For numerical calculations it is recommended to rewrite (5.5.4) as

$$W_q(x) = 1 - \sum_{k=0}^{\infty} e^{-\mu x} \frac{(\mu x)^k}{k!} \sum_{j=k+1}^{\infty} f_j, \quad x \ge 0, \tag{5.5.6}$$

by interchanging the order of summation. The series representation (5.5.6) converges faster than the series (5.5.4). Of course $\sum_{j=k+1}^{\infty} f_j$ should be replaced by $1 - \sum_{j=0}^{k} f_j$ in (5.5.6). The computational work in (5.5.5) and (5.5.6) can be reduced

by using asymptotic expansions for f_j as $j \to \infty$ and $1 - W_q(x)$ as $x \to \infty$; see Exercise 5.26.

Example 5.5.2 A finite-buffer storage problem

Data messages arrive at a transmission channel according to a Poisson process with rate λ. The transmission channel has a buffer to store arriving messages. The buffer has a finite capacity $K > 0$. An arriving message is only stored in the buffer when its length does not exceed the unoccupied buffer capacity, otherwise the whole message is rejected. Data are transmitted from the buffer at a constant rate of $\sigma > 0$. The message lengths are independent of each other and are assumed to have a continuous probability distribution function $F(x)$. An important performance measure is the long-run fraction of messages that are rejected. This model, which is known as the $M/G/1$ queue with *bounded sojourn time*, is very useful. It also applies to a finite-capacity production/inventory system in which production occurs at a constant rate as long as the inventory is below its maximum level and the demand process is a compound Poisson process, where demands occurring when the system is out of stock are completely lost.

A possible approach to solving the model is to discretize the model; see Exercise 9.9 for another approach. In the discretized model a message is represented by a batch consisting of a discrete number of data units. The probability of a batch of size k is given by

$$b_k(\Delta) = F(k\Delta) - F((k-1)\Delta), \quad k = 1, 2, \ldots$$

with $F(-\Delta) = 0$. The buffer only has room for $K(\Delta)$ data units, where

$$K(\Delta) = \frac{K}{\Delta}.$$

It is assumed that the number Δ is chosen such that $K(\Delta)$ is an integer. An arriving message is only stored in the buffer when its batch size does not exceed the number of unoccupied buffer places, otherwise the whole message is rejected. The data units are transmitted one at a time at a constant rate of $\sigma > 0$. The key step is now to take an *exponential distribution* with mean $1/\mu(\Delta) = \Delta/\sigma$ for the transmission time of a data unit. This approach is motivated by Theorem 5.5.1. A data unit leaves the buffer as soon as its transmission is completed. For the discretized model, let

$$\pi_\Delta(K) = \text{the long-run fraction of messages that are rejected.}$$

In view of Theorem 5.5.1 one might expect that $\pi_\Delta(K)$ is an excellent approximation to the rejection probability in the original model when Δ is chosen sufficiently

small. The discretized rejection probability $\pi_\Delta(K)$ is routinely found by using the continuous-time Markov chain approach. In the discretized model, let the random variable

$$X(t) = \text{the number of data units in the buffer at time } t.$$

The process $\{X(t)\}$ is a continuous-time Markov chain with the finite state space $I = \{0, 1 \ldots, K(\Delta)\}$. Denote the equilibrium distribution of the discretized process by $\{p_j(\Delta)\}$. The process has the property that for each state i the transition rate $q_{ij} = 0$ for $j \leq i - 2$. Hence $p_j(\Delta)$ can recursively be computed. By equating the rate out of the set $\{i, \ldots, K(\Delta)\}$ to the rate into this set,

$$\mu(\Delta) p_i(\Delta) = \sum_{j=0}^{i-1} p_j(\Delta) \left[\lambda \sum_{k=i-j}^{K(\Delta)-j} b_k(\Delta) \right], \quad i = 1, 2, \ldots, K(\Delta).$$

Using the PASTA property, we next obtain $\pi_\Delta(K)$ from

$$\pi_\Delta(K) = \sum_{i=0}^{K(\Delta)} p_i(\Delta) \sum_{k > K(\Delta)-i} b_k(\Delta).$$

The computational work is considerably reduced by noting that

$$\sum_{k=\ell}^{\infty} b_k(\Delta) = 1 - \sum_{k=0}^{\ell-1} b_k(\Delta) = 1 - F((\ell - 1)\Delta), \quad \ell = 1, 2, \ldots .$$

The accuracy of the discretization is improved by slightly modifying the definition of the batch-size probabilities $b_k(\Delta)$. It is recommended to take

$$b_k(\Delta) = \frac{1}{2}[F(k\Delta) - F((k-1)\Delta)] + \frac{1}{2}[F((k+1)\Delta) - F(k\Delta)]$$

for $k = 1, 2, \ldots$, in which case $\sum_{k \geq \ell} b_k(\Delta) = 1 - \frac{1}{2}F((\ell - 1)\Delta) - \frac{1}{2}F(\ell\Delta)$. It remains to decide how small to choose Δ in order to obtain a sufficiently close approximation to the rejection probability in the original model. In general one should search for a value of Δ such that the answers for the values Δ and $\Delta/2$ are sufficiently close to each other.

5.6 QUEUEING NETWORKS

Queueing network models are a useful analysis tool in a wide variety of areas such as computer performance evaluation, communication network design and production planning in flexible manufacturing. Generally speaking, a network of queues is a collection of service nodes with customers (jobs) moving between the nodes and making random requests for service at the nodes. Under appropriate conditions these networks can be modelled and analysed by means of continuous-time

Markov chains. The prominent result of the analysis is the product-form solution for the joint distribution of the numbers of customers present at the various nodes. Networks that can be described by a continuous-time Markov chain and have the product-form solution are often called *Jackson networks* after J.R. Jackson (1957, 1963), who discovered the product-form solution. In Section 5.6.1 we consider the open network model. A network is called *open* if external arrivals occur at one or more nodes and departures from the system occur at one or more nodes. A network is called *closed* when a fixed number of customers move around in the network. The closed network will be analysed in Section 5.6.2. For clarity of presentation the analysis is restricted to a single class of customers. In applications, however, one often encounters networks of queues with several customer classes. The results presented in this section can be extended to the case of multiple customer classes.

5.6.1 Open Network Model

As a prelude to the open queueing network model, consider the following medical application involving the analysis of emergency facilities. Patients arrive at an emergency room for late-night operations. Incoming patients are initially screened to determine their level of severity. On average, 10% of incoming patients require hospital admission. Twenty percent of incoming patients are sent to the ambulatory unit, 30% to the X-ray unit and 40% to the laboratory unit. Patients sent to the ambulatory unit are released after having received ambulatory care. Of those going to the X-ray unit, 25% require admission to the hospital, 20% are sent to the laboratory unit for additional testing, and 55% have no need of additional care and are thus released. Of patients entering the laboratory unit, 15% require hospitalization and 85% are released. This emergency system provides an example of a network of queues.

Consider now the following model for an open network of queues (*open Jackson network*):

- The network consists of K service stations numbered as $j = 1, \ldots, K$.
- External arrivals of new customers occur at stations $1, \ldots, K$ according to independent Poisson processes with respective rates $r_1, \ldots, r_K$.
- Each station is a single-server station with ample waiting room and at each station service is in order of arrival.
- The service times of the customers at the different visits to the stations are independent of each other, and the service time of a customer at each visit to station j has an exponential distribution with mean $1/\mu_j$ for $j = 1, \ldots, K$.
- Upon service completion at station i, the served customer moves with probability p_{ij} to station j for $j = 1, \ldots, K$ or leaves the system with probability $p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$.

The routing matrix $\mathbf{P} = (p_{ij})$, $i, j = 1, \ldots, K$, is assumed to be an irreducible substochastic matrix with the property that $\mathbf{P}^n \to \mathbf{0}$ as $n \to \infty$. Thus each newly arriving customer ultimately leaves the system with probability 1. To ensure that the process describing the numbers of customers present at the various stations has an equilibrium distribution, we need an assumption involving the composite (external and internal) arrival rates at the stations. Define the composite rates $\lambda_1, \ldots, \lambda_K$ as the unique solution to the linear equations

$$\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i p_{ij}, \quad j = 1, \ldots, K. \tag{5.6.1}$$

This system of linear equations has a unique solution since the matrix $\mathbf{P}$ is transient and so $(\mathbf{I} - \mathbf{P})^{-1}$ exists. The assumption is made that

$$\frac{\lambda_j}{\mu_j} < 1, \quad j = 1, \ldots, K. \tag{5.6.2}$$

The quantity λ_j can be interpreted as the total arrival rate at station j. In the long run we have for each station i that the average number of arrivals per time unit at station i must be equal to the average number of service completions per time unit at station i. In particular, $\lambda_i p_{ij}$ is the arrival rate of customers to station j of those coming from station i. Hence the total arrival rate at station j must satisfy (5.6.1). The equations (5.6.1) are called the *traffic equations.*

For $j = 1, \ldots, K$, define the random variable

$$X_j(t) = \text{the number of customers present at station } j \text{ at time } t.$$

The multidimensional process $X(t) = \{(X_1(t), \ldots, X_K(t))\}$ is a continuous-time Markov chain with state space $I = \{(n_1, \ldots, n_K) \mid n_1 > 0, \ldots, n_K > 0\}$. Since the routing probability p_{ii} is allowed to be positive, self-transitions can occur in the process $\{X(t)\}$. Under assumption (5.6.2) the process $\{X(t)\}$ has a unique equilibrium distribution to be denoted by $p(n_1, \ldots, n_K)$. We now state Theorem 5.6.1.

Theorem 5.6.1 *The equilibrium probabilities $p(n_1, \ldots, n_K)$ have the product-form property*

$$p(n_1, \ldots, n_K) = \prod_{k=1}^{K} \left(1 - \frac{\lambda_k}{\mu_k}\right) \left(\frac{\lambda_k}{\mu_k}\right)^{n_k}. \tag{5.6.3}$$

Proof Let us use the shorthand notation $\mathbf{n} = (n_1, \ldots, n_K)$. Let $\mathbf{e}_i$ denote the ith unit vector, that is, the ith component of $\mathbf{e}_i$ is 1 and the other components are zero. By equating the rate out of state $\mathbf{n}$ to the rate into state $\mathbf{n}$ (including self-transitions),

we get for the process $\{X(t)\}$ the equilibrium equations

$$p(\mathbf{n})\sum_{j=1}^{k} r_j + p(\mathbf{n})\sum_{j:n_j>0}\mu_j = \sum_{j:n_j>0}\left[\sum_{i=1}^{K} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j)\mu_i p_{ij} + p(\mathbf{n}-\mathbf{e}_j)r_j\right] + \sum_{j=1}^{K} p(\mathbf{n}+\mathbf{e}_j)\mu_j p_{j0}.$$

These equations are certainly satisfied by

$$p(\mathbf{n}) = \prod_{k=1}^{K}\left(1-\frac{\lambda_k}{\mu_k}\right)\left(\frac{\lambda_k}{\mu_k}\right)^{n_k} \tag{5.6.4}$$

when this product-form solution satisfies the partial balance equations

$$p(\mathbf{n})\sum_{j=1}^{K} r_j = \sum_{j=1}^{K} p(\mathbf{n}+\mathbf{e}_j)\mu_j p_{j0}, \tag{5.6.5}$$

$$p(\mathbf{n})\mu_j = \sum_{i=1}^{K} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j)\mu_i p_{ij} + p(\mathbf{n}-\mathbf{e}_j)r_j, \quad 1 \le j \le K. \tag{5.6.6}$$

For the product-form solution (5.6.4) we have

$$p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j) = \left(\frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_j}{\mu_j}\right)^{-1} p(\mathbf{n}) \quad \text{and} \quad p(\mathbf{n}-\mathbf{e}_j) = \left(\frac{\lambda_j}{\mu_j}\right)^{-1} p(\mathbf{n}). \tag{5.6.7}$$

After substitution of (5.6.7) in (5.6.6), it remains to verify whether the relation

$$\mu_j = \sum_{i=1}^{K}\left(\frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_j}{\mu_j}\right)^{-1}\mu_i p_{ij} + \left(\frac{\lambda_j}{\mu_j}\right)^{-1} r_j \tag{5.6.8}$$

holds for each $j = 1, \ldots, K$. This is indeed true since the relation (5.6.8) coincides with the traffic equation (5.6.1) after cancelling out common terms (verify). In a similar way we can verify that (5.6.5) holds. Substituting $p(\mathbf{n}+\mathbf{e}_j) = (\lambda_j/\mu_j)p(\mathbf{n})$ into (5.6.5), we get

$$\sum_{j=1}^{K} r_j = \sum_{j=1}^{K} \lambda_j p_{j0}.$$

This relation is indeed true since it states that the rate of new customers entering the system equals the rate of customers leaving the system. This completes the proof.

The partial balance equations (5.6.5) and (5.6.6) are characteristic for the product-form solution. These equations express that

$$\begin{aligned}&\text{the rate out of a state due to a change at node } j\\ &\quad = \text{the rate into that state due to a change at node } j\end{aligned} \tag{5.6.9}$$

for each $j = 0, 1, \ldots, K$, where node 0 corresponds to the outside world. This property of *node local balance* is in general not satisfied in a stochastic network, but can indeed be verified for the Jackson network model. The product-form solution (5.6.3) can be expressed as

$$p(\mathbf{n}) = p_1(n_1) \cdots p_K(n_K), \tag{5.6.10}$$

where for any k the probability distribution $\{p_k(n), n = 0, 1, \ldots\}$ of the number of customers present at station k is the same as the equilibrium distribution of the number of customers present in an $M/M/1$ queue with arrival rate λ_k and service rate μ_k. In other words, in steady state the number of customers at the different service stations are *independent* of each other and the number at station k behaves *as if* station k is an $M/M/1$ queue with arrival rate λ_k and service rate μ_k. The result (5.6.10) is remarkable in the sense that in the network model the composite arrival process at station k is in general *not* a Poisson process. An easy counterexample is provided by a single-station network with *feedback*; that is, a customer served at the station goes immediately back to the station with a positive probability. Suppose that in this network the arrival rate from outside is very small and the service rate is very large. Then, if the feedback probability is close to 1, two consecutive arrivals at the station are highly correlated and so the arrival process is not Poisson.

The Jackson network model can be generalized to allow each service station to have multiple servers with exponential service times. If station j has c_j servers, the ergodicity condition (5.6.2) is replaced by $\lambda_j/(c_j\mu_j) < 1$. Then the node local balance equation (5.6.9) can again be verified and the equilibrium distribution $\{p(\mathbf{n})\}$ of the numbers of customers present at the different stations has the product form (5.6.10), where the probability distribution $\{p_k(n)\}$ of the number of customers present at station k is the same as the equilibrium distribution of the number of customers present in an $M/M/c$ queue with arrival rate λ_k, service rate μ_k and $c = c_k$ servers. Note that the multi-server $M/M/c$ queue with service rate μ can be regarded as a single-server queue with state-dependent service rate $\mu(n) = \min(n, c)\mu$ when n customers are present. Indeed it can be shown that the product-form solution also applies to the Jackson network model with state-dependent service rates provided that the service rate at each station depends only on the number of customers present at that station. More about the product-form solution and its ramifications can be found in the books of Boucherie (1992) and Van Dijk (1993). In these references the product-form solution is also linked to the concept of insensitivity. Insensitivity of the stochastic network holds when the condition of node local balance is sharpened to job local balance, requiring that

the rate out of a state due to a particular job is equal to the rate into that state due to that same job.

BCMP extension for the product-form solution

The product form has been established under the assumption that each service station has the first-come first-served discipline and that the service times are exponentially distributed. In an important paper of Baskett *et al.* (1975) it has been shown that the product-form solution (5.6.10) also holds when each service station uses one of the following four service disciplines, or *BCMP disciplines*:

1. The service discipline is first-come-first-served and the service times of the customers are exponentially distributed (multiple servers or state-dependent service is allowed).
2. The service discipline is processor-sharing; that is, if n customers are present at the station, each customer is served and receives service at a rate of $1/n$. The service time of a customer is allowed to have a general probability distribution.
3. The service discipline is determined by an infinite number of servers; that is, each arriving customer gets immediately assigned a free server. The service time of a customer is allowed to have a general probability distribution.
4. The service discipline is pre-emptive resume, last-in first-out; that is, customers are served one at a time in reverse order of arrival and a newly arriving customer gets immediate service, pre-empting anyone in service. The service time of a customer is allowed to have a general probability distribution.

The product-form solution (5.6.10) remains valid but the marginal probability distribution $\{p_k(n), n = 0, 1, \dots\}$ of the number of customers present at station k depends on the service discipline at station k. Under service discipline 1 with c_k identical servers, the marginal distribution $\{p_k(n)\}$ is given by the equilibrium distribution of the number of customers present in the $M/M/c$ queue with arrival rate $\lambda = \lambda_k$, service rate $\mu = \mu_k$ and $c = c_k$ servers. Under service discipline 3 at station k the marginal distribution $\{p_k(n)\}$ is given by the Poisson distribution with mean $\lambda_k E(S_k)$, where the random variable S_k denotes the service time of a customer at each visit to station k. Under both service discipline 2 and service discipline 4, at station k the marginal distribution $\{p_k(n)\}$ is given by the geometric distribution $\{(1 - \rho_k)\rho_k^n, n = 0, 1, \dots\}$ with $\rho_k = \lambda_k E(S_k)$, where S_k denotes the service time of a customer at each visit to station k.

5.6.2 Closed Network Model

In the performance evaluation of computer systems and flexible manufacturing systems it is often more convenient to consider a closed network with a *fixed* number of customers (jobs). A job may leave the system but is then immediately replaced by a new one. The basic *closed Jackson network* is as follows:

- The network consists of K service stations numbered as $j = 1, \dots, K$.

- A fixed number of M identical customers move around in the network.
- Each station is a single-server station with ample waiting room and at each station service is in order of arrival.
- The service times of the customers at the different visits to the stations are independent of each other, and the service time of a customer at station j has an exponential distribution with mean $1/\mu_j$ for $j = 1, \ldots, K$.
- Upon service completion at station i, the served customer moves with probability p_{ij} to station j for $j = 1, \ldots, K$, where $\sum_{j=1}^{K} p_{ij} = 1$ for all $i = 1, \ldots, K$.

The routing matrix $\mathbf{P} = (p_{ij}), i, j = 1, \ldots, K$ is assumed to be an irreducible Markov matrix. Since the Markov matrix $\mathbf{P}$ is irreducible, its equilibrium distribution $\{\pi_j\}$ is the unique positive solution to the equilibrium equations

$$\pi_j = \sum_{i=1}^{K} \pi_i p_{ij}, \quad j = 1, \ldots, K \tag{5.6.11}$$

in conjunction with the normalizing equation $\sum_{j=1}^{K} \pi_j = 1$. The relative visit frequencies to the stations are proportional to these equilibrium probabilities. To see this, let

$$\lambda_j = \text{the long-run average arrival rate of customers at station } j.$$

Since λ_i is also the rate at which customers depart from station i, we have that $\lambda_i p_{ij}$ is the rate at which customers arrive at station j from station i. This gives the *traffic equations*

$$\lambda_j = \sum_{i=1}^{K} \lambda_i p_{ij}, \quad j = 1, \ldots, K. \tag{5.6.12}$$

The solution of the equilibrium equations (5.6.11) of the Markov matrix $\mathbf{P}$ is unique up to a multiplicative constant. Hence, for some constant $\gamma > 0$,

$$\lambda_j = \gamma \pi_j, \quad j = 1, \ldots, K. \tag{5.6.13}$$

Denote by $X_j(t)$ the number of customers present at station j at time t. The process $\{(X_1(t), \ldots, X_K(t))\}$ is a continuous-time Markov chain with the finite state space $I = \{(n_1, \ldots, n_K) \mid n_i \geq 0, \sum_{i=1}^{K} n_i = M\}$.

Theorem 5.6.2 *The equilibrium distribution of the continuous-time Markov chain* $\{\mathbf{X}(t) = (X_1(t), \ldots, X_K(t))\}$ *is given by*

$$p(n_1, \ldots, n_K) = C \prod_{k=1}^{K} \left(\frac{\pi_k}{\mu_k}\right)^{n_k} \tag{5.6.14}$$

for some constant $C > 0$.

Proof The proof is along the same lines as that of Theorem 5.6.1. The equilibrium equations of the Markov process $\{\mathbf{X}(t)\}$ are given by

$$p(\mathbf{n}) \sum_{j:n_j>0} \mu_j = \sum_{j:n_j>0} \left[\sum_{i=1}^{K} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j)\mu_i p_{ij} \right].$$

It suffices to verify that (5.6.14) satisfies the node local balance equations

$$p(\mathbf{n})\mu_j = \sum_{i=1}^{K} p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j)\mu_i p_{ij} \tag{5.6.15}$$

for each j. To do so, note that the solution (5.6.14) has the property

$$p(\mathbf{n}+\mathbf{e}_i-\mathbf{e}_j) = \left(\frac{\pi_i}{\mu_i}\right)\left(\frac{\pi_j}{\mu_j}\right)^{-1} p(\mathbf{n}). \tag{5.6.16}$$

Hence, after substitution of (5.6.16) in (5.6.15), it suffices to verify that

$$\mu_j = \sum_{i=1}^{K} \left(\frac{\pi_i}{\mu_i}\right)\left(\frac{\pi_j}{\mu_j}\right)^{-1} \mu_i p_{ij}, \quad j = 1, \ldots, K.$$

This relation is indeed true since it coincides with the equilibrium equation (5.6.11). This completes the proof.

A computational difficulty in applying the product-form solution (5.6.14) is the determination of the normalization constant C. Theoretically this constant can be found by summing $p(n_1, \ldots, n_K)$ over all possible states $(n_1, \ldots, n_K)$. However, the number of possible states $(n_1, \ldots, n_K)$ such that $\sum_{i=1}^{K} n_i = M$ equals $\binom{M+K-1}{M}$. This is an enormous number even for modest values of K and M. Hence a direct summation to compute the constant C is only feasible for relatively small values of K and M. There are several approaches to handle the dimensionality problem, including the Gibbs sampler from Section 3.4.3. We discuss here only the mean-value algorithm.

Mean-value analysis

The mean-value algorithm is a numerically stable method for the calculation of the average number of customers at station j, the average amount of time a customer spends at station j on each visit and the average throughput at station j. The so-called arrival theorem underlies the mean-value algorithm. To formulate this theorem, it is convenient to express explicitly the dependency of the state probability $p(n_1, \ldots, n_K)$ on the number of customers in the network. We write $p(n_1, \ldots, n_K) = p_m(n_1, \ldots, n_K)$ for the network with a fixed number of m customers. For any state $(n_1, \ldots, n_K)$ with $n_1 + \cdots + n_K = M$ the equilibrium probability $p_M(n_1, \ldots, n_K)$ can be interpreted as the long-run fraction of time

that simultaneously n_1 customers are present at station 1, n_2 customers at station $2, \ldots, n_K$ customers at station K. Define the customer-average probability

$$\pi_j(n_1, \ldots, n_K) = \text{the long-run fraction of arrivals at station } j \text{ that see } n_\ell \text{ other customers present at station } \ell \text{ for } \ell = 1, \ldots, K.$$

Note that in this definition $n_1 + \cdots + n_K = M - 1$.

Theorem 5.6.3 (arrival theorem) *For any* $(n_1, \ldots, n_K)$ *with* $\sum_{\ell=1}^{K} n_\ell = M-1$,

$$\pi_j(n_1, \ldots, n_K) = p_{M-1}(n_1, \ldots, n_K).$$

Proof By part (b) of Corollary 4.3.2,

$$\text{the long-run average number of arrivals per time unit at station } j \text{ that find } n_\ell \text{ other customers present at station } \ell \text{ for } \ell = 1, \ldots, K$$
$$= \sum_{i=1}^{K} \mu_i p_{ij} p_M(n_1, \ldots, n_i + 1, \ldots, n_K)$$

for any $(n_1, \ldots, n_K)$ with $\sum_{\ell=1}^{K} n_\ell = M - 1$. In particular,

$$\text{the long-run average number of arrivals per time unit at station } j$$
$$= \sum_{\mathbf{m} \in I_{M-1}} \sum_{i=1}^{K} \mu_i p_{ij} p_M(m_1, \ldots, m_i + 1, \ldots, m_K)$$

where $\mathbf{m} = (m_1, \ldots, m_K)$ and $I_{M-1} = \{\mathbf{m} \mid \mathbf{m} \geq 0 \text{ and } m_1 + \ldots + m_K = M - 1\}$. Thus

$$\pi_j(n_1, \ldots, n_K) = \frac{\displaystyle\sum_{i=1}^{K} \mu_i p_{ij} p_M(n_1, \ldots, n_i + 1, \ldots, n_K)}{\displaystyle\sum_{\mathbf{m} \in I_{M-1}} \sum_{i=1}^{K} \mu_i p_{ij} p_M(m_1, \ldots, m_i + 1, \ldots, m_K)}.$$

By Theorem 5.6.2,

$$p_M(m_1, \ldots, m_i + 1, \ldots, m_K) = \frac{\pi_i}{\mu_i} C \prod_{k=1}^{K} \left(\frac{\pi_k}{\mu_k}\right)^{m_k}.$$

Substituting this in the numerator and the denominator of the expression for $\pi_j(n_1, \ldots, n_K)$ and cancelling out the common term $\sum_{i=1}^{K} \pi_i p_{ij}$, we find

$$\pi_j(n_1, \ldots, n_K) = C_{M-1} \prod_{k=1}^{K} \left(\frac{\pi_k}{\mu_k}\right)^{n_k}.$$

for some constant C_{M-1}. The desired result now follows from Theorem 5.6.2.

In other words, the arrival theorem states that in steady state the customer-average probability distribution of the state seen by an arriving customer (not counting this customer) is the same as the time-average probability distribution of the state in the closed network with one customer less. A special case of the arrival theorem was encountered in the Engset model; see relation (5.2.4). The product-form solution is crucial for the arrival theorem. It is noted that the arrival theorem remains valid for the closed network with a BCMP service discipline at each station. Then the product-form solution

$$p(n_1, \ldots, n_K) = C p_1(n_1) \cdots p_K(n_K)$$

holds for appropriate probability distributions $\{p_1(n_1)\}, \ldots, \{p_K(n_K)\}$.

To calculate the average number of customers and the average sojourn times at the different stations, we take the fixed number of customers moving around in the network as parameter. For the closed network with a fixed number of m customers, define the following long-run averages:

$L_m(j) =$ the average number of customers present at station j,

$W_m(j) =$ the average sojourn time of a customer at station j on each visit,

$\lambda_m(j) =$ the average number of arrivals per time unit at station j.

Note that $\lambda_m(j)$ also gives the average throughput at station j. Also, by Little's formula, $\lambda_m(j)/\mu_j$ gives the long-run fraction of time the server at station j is busy. For a constant $\gamma_m > 0$, we have by (5.6.13) that

$$\lambda_m(j) = \gamma_m \pi_j, \quad j = 1, \ldots, K, \tag{5.6.17}$$

where the π_j are the equilibrium probabilities associated with the Markov matrix $\mathbf{P} = (p_{ij})$. By Little's formula,

$$L_m(j) = \lambda_m(j) W_m(j), \quad j = 1, \ldots, K. \tag{5.6.18}$$

Obviously, we have

$$\sum_{j=1}^{K} L_m(j) = m. \tag{5.6.19}$$

The arrival theorem implies the key relation

$$W_m(j) = \frac{1}{\mu_j}\left[1 + L_{m-1}(j)\right], \quad j = 1, \ldots, K. \tag{5.6.20}$$

To see this, note that an arriving customer at node j sees on average

$$\sum_{\substack{(n_1,\ldots,n_K):\\ n_1+\cdots+n_K=m-1}} n_j \pi_j(n_1, \ldots, n_K) = \sum_{\substack{(n_1,\ldots,n_K):\\ n_1+\cdots+n_K=m-1}} n_j p_{m-1}(n_1, \ldots, n_K) = L_{m-1}(j)$$

other customers at node j. By the memoryless property of the exponential distribution and the assumption of service in order of arrival, the relation (5.6.20) now follows. This relation enables us to calculate $L_m(j)$, $W_m(j)$ and $\lambda_m(j)$ in a recursive manner. A direct consequence of (5.6.17) to (5.6.19) is the relation

$$\gamma_m = \frac{m}{\sum_{j=1}^{K} \pi_j W_m(j)}. \tag{5.6.21}$$

Mean-value algorithm

Step 0. Calculate first the equilibrium probabilities π_j associated with the Markov matrix $\mathbf{P} = (p_{ij})$. Calculate $W_1(j) = 1/\mu_j$ for $j = 1, \ldots, K$. Let $m := 1$.
Step 1. Calculate the constant γ_m from (5.6.21). Next calculate $\lambda_m(j)$ and $L_m(j)$ for $j = 1, \ldots, K$ from (5.6.17) and (5.6.18). If $m < M$, then go to step 2.
Step 2. $m := m+1$. Calculate $W_m(j)$ for $j = 1, \ldots K$ from (5.6.20). Repeat step 1.

EXERCISES

5.1 Consider the $M/M/c/c + N$ queueing model with finite waiting room. This model is the same as the $M/M/c$ model except that there are only N waiting places for customers to await service. An arriving customer who finds all c servers busy and all N waiting places occupied is rejected. Denote by $\{p_j, 0 \le j \le N + c\}$ the equilibrium distribution of the number of customers present.

(a) Give a recursion scheme for the computation of the p_j.

(b) Verify that the limiting distribution of the delay in queue of an accepted customer is given by

$$W_q(x) = 1 - \frac{1}{p_{N+c}} \sum_{j=c}^{N+c-1} p_j \sum_{k=0}^{j-c} e^{-c\mu x} \frac{(c\mu x)^k}{k!}, \quad x \ge 0.$$

5.2 In the machine-repair queueing model there are N identical machines which are attended by c repairmen, where $N > c$. The running time of a machine is exponentially distributed with mean $1/\nu$. The running times of the machines are independent of each other. A stopped machine is attended as soon as possible by a free repairman. Each repairman can handle only one machine at a time. The service time of a machine is exponentially distributed with mean $1/\mu$.

(a) Let $p_j = \lim_{t\to\infty} P\{j$ service requests are present at time $t\}$ for $0 \le j \le N$. Give a recursion scheme to compute the p_j.

(b) Let π_j denote the long-run fraction of service requests finding j other requests present upon occurrence. Argue that $\pi_j = (N - j)p_j / \sum_{k=0}^{N}(N - k)p_k$ for $0 \le j \le N - 1$.

(c) What is the limiting distribution of the delay in queue of a service request when service is in order of arrival? What is the long-run average number of busy repairmen?

5.3 Consider the following modification of the call-centre problem dealt with in Section 5.3. If the service of a customer has not yet started, the customer becomes impatient after an exponentially distributed time with mean $1/\theta$ and then leaves the system. It is assumed that the impatience time of the customer does not depend on their position in the queue (call-centre customers cannot see each other).

(a) Give a recursive relation for the computation of the equilibrium distribution $\{p_j\}$ of the number of customers present.

(b) What is the long-run fraction of customers who are delayed? Can you explain why $\theta \sum_{j=c}^{\infty}(j-c)p_j/\lambda$ gives the long-run fraction of customers who prematurely leave the system?

5.4 An information centre provides service in a bilingual environment. Requests for service arrive by telephone. Service requests of major-language customers and minor-language customers arrive according to independent Poisson processes with respective rates λ_1 and λ_2. There are c bilingual agents to handle the service requests. Each service request finding all c agents occupied upon arrival waits in queue until a free agent becomes available. The service time of a major-language request is exponentially distributed with mean $1/\mu_1$ and that of a minor-language request has an exponential distribution with mean $1/\mu_2$. Let $p(i, i_1, i_2)$ denote the joint equilibrium probability that simultaneously i_1 agents are servicing major-language customers, i_2 agents are servicing minor-language customers and i service requests are waiting in queue. Use the equilibrium equations of an appropriately chosen continuous-time Markov chain and use generating functions to prove that for any $i_1 = 0, 1, \ldots, c$ there is a constant $\gamma(i_1)$ such that

$$p(i, i_1, c - i_1) \sim \gamma(i_1)\tau^{-i} \quad \text{as } i \to \infty,$$

with $\tau = 1 + \delta/\lambda$, where $\lambda = \lambda_1 + \lambda_2$ and δ is the unique solution of

$$\delta^2 - (c\mu_1 + c\mu_2 - \lambda)\delta + c^2\mu_1\mu_2 - c\lambda_1\mu_2 - c\lambda_2\mu_1 = 0$$

on the interval $(0, c\min(\mu_1, \mu_2))$.

5.5 Consider the following modification of Example 2.5.1. Overflow is allowed from one loo to another when there is a queue at one of the loos and there is nobody at the other loo. It is assumed that the occupation times at the loos are exponentially distributed. Formulate a continuous-time Markov chain to analyse the new situation. Assume the numerical data $\lambda_w = \lambda_m = 0.6$, $\mu_w = 1.5$ and $\mu_m = 0.75$. Solve the equilibrium equations and compare the average queue sizes for the women's loo and the men's loo with the average queue sizes in the situation of strictly separated loos.

5.6 Jobs of types 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. Each job type has its own queue. Both queues are simultaneously served, where service is only provided to the job at the head of the queue. If both queues are not empty, service is provided at unity rate at each queue. A non-empty queue for type i jobs receives service at a rate of $r_i \geq 1$ when the other queue is empty ($i = 1, 2$). The service requirement of a type i job has an exponential distribution with mean $1/\mu_i$. The service requirements of the jobs are independent of each other. It is assumed that $\rho_i = \lambda_i/\mu_i$ is less than 1 for $i = 1, 2$. Let $p(i_1, i_2)$ be the joint equilibrium probability of having i_1 jobs at queue 1 and i_2 jobs at queue 2. Set up the equilibrium equations for the probabilities $p(i_1, i_2)$. Do numerical investigations to find out whether or not $p(i_1, i_2) \sim \gamma\rho_1^{i_1}\rho_2^{i_2}$ as $i_1 \to \infty$ and $i_2 \to \infty$ for some constant γ.

5.7 Consider a production hall with two machines. Jobs arrive according to a Poisson process with rate λ. Upon arrival a job has to be assigned to one of the two machines. Each machine has ample waiting space for jobs that have to wait. Each machine can handle only one job at a time. If a job is assigned to machine i, its processing time is exponentially distributed with mean $1/\mu_i$ for $i = 1, 2$. The control rule is to assign an arriving job to the machine with the shortest queue (if both queues are equal, machine group 1 is chosen). Jockeying of the jobs is not possible. Use Markov-chain analysis to find the equilibrium probability that the delay of a job in queue is longer than a given time t_0.

5.8 Consider an irreducible continuous-time Markov chain with state space I and infinitesimal transition rates q_{ij}. Let $\{p_j, j \in I\}$ be the equilibrium distribution of the Markov chain. Assume that the Markov chain is time reversible and thus has the detailed balance property (5.1.15). Suppose that the Markov chain is truncated to the subset $A \subset I$. That is, q_{ij} is changed to 0 for all $i \in A$ and $j \notin A$. Prove that the equilibrium distribution of the truncated Markov chain is given by

$$p_i^A = \frac{p_i}{\sum_{k \in A} p_k}, \quad i \in A.$$

This important result is due to Kelly (1979).

5.9 Suppose we wish to determine the capacity of a stockyard at which containers arrive according to a Poisson process with a rate of $\lambda = 1$ per hour. A container finding a full yard upon arrival is brought elsewhere. The time that a container is stored in the yard is exponentially distributed with mean $1/\mu = 10$ hours. Determine the required capacity of the yard so that no more than 1% of the arriving containers find the yard full. How does the answer change when the time that a container is stored in the yard is uniformly distributed between 5 and 15 hours?

5.10 Long-term parkers and short-term parkers arrive at a parking place for cars according to independent Poisson processes with respective rates $\lambda_1 = 4$ and $\lambda_2 = 6$ per hour. The parking place has room for $N = 10$ cars. Each arriving car which finds all places occupied goes elsewhere. The parking time of long-term parkers is uniformly distributed between 1 and 2 hours, while the parking time of short-term parkers has a uniform distribution between 20 and 60 minutes. Calculate the probability that a car finds all parking places occupied upon arrival.

5.11 Consider the loss version of the delay model from Exercise 5.4. In the loss model each service request finding all c agents occupied upon arrival is lost and has no further influence on the system. Let $p(i_1, i_2)$ denote the long-run fraction of time that simultaneously i_1 major-language customers are in service and i_2 minor-language customers are in service. Verify from the equilibrium equations for the state probabilities $p(i_1, i_2)$ that, for some constant $C > 0$,

$$p(i_1, i_2) = C \frac{(\lambda_1/\mu_1)^{i_1}}{i_1!} \frac{(\lambda_2/\mu_2)^{i_2}}{i_2!}$$

for all i_1, i_2 with $0 \leq i_1 + i_2 \leq c$. Next conclude that the equilibrium distribution of the number of occupied agents is given by formula (5.2.1) with $\lambda = \lambda_1 + \lambda_2$ and $1/\mu = (\lambda_1/\lambda) \times (1/\mu_1) + (\lambda_2/\lambda) \times (1/\mu_2)$.

5.12 Units offered for repair arrive at a repair facility according to a Poisson process with rate λ. There are c repairmen. Each repairman can handle only one unit at a time. An offered unit finding all repairmen busy is rejected and handled elsewhere. The repair time of a unit consists of two phases. The first phase is exponentially distributed with mean $1/\mu_1$ and the second one is exponentially distributed with mean $1/\mu_2$.

(a) Let $p(i_1, i_2)$ be the equilibrium probability of having i_1 units in repair phase 1 and i_2 units in repair phase 2. Verify that, for some constant C, the probability $p(i_1, i_2) = C(\lambda/\mu_1)^{i_1}(\lambda/\mu_2)^{i_2}/(i_1!i_2!)$ for all i_1, i_2.

(b) What is the equilibrium distribution of the number of busy repairmen?

(c) What is the long-run fraction of offered units that are rejected? Does this loss probability increase when the two repair phases are more variable than the exponential phases but have the same means as the exponential phases?

5.13 Consider a continuous-review inventory system in which customers asking for a certain item arrive according to a Poisson process with rate λ. Each customer asks for one unit of the item. Customer demands occurring when the system is out of stock are lost. The $(S-1, S)$ control rule is used. Under this control rule the base stock is S and a replenishment for

exactly one unit is placed each time the on-hand inventory decreases by one unit. The lead times of the replenishments are independent and identically distributed random variables with mean τ. Establish an equivalence with the Erlang loss model and give expressions for the long-run average on-hand inventory and the long-run fraction of demand that is lost.

5.14 In an electronic system there are c elements of a crucial component connected in parallel to increase the reliability of the system. Each component is switched on and the lifetimes of the components have an exponential distribution with mean $1/\alpha$. The lifetimes of the components are independent of each other. The electronic system is working as long as at least one of the components is functioning, otherwise the system is down. A component that fails is replaced by a new one. It takes an exponentially distributed time with mean $1/\beta$ to replace a failed component. Only one failed component can be replaced at a time.

(a) Use a continuous-time Markov chain to calculate the long-run fraction of time the system is down. Specify the transition rate diagram first.

(b) Does the answer in (a) change when the replacement time of a failed component has a general probability distribution with mean $1/\alpha$? (*Hint*: compare the transition rate diagram with the transition rate diagram in the Erlang loss model.)

5.15 Reconsider Exercise 5.14 but this time assume there are ample repairmen to replace failed components.

(a) Use a continuous-time Markov chain to calculate the long-run fraction of time the system is down. Specify the transition rate diagram first.

(b) What happens to the answer in (a) when the replacement time is fixed rather than exponentially distributed? (*Hint*: compare the transition rate diagram with the transition rate diagram in the Engset loss model.)

5.16 Suppose you have two groups of servers each without waiting room. The first group consists of c_1 identical servers each having an exponential service rate μ_1 and the second group consists of c_2 identical servers each having an exponential service rate μ_2. Customers for group i arrive according to a Poisson process with rate λ_i $(i = 1, 2)$. A customer who finds all servers in his group busy upon arrival is served by a server in the other group, provided one is free, otherwise the customer is lost. Show how to calculate the long-run fraction of customers lost.

5.17 Consider a conveyor system at which items for processing arrive according to a Poisson process with rate λ. The service requirements of the items are independent random variables having a common exponential distribution with mean $1/\mu$. The conveyor system has two work stations 1 and 2 that are placed according to this order along the conveyor. Workstation i consists of s_i identical service channels, each having a constant processing rate of σ_i $(i = 1, 2)$; that is, an item processed at workstation i has an average processing time of $1/(\sigma_i \mu)$. Both workstations have no storage capacity and each service channel can handle only one item at a time. An arriving item is processed by the first workstation in which a service channel is free and is lost when no service channel is available at either of the stations. Show how to calculate the fraction of items lost and solve for the numerical data $\lambda = 10$, $\mu = 1$, $\sigma_1 = 2$, $\sigma_2 = 1.5$, $s_1 = 5$ and $s_2 = 5$ (*Answer*: 0.0306). Verify experimentally that the loss probability is nearly insensitive to the distributional form of the service requirement (e.g. compute the loss probability 0.0316 for the data when the service requirement has an H_2 distribution with balanced means and a squared coefficient of variation of 4).

5.18 Consider a stochastic service system with Poisson arrivals at rate λ and two different groups of servers, where each arriving customer simultaneously requires a server from both groups. An arrival not finding that both groups have a free server is lost and has no further influence on the system. The ith group consists of s_i identical servers $(i = 1, 2)$ and each server can handle only one customer at a time. An entering customer occupies the two assigned servers from the groups 1 and 2 during independently exponentially distributed times with respective means $1/\mu_1$ and $1/\mu_2$. Show how to calculate the loss probability

and solve for the numerical data $\lambda = 1$, $1/\mu_1 = 2$, $1/\mu_2 = 5$, $s_1 = 5$ and $s_2 = 10$. (*Answer*: 0.0464.) Verify experimentally that the loss probability is nearly insensitive to the distributional form of the service times (e.g. compute the loss probability 0.0470 for the above data when the service time in group 1 has an E_2 distribution and the service time in group 2 has an H_2 distribution with balanced means and a squared coefficient of variation of 4).

5.19 Customers of the types $1, \ldots, m$ arrive at a service centre according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_m$. The service centre has c identical servers. An arriving customer of type j requires b_j servers and is lost when there are no b_j servers available. A customer of type j has an exponentially distributed service time with mean $1/\mu_j$ for $j = 1, \ldots, m$. The customer keeps all of the assigned b_j servers busy during his service time and upon completion of the service time the b_j servers are *simultaneously* released. Let $p(n_1, \ldots, n_m)$ be the long-run fraction of time that n_j groups of b_j servers are handling type j customers for $j = 1, \ldots, m$.

(a) Verify from the equilibrium equations for the probabilities $p(n_1, \ldots, n_m)$ that, for some constant $C > 0$,

$$p(n_1, \ldots, n_m) = C \prod_{j=1}^{m} \frac{(\lambda_j/\mu_j)^{n_j}}{n_j!}$$

for all $(n_1, \ldots, n_m)$ with $n_1 b_1 + \cdots + n_m b_m \leq c$.

(b) What is the long-run fraction of type j customers who are lost?

The above product-form solution can also be proved by considering the process $\{(X_1(t), \ldots, X_m(t))\}$ in the infinite-server model ($c = \infty$) with $X_j(t)$ denoting the number of type j customers present at time t. The processes $\{X_1(t)\}, \ldots, \{X_m(t)\}$ are independent of each other and each separate process $\{X_j(t)\}$ constitutes an $M/M/\infty$ queueing process having a Poisson distribution with mean λ_j/μ_j as equilibrium distribution. Noting that the process $\{(X_1(t), \ldots, X_m(t))\}$ is time reversible, it can be concluded from the result in Exercise 5.8 that the above product-form solution holds. The normalization constant C can be computed as follows. Let $\{p_j,\ 0 \leq j \leq c\}$ denote the equilibrium distribution of the numbers of busy servers in the loss model with c servers and let $\{p_j^{(\infty)}\}$ denote the equilibrium distribution of the number of busy servers in the infinite-server model. Then

$$p_j = \frac{p_j^{(\infty)}}{\sum_{k=0}^{c} p_k^{(\infty)}}, \quad j = 0, 1, \ldots, c.$$

The normalization constant C is given by p_0. It is left to the reader to verify that $\{p_j^{(\infty)}\}$ can be computed as the convolution of m compound Poisson distributions. The jth compound Poisson distribution represents the limiting distribution of the numbers of busy servers in a batch arrival $M^X/G/\infty$ queue with group service, where the arrival rate of batches is λ_j, each batch consists of b_j customers and the mean service time of the customers from the same batch is $1/\mu_j$; see part (b) of Exercise 1.15. Finally, it is noted that the loss model has the insensitivity property.

5.20 Batches of containers arrive at a stockyard according to a Poisson process with a rate of $\lambda = 15$ batches per day. Each batch consists of two or three containers with respective probabilities of $\frac{2}{3}$ and $\frac{1}{3}$. The stockyard has space for only 50 containers. An arriving batch finding not enough space is lost and is brought elsewhere. Containers from the same batch are removed simultaneously after a random time. The holding times of the batches are independent random variables and have a lognormal distribution with a mean of 1 day and a standard deviation of 2 days for batches of size 3 and a mean of 1 day and a standard deviation of $\frac{1}{2}$ day for batches of size 3. Calculate the long-run fraction of batches of size 2 that are lost and the long-run fraction of batches of size 3 that are lost.

5.21 Consider the following modification of Example 5.4.2. Instead of infinite-source input, there is finite-source input for each of the two message types. The source of messages of type j has M_j users, where each user generates a new message after an exponentially distributed think time with mean $1/\lambda_j$ provided the user has no message in service at the communication system. Assume the numerical data $c = 10$, $M_1 = M_2 = 10$, $\lambda_1 = 3$, $\lambda_2 = 1$, $\mu_1 = 4$, $\mu_2 = 1$. Use continuous-time Markov chain analysis to compute the L-policy for which the average throughput is maximal. Does the result change when the transmission times are constant rather than exponentially distributed?

5.22 Suppose a production facility has M operating machines and a buffer of B standby machines. Machines in operation are subject to breakdowns. The running times of the operating machines are independent of each other and have a common exponential distribution with mean $1/\lambda$. An operating machine that breaks down is replaced by a standby machine if one is available. A failed machine immediately enters repair. There are ample repair facilities so that any number of machines can be repaired simultaneously. The repair time of a failed machine is assumed to have an exponential distribution with mean $1/\mu$. For given values of μ, λ and M, demonstrate how to calculate the minimum buffer size B in order to achieve that the long-run fraction of time that less than M machines are operating is no more than a specific value β. Do you expect the answer to depend on the specific form of the repair-time distribution?

5.23 Suppose a communication system has c transmission channels at which messages arrive according to a Poisson process with rate λ. Each message that finds all of the c channels busy is lost upon arrival, otherwise the message is randomly assigned to one of the free channels. The transmission length of an accepted message has an exponential distribution with mean $1/\mu$. However, each separate channel is subject to a randomly changing environment that influences the transmission rate of the channel. Independently of each other, the channels alternate between periods of good condition and periods of bad condition. These alternating periods are independent of each other and have exponential distributions with means $1/\gamma_g$ and $1/\gamma_b$. The transmission rate of a channel being in good (bad) condition is σ_g (σ_b). Set up the balance equations for calculating the fraction of messages that are lost. Noting that $\sigma = (\sigma_b\gamma_g + \sigma_g\gamma_b)/(\gamma_g + \gamma_b)$ is the average transmission rate used by a channel, make some numerical comparisons with the case of a fixed transmission rate σ.

5.24 Jobs have to undergo tooling at two stations, 1 and 2, which are linked in series. New jobs arrive at station 1 according to a Poisson process with rate λ. At station 1 they undergo their first tooling. Upon completion of the tooling at station 2, there is a given probability p that both toolings have to be done anew. In this case the job rejoins the queue at station 1, otherwise the job leaves the system. The handling times of a job at stations 1 and 2 are independent random variables having exponential distributions with respective means $1/\mu_1$ and $1/\mu_2$. Each station can handle only one job at a time. What is the long-run average amount of time spent in the system by a newly arriving job?

5.25 Consider a closed queueing network as in Section 5.6.2. Assume now that the service rate at station i is a function $\mu_i(n_i)$ of the number (n_i) of customers present at station i. Verify that the product-form solution is given by

$$p(n_1, \ldots n_K) = C \prod_{i=1}^{K} \left[\lambda_i^{n_i} / \prod_{l=1}^{n_i} \mu_i(l) \right].$$

5.26 Consider the $M/G/1$ queue with Erlangian services from Example 5.5.1. Define the generating functions $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ and $F(z) = \sum_{j=0}^{\infty} f_j z^j$. Let R be the convergence radius of the series $\sum_{j=1}^{\infty} \beta_j z^j$. It is assumed that $R > 1$.

(a) Verify that

$$F(z) = \frac{\mu f_0(1-z)}{\mu(1-z) - \lambda z(1-\beta(z))}.$$

(b) Use Theorem C.1 in Appendix C to prove that $f_j \sim \gamma \eta^j$ as $j \to \infty$ for some constant γ, where η is the reciprocal of the smallest root of $\mu(1-x) - \lambda x(1-\beta(x)) = 0$ on $(1, R)$.

(c) Verify that $1 - W_q(x) \sim \gamma_1 e^{-\mu(1-\eta)x}$ as $x \to \infty$ for some constant $\gamma_1 > 0$.

5.27 Consider the so-called $MAP/G/1$ queue with a Markov modulated Poisson arrival process (an important application of this model in teletraffic analysis is the buffering of independent on-off sources at a statistical multiplexer). The arrival rate of customers is governed by an exogenous phase process. The phase process is a continuous-time Markov chain with finitely many states $s = 0, 1, \ldots, m$ and infinitesimal transition rates α_{st}. It is assumed that the phase process is irreducible and thus has a unique equilibrium distribution which is denoted by $\{e_s\}$. If the phase process is in state s, customers arrive according to a Poisson process with rate λ_s. The service times of the customers are independent random variables which are also independent of the arrival process. Customers are served in order of arrival. It is assumed that the service time of a customer has the same probability distribution function (5.5.3) as in Example 5.5.1. Letting $\rho = \left(\sum_{s=0}^m \lambda_s e_s\right) \times \left(\mu^{-1} \sum_{j=1}^\infty j\beta_j\right)$, it is assumed that the server utilization ρ is less than 1. Also it is assumed that the convergence radius R of the power series $\beta(z) = \sum_{j=1}^\infty \beta_j z^j$ is larger than 1.

(a) Let $p(i, s)$ denote the joint equilibrium probability that i customers are present and the arrival process is in phase s. Verify that for any s there is a constant γ_s such that

$$p(i, s) \sim \gamma_s \eta^i \quad \text{as } i \to \infty,$$

where η is the reciprocal of the smallest root τ of $\det \mathbf{A}(x) = 0$ on $(1, R)$. Here the $(m+1) \times (m+1)$ matrix $\mathbf{A}(z)$ is given by

$$\mathbf{A}(z) = \mu(1-z)\mathbf{I} - z(1-\beta(z))\Lambda + z\mathbf{Q}^T,$$

where Λ is the diagonal matrix $\Lambda = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_m)$ and $\mathbf{Q}^T$ is the transpose of the transition matrix $\mathbf{Q} = (q_{st})$, $s, t = 0, 1, \ldots, m$ with $q_{st} = \alpha_{st}$ for $t \neq s$ and $q_{ss} = -\sum_{t \neq s} \alpha_{st}$. For the special case of $m = 1$ with $\alpha_{01} = \omega_1$ and $\alpha_{10} = \omega_2$ (switched Poisson process), verify that the determination of τ reduces to finding the smallest root of

$$[(\lambda_1 + \mu + \omega_1)z - \lambda_1 z\beta(z) - \mu][(\lambda_2 + \mu + \omega_2)z - \lambda_2 z\beta(z) - \mu] - \omega_1\omega_2 z^2 = 0$$

on the interval $(1, R)$. Conclude that the geometric tail approach can be applied to calculate the state probabilities $p(i, s)$.

(b) Let π_j denote the long-run fraction of customers who find j other customers present upon arrival. Argue that $\pi_j = \sum_{s=0}^m \lambda_s p(j, s) / \sum_{s=0}^m \lambda_s e_s$.

(c) Let $W_q(x)$ denote the limiting probability distribution function of the delay in queue of a customer. Verify that $1 - W_q(x) \sim \gamma e^{-\mu(1-\eta)x}$ as $x \to \infty$ for some constant γ.

BIBLIOGRAPHIC NOTES

Queueing problems have laid the foundation for the continuous-time Markov chain model. The Erlang delay model and the Erlang loss model stem from teletraffic analysis. The square-root rule is discussed in many papers and was obtained by A.K. Erlang in an unpublished paper in 1924. Recommended references are Borst

et al. (2003), Halfin and Whitt (1981), Jennings *et al*. (1996) and Whitt (1992). Influential papers showing Poisson departures for the $M/M/c$ queue are Burke (1956) and Reich (1957). Insensitivity is a fundamental concept in stochastic service systems with no queueing. The illustrative problem from Example 5.4.2 is adapted from Foschini *et al*. (1981). A general discussion of the insensitivity phenomenon in stochastic networks can be found in Kelly (1979, 1991) and Van Dijk (1993). The book of Kelly (1979) makes extensive use of the concept of time-reversible Markov chains. The method of phases using fictitious stages with exponentially distributed lifetimes has its origin in the pioneering work of Erlang on stochastic processes in the early 1900s. The scope of this method was considerably enlarged by Schassberger (1973), who showed that the probability distribution of any non-negative random variable can be represented as the limit of a sequence of mixtures of Erlangian distributions with the same scale parameters. This result is very useful for both analytical and computational purposes. The product-form solution was first obtained in the paper of R.R.P. Jackson (1954) for a tandem queue consisting of two single-server stations. This work was considerably extended by J.R. Jackson (1957, 1963) to produce what have come to be known as Jackson networks. More material on queueing networks and their applications in computer and communication networks can be found in the books of Hayes (1984) and Kleinrock (1976).

REFERENCES

Baskett, F., Chandy, K.M., Muntz, R.R. and Palacios, F.G. (1975) Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.*, **22**, 248–260.

Borst, S., Mandelbaum, A. and Reiman, M.I. (2003) Dimensioning large call centers. *Operat. Res.*, **51**, 000–000.

Boucherie, R.J. (1992) *Product-Form in Queueing Networks*. Tinbergen Institute, Amsterdam.

Burke, P.J. (1956) The output of queueing systems. *Operat. Res.*, **4**, 699–704.

Cohen, J.W. (1976) *On Regenerative Processes in Queueing Theory*. Springer-Verlag, Berlin.

Cohen, J.W. (1979) The multiple phase service network with generalized processor sharing. *Acta Informatica*, **12**, 245–289.

Engset, T. (1918) Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wähleranzahl in automatischen Fernsprechämtern. *Elektrotechn. Zeitschrift*, **31**, 304–306.

Erlang, A.K. (1917) Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electr. Engin. J.*, **10**, 189–197. Reprinted in E. Brockmeyer, H.L. Halstrøm and A. Jensen, *The Life and Works of A.K. Erlang*, 2nd ed, Acta Polytechnica Scandinavica, Copenhagen.

Foschini, G.J., Gopinath, B. and Hayes, J.F. (1981) Optimum allocation of servers to two types of competing customers. *IEEE Trans. Commun.*, **29**, 1051–1055.

Halfin, S. and Whitt, W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operat. Res.*, **29**, 567–588.

Hayes, F.J. (1984) *Modelling and Analysis of Computer Communication Networks*. Plenum Press, New York.

Hordijk, A. and Schassberger, R. (1982) Weak convergence of generalized semi-Markov processes. *Stoch. Proc. Appl.*, **12**, 271–291.

Jackson, J.R. (1957) Networks of waiting times. *Operat. Res.*, **5**, 518–521.

Jackson, J.R. (1963) Jobshop-like queueing systems. *Management Sci.*, **10**, 131–142.

Jackson, R.R.P. (1954) Queueing systems with phase-type services. *Operat. Res. Quart.*, **5**, 109–120.

Jennings, O.B., Mandelbaum, A., Massey, W.A. and Whitt, W. (1996) Server staffing to meet time-varying demand. *Management Sci.*, **10**, 1383–1394.

Kelly, F.P. (1979) *Reversibility and Stochastic Networks*. John Wiley & Sons, Inc., New York.

Kelly, F.P. (1991) Loss networks. *Ann. Appl. Prob.*, **1**, 319–378.

Kleinrock, L. (1976) *Queueing Systems,* Vol II*, Computer Applications*. John Wiley & Sons, Inc., New York.

Reich, E. (1957) Waiting-times when queues are in tandem. *Ann. Math. Statist.*, **28**, 768–773.

Ross, K.W. (1995) *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, Berlin.

Schassberger, R. (1973) *Warteschlangen* (in German). Springer-Verlag, Berlin.

Schassberger, R. (1986) Two remarks on insensitive stochastic models. *Adv. Appl. Prob.*, **18**, 791–814.

Takács, L. (1969) On Erlang's formula. *Ann. Math. Statist.*, **40**, 71–78.

Van Dijk, N.M. (1993) *Queueing Networks and Product Form*. John Wiley & Sons, Ltd, Chichester.

Van Dijk, N.M. and Tijms, H.C. (1986) Insensitivity in two-node blocking network models with applications. *Teletraffic Analysis and Computer Performance Evaluation*, edited by O.J. Boxma, J.W. Cohen and H.C. Tijms, pp. 329–340. North-Holland, Amsterdam.

Whitt, W. (1980) Continuity of generalized semi-Markov processes. *Math. Operat. Res.*, **5**, 494–501.

Whitt, W. (1992) Understanding the efficiency of multi-server service systems. *Management Sci.*, **38**, 708–723.

Whittle, P. (1985) Partial balance and insensitivity. *J. Appl. Prob.*, **22**, 168–176.

CHAPTER 6

Discrete-Time Markov Decision Processes

6.0 INTRODUCTION

In the previous chapters we saw that in the analysis of many operational systems the concepts of a state of a system and a state transition are of basic importance. For dynamic systems with a *given* probabilistic law of motion, the simple Markov model is often appropriate. However, in many situations with uncertainty and dynamism, the state transitions can be controlled by taking a sequence of actions. The Markov decision model is a versatile and powerful tool for analysing probabilistic sequential decision processes with an infinite planning horizon. This model is an outgrowth of the Markov model and dynamic programming. The latter concept, being developed by Bellman in the early 1950s, is a computational approach for analysing sequential decision processes with a finite planning horizon. The basic ideas of dynamic programming are states, the principle of optimality and functional equations.

In fact dynamic programming is a recursion procedure for calculating optimal value functions from a functional equation. This functional equation reflects the principle of optimality, stating that an optimal policy has the property that whatever the initial state and the initial decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first transition. This principle is always valid when the number of states and the number of actions are finite. At much the same time as Bellman (1957) popularized dynamic programming, Howard (1960) used basic principles from Markov chain theory and dynamic programming to develop a policy-iteration algorithm for solving probabilistic sequential decision processes with an infinite planning horizon. In the two decades following the pioneering work of Bellman and Howard, the theory of Markov decision processes has expanded at a fast rate and a powerful technology has developed. However, in that period relatively little effort was put into applying the quite useful Markov decision model to practical problems.

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

The Markov decision model has many potential applications in inventory control, maintenance, manufacturing and telecommunication among others. Perhaps this versatile model will see many more significant applications when it becomes more familiar to engineers, operations research analysts, computer science people and others. To that end, Chapters 6 and 7 focus on the algorithmic aspects of Markov decision theory and illustrate the wide applicability of the Markov decision model to a variety of realistic problems. The presentation is confined to the optimality criterion of the *long-run average cost* (reward) per time unit. For many applications of Markov decision theory this criterion is the most appropriate optimality criterion. The average cost criterion is particularly appropriate when many state transitions occur in a relatively short time, as is typically the case for stochastic control problems in computer systems and telecommunication networks. Other criteria are the expected total cost and the expected total discounted cost. These criteria are discussed in length in Puterman (1994) and will not be addressed in this book.

This chapter deals with the discrete-time Markov decision model in which decisions can be made only at fixed equidistant points in time. The semi-Markov decision model in which the times between the decision epochs are random will be the subject of the next chapter. In Section 6.1 we present the basic elements of the discrete-time Markov decision model. A policy-improvement procedure is discussed in Section 6.2. This procedure is the key to various algorithms for computing an average cost optimal polity. The so-called relative values of a given policy play an important role in the improvement procedure. The relative values and their interpretation are the subject of Section 6.3. In Section 6.4 we present the policy-iteration algorithm which generates a sequence of improved policies. Section 6.5 discusses the linear programming formulation for the Markov decision model, including a formulation to handle probabilistic constraints on the state-action frequencies. The policy-iteration algorithm and the linear programming formulation both require the solving of a system of linear equations in each iteration step. In Section 6.6 we discuss the alternative method of value iteration which avoids the computationally burdensome solving of systems of linear equations but involves only recursive computations. The value-iteration algorithm endowed with quickly converging lower and upper bounds on the minimal average cost is usually the most effective method for solving Markov decision problems with a large number of states. Section 6.7 gives convergence proofs for the policy-iteration algorithm and the value-iteration algorithm.

6.1 THE MODEL

In Chapter 3 we have considered a dynamic system that evolves over time according to a *fixed* probabilistic law of motion satisfying the Markovian assumption. This assumption states that the next state to be visited depends only on the present state of the system. In this chapter we deal with a dynamic system evolving over time where the probabilistic law of motion can be controlled by taking decisions. Also, costs are incurred (or rewards are earned) as a consequence of the decisions

that are sequentially made when the system evolves over time. An *infinite planning horizon* is assumed and the goal is to find a control rule which minimizes the *long-run average cost per time unit.*

A typical example of a controlled dynamic system is an inventory system with stochastic demands where the inventory position is periodically reviewed. The decisions taken at the review times consist of ordering a certain amount of the product depending on the inventory position. The economic consequences of the decisions are reflected in ordering, inventory and shortage costs.

We now introduce the Markov decision model. Consider a dynamic system which is reviewed at equidistant points of time $t = 0, 1, \ldots$. At each review the system is classified into one of a possible number of states and subsequently a decision has to be made. The set of possible states is denoted by I. For each state $i \in I$, a set $A(i)$ of decisions or actions is given. The state space I and the action sets $A(i)$ are assumed to be *finite*. The economic consequences of the decisions taken at the review times (decision epochs) are reflected in costs. This controlled dynamic system is called a *discrete-time Markov model* when the following Markovian property is satisfied. If at a decision epoch the action a is chosen in state i, then regardless of the past history of the system, the following happens:

(a) an immediate cost $c_i(a)$ is incurred,

(b) at the next decision epoch the system will be in state j with probability $p_{ij}(a)$, where

$$\sum_{j \in I} p_{ij}(a) = 1, \quad i \in I.$$

Note that the one-step costs $c_i(a)$ and the one-step transition probabilities $p_{ij}(a)$ are assumed to be time homogeneous. In specific problems the 'immediate' costs $c_i(a)$ will often represent the expected cost incurred until the next decision epoch when action a is chosen in state i. Also, it should be emphasized that the choice of the state space and of the action sets often depends on the cost structure of the specific problem considered. For example, in a production/inventory problem involving a fixed set-up cost for restarting production after an idle period, the state description should include a state variable indicating whether the production facility is on or off. Many practical control problems can be modelled as a Markov decision process by an appropriate choice of the state space and action sets. Before we develop the required theory for the average cost criterion, we give a typical example of a Markov decision problem.

Example 6.1.1 A maintenance problem

At the beginning of each day a piece of equipment is inspected to reveal its actual working condition. The equipment will be found in one of the working conditions $i = 1, \ldots, N$, where the working condition i is better than the working condition

$i + 1$. The equipment deteriorates in time. If the present working condition is i and no repair is done, then at the beginning of the next day the equipment has working condition j with probability q_{ij}. It is assumed that $q_{ij} = 0$ for $j < i$ and $\sum_{j \geq i} q_{ij} = 1$. The working condition $i = N$ represents a malfunction that requires an enforced repair taking two days. For the intermediate states i with $1 < i < N$ there is a choice between preventively repairing the equipment and letting the equipment operate for the present day. A preventive repair takes only one day. A repaired system has the working condition $i = 1$. The cost of an enforced repair upon failure is C_f and the cost of a pre-emptive repair in working condition i is C_{pi}. We wish to determine a maintenance rule which minimizes the long-run average repair cost per day.

This problem can be put in the framework of a discrete-time Markov decision model. Also, since an enforced repair takes two days and the state of the system has to be defined at the beginning of each day, we need an auxiliary state for the situation in which an enforced repair is in progress already for one day. Thus the set of possible states of the system is chosen as

$$I = \{1, 2, \ldots, N, N + 1\}.$$

State i with $1 \leq i \leq N$ corresponds to the situation in which an inspection reveals working condition i, while state $N + 1$ corresponds to the situation in which an enforced repair is in progress already for one day. Define the actions

$$a = \begin{cases} 0 & \text{if no repair is done,} \\ 1 & \text{if a preventive repair is done,} \\ 2 & \text{if an enforced repair is done.} \end{cases}$$

The set of possible actions in state i is chosen as

$$A(1) = \{0\}, \quad A(i) = \{0, 1\} \text{ for } 1 < i < N, \quad A(N) = A(N + 1) = \{2\}.$$

The one-step transition probabilities $p_{ij}(a)$ are given by

$$p_{ij}(0) = q_{ij} \quad \text{for } 1 \leq i < N,$$

$$p_{i1}(1) = 1 \quad \text{for } 1 < i < N,$$

$$p_{N,N+1}(2) = p_{N+1,1}(2) = 1,$$

and the other $p_{ij}(a) = 0$. The one-step costs $c_i(a)$ are given by

$$c_i(0) = 0, \quad c_i(1) = C_{pi}, \quad c_N(2) = C_f \quad \text{and} \quad c_{N+1}(2) = 0.$$

Stationary policies

We now introduce some concepts that will be needed in the algorithms to be described in the next sections. *A rule or policy* for controlling the system is a prescription for taking actions at each decision epoch. In principle a control rule

may be quite complicated in the sense that the prescribed actions may depend on the whole history of the system. An important class of policies is the subclass of stationary policies. A *stationary policy* R is a policy that assigns to each state i a fixed action $a = R_i$ and always uses this action whenever the system is in state i. For example, in the maintenance problem with $N = 5$, the policy R prescribing a preventive repair only in the states 3 and 4 is given by $R_1 = 0$, $R_2 = 0$, $R_3 = R_4 = 1$ and $R_5 = R_6 = 2$.

For $n = 0, 1, \ldots$, define

$$X_n = \text{the state of the system at the } n\text{th decision epoch.}$$

Under a given stationary policy R, we have

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

regardless of the past history of the system up to time n. Hence under a given stationary policy R the stochastic process $\{X_n\}$ is a discrete-time Markov chain with one-step transition probabilities $p_{ij}(R_i)$. This Markov chain incurs a cost $c_i(R_i)$ each time the system visits state i. Thus we can invoke results from Markov chain theory to specify the long-run average cost per time unit under a given stationary policy.

In view of the Markov assumption made and the fact that the planning horizon is infinitely long, it will be intuitively clear that it is sufficient to consider only the class of stationary policies. However, other policies are conceivable: policies whose actions depend on the past states or policies whose actions are determined by a random mechanism. This issue raises a fundamental question in Markov decision theory: does there exist an optimal policy among the class of all conceivable policies and, if an optimal policy exists, is such a policy a stationary policy? The answer to these questions is yes for the average-cost Markov decision model with a *finite* state space and *finite* action sets. However, a mathematical proof requires rather deep arguments. The interested reader is referred to Derman (1970) and Puterman (1994) for a proof. From these books the reader will learn that the issue of the existence of an optimal (stationary) policy is a very subtle one. Especially for the average cost criterion, the optimality questions become very complicated when the state space is not finite but countably infinite. Even in simple countable-state models, average cost optimal policies need not exist and, when they do, they need not be stationary; see Puterman (1994). In the average-cost Markov decision model with a finite state space and finite action sets these difficulties do not arise and the analysis can be restricted to the class of stationary policies.

6.2 THE POLICY-IMPROVEMENT IDEA

In this section we will establish a key result that underlies the various algorithms for the computation of an average cost optimal policy. Before doing this, we discuss the long-run average cost per time unit for a stationary policy.

Average cost for a given stationary policy

Fix a stationary policy R. Under policy R each time the action $a = R_i$ is taken whenever the system is in state i at a decision epoch. The process $\{X_n\}$ describing the state of the system at the decision epochs is a Markov chain with one-step transition probabilities $p_{ij}(R_i)$, $i, j \in I$ when policy R is used. Denote the n-step transition probabilities of this Markov chain by

$$p_{ij}^{(n)}(R) = P\{X_n = j \mid X_0 = i\}, \quad i, j \in I \text{ and } n = 1, 2, \ldots .$$

Note that $p_{ij}^{(1)}(R) = p_{ij}(R_i)$. By the equations (3.2.1),

$$p_{ij}^{(n)}(R) = \sum_{k \in I} p_{ik}^{(n-1)}(R) p_{kj}(R_k), \quad n = 1, 2, \ldots , \tag{6.2.1}$$

where $p_{ij}^{(0)}(R) = 1$ for $j = i$ and $p_{ij}^{(0)}(R) = 0$ for $j \neq i$. Also, define the expected cost function $V_n(i, R)$ by

$$V_n(i, R) = \text{the total expected costs over the first } n \text{ decision epochs when the initial state is } i \text{ and policy } R \text{ is used.}$$

Obviously, we have

$$V_n(i, R) = \sum_{t=0}^{n-1} \sum_{j \in I} p_{ij}^{(t)}(R) c_j(R_j), \tag{6.2.2}$$

Next we define the average cost function $g_i(R)$ by

$$g_i(R) = \lim_{n \to \infty} \frac{1}{n} V_n(i, R), \quad i \in I. \tag{6.2.3}$$

This limit exists by Theorem 3.3.1 and represents the long-run average *expected* cost per time unit when the system is controlled by policy R and the initial state is i. A state i is said to be *recurrent* under policy R if the system ultimately returns to state i with probability 1 when the system starts in state i and policy R is used; see Section 3.2.3. Otherwise, state i is said to be *transient* under policy R. If state i is recurrent under policy R, then $g_i(R)$ allows for the stronger interpretation

$$\text{the long-run } \textit{actual} \text{ average cost per time unit} = g_i(R) \tag{6.2.4}$$

with probability 1 when the initial state is i and policy R is used. This is a direct consequence of the theory for finite-state Markov chains. For the Markov chain $\{X_n\}$ corresponding to policy R, the state space can be uniquely split up into a finite number of disjoint irreducible sets of recurrent states and a (possibly empty) set of transient states; see Section 3.5.1. Denote the recurrent subclasses by $I_1(R), \ldots , I_f(R)$ and the set of transient states by $T(R)$. Since the system cannot

leave a closed set, the process $\{X_n\}$ restricted to any recurrent subclass $I_\ell(R)$ is a Markov chain itself with its own equilibrium distribution. Since the restricted Markov chain on $I_\ell(R)$ is irreducible, it follows from Theorem 3.3.3 that (6.2.4) holds for $i \in I_\ell(R)$, $\ell = 1, \ldots, f$. Moreover,

$$g_i(R) = g_j(R), \quad i, j \in I_\ell(R).$$

Let $g^{(\ell)}(R)$ denote the common value of $g_i(R)$ for $i \in I_\ell(R)$. For a transient initial state i, the long-run average cost per time unit is a random variable. This random variable assumes the value $g^{(\ell)}(R)$ with probability $f_i^{(\ell)}(R)$, where $f_i^{(\ell)}(R)$ is the probability that the system is ultimately absorbed in the recurrent subclass $I_\ell(R)$ when the initial state is i and policy R is used. Obviously, $g_i(R) = \sum_{\ell=1}^{f} g^{(\ell)}(R) f_i^{(\ell)}(R)$ for $i \in T(R)$.

The above technical discussion involves rather heavy notation and might be intimidating for some readers. This discussion greatly simplifies when the Markov chain $\{X_n\}$ corresponding to policy R is unichain as is mostly the case in practical situations. The Markov chain is said to be *unichain* if it has no two disjoint closed sets. In the unichain case the Markov chain $\{X_n\}$ has a unique equilibrium distribution $\{\pi_j(R), j \in I\}$. For any $j \in I$,

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)}(R) = \pi_j(R), \tag{6.2.5}$$

independently of the initial state i. The $\pi_j(R)$ are the unique solution to the system of equilibrium equations

$$\pi_j(R) = \sum_{i\in I} p_{ij}(R_i)\pi_i(R), \quad j \in I, \tag{6.2.6}$$

in conjunction with $\sum_{j\in I} \pi_j(R) = 1$. By (6.2.2), (6.2.3) and (6.2.5),

$$g_i(R) = g(R) \quad \text{for all } i \in I$$

with

$$g(R) = \sum_{j\in I} c_j(R_j)\pi_j(R). \tag{6.2.7}$$

We defined $g_i(R)$ as an average *expected* cost. For the unichain case, it follows from renewal-reward theory that the long-run average *actual* cost per time unit equals $g(R)$ with probability 1 when policy R is used, independently of the initial state.

In practical applications the Markov chain $\{X_n\}$ associated with an optimal stationary policy will typically be unichain. The reader might wonder why we are still paying attention to the multichain case. The reason is that in some applications non-optimal policies may have multiple recurrent subclasses and those policies may

show up in intermediate steps of the algorithms for computing an optimal policy. However, in most practical applications the Markov chain $\{X_n\}$ is unichain for each stationary policy.

Policy-improvement idea

A stationary policy R^* is said to be *average cost optimal* if

$$g_i(R^*) \leq g_i(R)$$

for each stationary policy R, uniformly in the initial state i. It is stated without proof that an average cost optimal stationary policy R^* always exists. Moreover, policy R^* is not only optimal among the class of stationary policies but it is also optimal among the class of all conceivable policies.

In most applications it is computationally not feasible to find an average cost optimal policy by computing the average cost for each stationary policy separately. For example, if the number of states is N and there are two actions in each state, then the number of possible stationary policies is 2^N and this number grows quickly beyond any practical bound. However, several algorithms can be given that lead in an effective way to an average cost optimal policy. Policy iteration and value iteration are the most widely used algorithms to compute an average cost optimal policy. The first method works on the policy space and generates a sequence of improved policies, whereas the second method approximates the minimal average cost through a sequence of value functions. In both methods a key role is played by the so-called relative values. The relative values are the basis for a powerful improvement step. The improvement step is motivated through a heuristic discussion of the relative values of a given policy R. In the next section a rigorous treatment will be presented for the relative values.

Let us fix any stationary policy R. It is assumed that the Markov chain $\{X_n\}$ associated with policy R has no two disjoint closed sets. Then the average cost $g_i(R) = g(R)$, independently of the initial state $i \in I$. The starting point is the obvious relation $\lim_{n\to\infty} V_n(i, R)/n = g(R)$ for all i, where $V_n(i, R)$ denotes the total expected costs over the first n decision epochs when the initial state is i and policy R is used. This relation motivates the heuristic assumption that bias values $v_i(R)$, $i \in I$, exist such that, for each $i \in I$,

$$V_n(i, R) \approx ng(R) + v_i(R) \quad \text{for } n \text{ large.} \tag{6.2.8}$$

Note that $v_i(R) - v_j(R) \approx V_n(i, R) - V_n(j, R)$ for n large. Thus $v_i(R) - v_j(R)$ measures the difference in total expected costs when starting in state i rather than in state j, given that policy R is followed. This explains the name of *relative values* for the $v_i(R)$. We have the recursion equation

$$V_n(i, R) = c_i(R_i) + \sum_{j\in I} p_{ij}(R_i)V_{n-1}(j, R), \quad n \geq 1 \text{ and } i \in I$$

with $V_0(i, R) = 0$. This equation follows by conditioning on the next state that occurs when action $a = R_i$ is made in state i when n decision epochs are to go. A cost $c_i(R_i)$ is incurred at the first decision epoch and the total expected cost over the remaining $n - 1$ decision epochs is $V_{n-1}(j, R)$ when the next state is j. By substituting the asymptotic expansion (6.2.8) in the recursion equation, we find, after cancelling out common terms,

$$g(R) + v_i(R) \approx c_i(R_i) + \sum_{j \in I} p_{ij}(R_i) v_j(R), \quad i \in I. \tag{6.2.9}$$

The intuitive idea behind the procedure for improving the given policy R is to consider the following difference in costs:

$$\Delta(i, a, R) = \text{the difference in total expected costs over an infinitely long period of time by taking first action } a \text{ and next using policy } R \text{ rather than using policy } R \text{ from scratch when the initial state is } i.$$

This difference is equal to zero when action $a = R_i$ is chosen. We wish to make the difference $\Delta(i, a, R)$ as negative as possible. This difference is given by

$$\Delta(i, a, R) = \lim_{n \to \infty} \left[c_i(a) + \sum_{j \in I} p_{ij}(a) V_{n-1}(j, R) - \{c_i(R_i) + \sum_{j \in I} p_{ij}(R_i) V_{n-1}(j, R)\} \right].$$

Substituting (6.2.8) into the expression between brackets, we find that for large n this expression is approximately equal to

$$c_i(a) + \sum_{j \in I} p_{ij}(a) v_j(R) - (n-1) g(R) - \left\{ c_i(R_i) + \sum_{j \in I} p_{ij}(R_i) v_j(R) - (n-1) g(R) \right\}.$$

This gives

$$\Delta(i, a, R) \approx c_i(a) + \sum_{j \in I} p_{ij}(a) v_j(R) - c_i(R_i) - \sum_{j \in I} p_{ij}(R_i) v_j(R).$$

Thus, by using (6.2.9),

$$\Delta(i, a, R) \approx c_i(a) + \sum_{j \in I} p_{ij}(a) v_j(R) - g(R) - v_i(R).$$

This relation and the definition of $\Delta(i, a, R)$ suggest we should look for an action a in state i so that the quantity

$$c_i(a) - g(R) + \sum_{j \in I} p_{ij}(a) v_j(R) \tag{6.2.10}$$

is as small as possible. The quantity in (6.2.10) is called the *policy-improvement quantity*. The above heuristic discussion suggests a main theorem that will be the basis for the algorithms to be discussed later. A direct proof of this theorem can be given without using any of the heuristic assumptions made above.

Theorem 6.2.1 (improvement theorem) *Let g and v_i, $i \in I$, be given numbers. Suppose that the stationary policy $\overline{R}$ has the property*

$$c_i(\overline{R}_i) - g + \sum_{j \in I} p_{ij}(\overline{R}_i) v_j \leq v_i \quad \text{for each } i \in I. \tag{6.2.11}$$

Then the long-run average cost of policy $\overline{R}$ satisfies

$$g_i(\overline{R}) \leq g, \quad i \in I, \tag{6.2.12}$$

where the strict inequality sign holds in (6.2.12) for $i = r$ when state r is recurrent under policy $\overline{R}$ and the strict inequality sign holds in (6.2.11) for $i = r$. The result is also true when the inequality signs in (6.2.11) and (6.2.12) are reversed.

Proof We first give an intuitive explanation of the theorem and next we give a formal proof. Suppose that a control cost of $c_i(a) - g$ is incurred each time the action a is chosen in state i, while a terminal cost of v_j is incurred when the control of the system is stopped and the system is left behind in state j. Then (6.2.11) states that controlling the system for one step according to rule $\overline{R}$ and stopping next is preferable to stopping directly when the initial state is i. Since this property is true for each initial state, a repeated application of this property yields that controlling the system for m steps according to rule $\overline{R}$ and stopping after that is preferable to stopping directly. Thus, for each initial state $i \in I$,

$$V_m(i, \overline{R}) - mg + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) v_j \leq v_i, \quad m = 1, 2, \ldots .$$

Dividing both sides of this inequality by m and letting $m \to \infty$, we get (6.2.12). Next we give a formal proof that yields the result with the strict inequality sign as well. The proof is first given under the assumption that the Markov chain $\{X_n\}$ associated with policy $\overline{R}$ is unichain. Then this Markov chain has a unique equilibrium distribution $\{\pi_j(\overline{R}), j \in I\}$, where $\pi_j(\overline{R}) > 0$ only if state j is recurrent under policy $\overline{R}$. Multiply both sides of (6.2.11) by $\pi_i(\overline{R})$ and sum over i. This gives

$$\sum_{i \in I} \pi_i(\overline{R}) c_i(\overline{R}_i) - g + \sum_{i \in I} \pi_i(\overline{R}) \sum_{j \in I} p_{ij}(\overline{R}_i) v_j \leq \sum_{i \in I} \pi_i(\overline{R}) v_i,$$

where the strict inequality sign holds when the strict inequality sign holds in (6.2.11) for some i with $\pi_i(\overline{R}) > 0$. Interchanging the order of summation in the above inequality and using (6.2.6) and (6.2.7) with R replaced by $\overline{R}$, we find

$$g(\overline{R}) - g + \sum_{j \in I} \pi_j(\overline{R}) v_j \leq \sum_{i \in I} \pi_i(\overline{R}) v_i,$$

where the strict inequality sign holds when the strict inequality sign holds in (6.2.11) for some i with $\pi_i(\overline{R}) > 0$. This verifies the theorem for the case of a unichain policy $\overline{R}$. Next it is easy to establish the theorem for the case of a multichain policy $\overline{R}$. Letting $I_1(\overline{R}), \ldots, I_f(\overline{R})$ denote the recurrent subclasses of the Markov chain associated with policy $\overline{R}$, the above proof shows that for any $\ell = 1, \ldots, f$ the inequality (6.2.12) holds for all $i \in I_\ell(R)$. The proof of the theorem is next completed by noting that for each transient state i the average expected cost $g_i(R)$ is a linear combination of the average costs on the recurrent subclasses.

6.3 THE RELATIVE VALUE FUNCTION

In Section 6.2 we introduced in a heuristic way the relative values for a given stationary policy R. In this section we give a rigorous treatment. This will be done for the case of a unichain policy R. Let r be any recurrent state of policy R. In view of the unichain assumption, the Markov chain $\{X_n\}$ associated with policy R will visit state r after finitely many transitions, regardless of the initial state. Thus we can define, for each state $i \in I$,

$$T_i(R) = \text{the expected time until the first return to state } r \text{ when starting in state } i \text{ and using policy } R.$$

In particular, letting a cycle be the time elapsed between two consecutive visits to the regeneration state r under policy R, we have that $T_r(R)$ is the expected length of a cycle. Also define, for each $i \in I$,

$$K_i(R) = \text{the expected costs incurred until the first return to state } r \text{ when starting in state } i \text{ and using policy } R.$$

We use the convention that $K_i(R)$ includes the cost incurred when starting in state i but excludes the cost incurred when returning to state r. By the theory of renewal-reward processes, the average cost per time unit equals the expected costs incurred in one cycle divided by the expected length of one cycle and so

$$g(R) = \frac{K_r(R)}{T_r(R)}.$$

Next we define the particular relative value function

$$w_i(R) = K_i(R) - g(R) T_i(R), \quad i \in I. \tag{6.3.1}$$

Note, as a consequence of (6.3.1), the normalization

$$w_r(R) = 0.$$

In accordance with the heuristic result (6.2.9), the next theorem shows that the average cost $g = g(R)$ and the relative values $v_i = w_i(R)$, $i \in I$ satisfy a system of linear equations.

Theorem 6.3.1 *Let R be a given stationary policy such that the associated Markov chain $\{X_n\}$ has no two disjoint closed sets. Then*

(a) The average cost $g(R)$ and the relative values $w_i(R)$, $i \in I$, satisfy the following system of linear equations in the unknowns g and v_i, $i \in I$:

$$v_i = c_i(R_i) - g + \sum_{j \in I} p_{ij}(R_i) v_j, \quad i \in I. \tag{6.3.2}$$

(b) Let the numbers g and v_i, $i \in I$, be any solution to (6.3.2). Then

$$g = g(R)$$

and, for some constant c,

$$v_i = w_i(R) + c, \quad i \in I.$$

(c) Let s be an arbitrarily chosen state. Then the linear equations (6.3.2) *together with the normalization equation $v_s = 0$ have a unique solution.*

Proof (a) By conditioning on the next state following the initial state i, it can be seen that

$$T_i(R) = 1 + \sum_{j \neq r} p_{ij}(R_i) T_j(R), \quad i \in I,$$

$$K_i(R) = c_i(R_i) + \sum_{j \neq r} p_{ij}(R_i) K_j(R), \quad i \in I.$$

This implies that

$$K_i(R) - g(R)T_i(R) = c_i(R_i) - g(R) + \sum_{j \neq r} p_{ij}(R_i)\{K_j(R) - g(R)T_j(R)\}.$$

Hence, by $w_r(R) = 0$, we find

$$w_i(R) = c_i(R_i) - g(R) + \sum_{j \in I} p_{ij}(R_i) w_j(R), \quad i \in I.$$

(b) Let $\{g, v_i\}$ be any solution to (6.3.2). We first verify by induction that the following identity holds for each $m = 1, 2, \ldots$.

$$v_i = \sum_{t=0}^{m-1} \sum_{j \in I} p_{ij}^{(t)}(R) c_j(R_j) - mg + \sum_{j \in I} p_{ij}^{(m)}(R) v_j, \quad i \in I. \tag{6.3.3}$$

Clearly, (6.3.3) is true for $m = 1$. Suppose that (6.3.3) is true for $m = n$. Substituting equations (6.3.2) into the right-hand side of (6.3.3) with $m = n$, it follows that

$$\begin{aligned} v_i &= \sum_{t=0}^{n-1} \sum_{j \in I} p_{ij}^{(t)}(R) c_j(R_j) - ng + \sum_{j \in I} p_{ij}^{(n)}(R) \left\{ c_j(R_j) - g + \sum_{k \in I} p_{jk}(R_j) v_k \right\} \\ &= \sum_{t=0}^{n} \sum_{j \in I} p_{ij}^{(t)}(R) c_j(R_j) - (n+1)g + \sum_{k \in I} \left\{ \sum_{j \in I} p_{ij}^{(n)}(R) p_{jk}(R_j) \right\} v_k, \quad i \in I. \end{aligned}$$

where the latter equality involves an interchange of the order of summation. Next, using (6.2.1), we get (6.3.3) for $m = n + 1$, which completes the induction step.

Using the relation (6.2.2) for the total expected costs over the first m decision epochs, we can rewrite (6.3.3) in the more convenient form

$$v_i = V_m(i, R) - mg + \sum_{j \in I} p_{ij}^{(m)}(R) v_j, \quad i \in I. \tag{6.3.4}$$

Since $V_m(i, R)/m \to g(R)$ as $m \to \infty$ for each i, the result $g = g(R)$ follows by dividing both sides of (6.3.4) by m and letting $m \to \infty$. To prove the second part of assertion (b), let $\{g, v_i\}$ and $\{g', v_i'\}$ be any two solutions to (6.3.1). Since $g = g' = g(R)$, it follows from the representation (6.3.4) that

$$v_i - v_i' = \sum_{j \in I} p_{ij}^{(m)}(R)\{v_j - v_j'\}, \quad i \in I \text{ and } m \geq 1.$$

By summing both sides of this equation over $m = 1, \ldots, n$ and then dividing by n, it follows after an interchange of the order of summation that

$$v_i - v_i' = \sum_{j \in I} \left\{ \frac{1}{n} \sum_{m=1}^{n} p_{ij}^{(m)}(R) \right\} (v_j - v_j'), \quad i \in I \text{ and } n \geq 1.$$

Next, by letting $n \to \infty$ and using (6.2.5), we obtain

$$v_i - v_i' = \sum_{j \in I} \pi_j(R)(v_j - v_j'), \quad i \in I.$$

The right-hand side of this equation does not depend on i. This proves part (b).

(c) Since $\sum_j p_{ij}(R_i) = 1$ for each $i \in I$, it follows that for any constant c the numbers g and $v_i = w_i(R) + c$, $i \in I$, satisfy (6.3.2). Hence the equations (6.3.2)

together with $v_s = 0$ for some s must have a solution. In view of assertion (b), this solution is unique. This completes the proof of the theorem.

Interpretation of the relative values

The equations (6.3.2) are referred to as the *value-determination equations.* The relative value function v_i, $i \in I$ is unique up to an additive constant. The particular solution (6.3.1) can be interpreted as the total expected costs incurred until the first return to state r when policy R is used and the one-step costs are given by $c_i'(a) = c_i(a) - g$ with $g = g(R)$. If the Markov chain $\{X_n\}$ associated with policy R is aperiodic, two other interpretations can be given to the relative value function. The first interpretation is that, for any two states $i, j \in I$,

$$v_i - v_j = \text{the difference in total expected costs over an infinitely long period of time by starting in state } i \text{ rather than in state } j \text{ when using policy } R.$$

In other words, $v_i - v_j$ is the maximum amount that a rational person is willing to pay to start the system in state j rather than in state i when the system is controlled by rule R. This interpretation is an easy consequence of (6.3.3). Using the assumption that the Markov chain $\{X_n\}$ is aperiodic, we have that $\lim_{m\to\infty} p_{ij}^{(m)}(R)$ exists. Moreover this limit is independent of the initial state i, since R is unichain. Thus, by (6.3.3),

$$v_i = \lim_{m\to\infty} \{V_m(i, R) - mg\} + \sum_{j\in I} \pi_j(R) v_j. \tag{6.3.5}$$

This implies that $v_i - v_j = \lim_{m\to\infty}\{V_m(i, R) - V_m(j, R)\}$, yielding the above interpretation. A special interpretation applies to the relative value function v_i, $i \in I$ with the property $\sum_{j\in I} \pi_j(R) v_j = 0$. Since the relative value function is unique up to an additive constant, there is a unique relative value function with this property. Denote this relative value function by h_i, $i \in I$. Then, by (6.3.5),

$$h_i = \lim_{m\to\infty} \{V_m(i, R) - mg\}. \tag{6.3.6}$$

The bias h_i can also be interpreted as the difference in total expected costs between the system whose initial state is i and the system whose initial state is distributed according to the equilibrium distribution $\{\pi_j(R),\ j \in I\}$ when both systems are controlled by policy R. The latter system is called the stationary system. This system has the property that at any decision epoch the state is distributed as $\{\pi_j(R)\}$; see Section 3.3.2. Thus, for the stationary system, the expected cost incurred at any decision epoch equals $\sum_{j\in I} c_j(R_j)\pi_j(R)$ being the average cost $g = g(R)$ of policy R. Consequently, in the stationary system the total expected costs over the first m decision epochs equals mg. This gives the above interpretation of the bias h_i.

6.4 POLICY-ITERATION ALGORITHM

For ease of presentation we will discuss the policy-iteration algorithm under a unichain assumption that is satisfied in most applications.

Unichain assumption *For each stationary policy the associated Markov chain* $\{X_n\}$ *has no two disjoint closed sets.*

The relative values associated with a given policy R provide a tool for constructing a new policy $\overline{R}$ whose average cost is no more than that of the current policy R. In order to improve a given policy R whose average cost $g(R)$ and relative values $v_i(R)$, $i \in I$, have been computed, we apply Theorem 6.2.1 with $g = g(R)$ and $v_i = v_i(R)$, $i \in I$. By constructing a new policy $\overline{R}$ such that, for each state $i \in I$,

$$c_i(\overline{R}_i) - g(R) + \sum_{j \in I} p_{ij}(\overline{R}_i) v_j \leq v_i, \tag{6.4.1}$$

we obtain an improved rule $\overline{R}$ according to $g(\overline{R}) \leq g(R)$. In constructing such an improved policy $\overline{R}$ it is important to realize that for each state i *separately* an action $\overline{R}_i$ satisfying (6.4.1) can be determined. As a side remark, we point out that this flexibility of the policy-improvement procedure may be exploited in specific applications to generate a sequence of improved policies within a subclass of policies having a simple structure. A particular way to find for state $i \in I$ an action $\overline{R}_i$ satisfying (6.4.1) is to minimize

$$c_i(a) - g(R) + \sum_{j \in I} p_{ij}(a) v_j(R) \tag{6.4.2}$$

with respect to $a \in A(i)$. Noting that the expression in (6.4.2) equals $v_i(R)$ for $a = R_i$, it follows that (6.4.1) is satisfied for the action $\overline{R}_i$ which minimizes (6.4.2) with respect to $a \in A(i)$. We are now in a position to formulate the following algorithm.

Policy-iteration algorithm

Step 0 (initialization). Choose a stationary policy R.
Step 1 (value-determination step). For the current rule R, compute the unique solution $\{g(R), v_i(R)\}$ to the following system of linear equations:

$$v_i = c_i(R_i) - g + \sum_{j \in I} p_{ij}(R_i) v_j, \quad i \in I,$$

$$v_s = 0,$$

where s is an arbitrarily chosen state.

Step 2 (policy-improvement step). For each state $i \in I$, determine an action a_i yielding the minimum in

$$\min_{a \in A(i)} \left\{ c_i(a) - g(R) + \sum_{j \in I} p_{ij}(a) v_j(R) \right\}.$$

The new stationary policy $\overline{R}$ is obtained by choosing $\overline{R}_i = a_i$ for all $i \in I$ with the convention that $\overline{R}_i$ is chosen equal to the old action R_i when this action minimizes the policy-improvement quantity.
Step 3 (convergence test). If the new policy $\overline{R} = R$, then the algorithm is stopped with policy R. Otherwise, go to step 1 with R replaced by $\overline{R}$.

The policy-iteration algorithm converges after a finite number of iterations to an average cost optimal policy. We defer the proof to Section 6.7. The policy-iteration algorithm is empirically found to be a remarkably robust algorithm that converges very fast in specific problems. The number of iterations is *practically independent* of the number of states and varies typically between 3 and 15, say. Also, it can be roughly stated that the average costs of the policies generated by policy iteration converge at least exponentially fast to the minimum average cost, with the greatest improvements in the first few iterations.

Remark 6.4.1 The average cost optimality equation

Since the policy-iteration algorithm converges after finitely many iterations, there exist numbers g^* and v_i^*, $i \in I$, such that

$$v_i^* = \min_{a \in A(i)} \left\{ c_i(a) - g^* + \sum_{j \in I} p_{ij}(a) v_j^* \right\}, \quad i \in I. \tag{6.4.3}$$

This functional equation is called the *average cost optimality equation*. Using Theorem 6.2.1, we can directly verify that any stationary policy R^* for which the action R_i^* minimizes the right-hand side of (6.4.3) for all $i \in I$ is average cost optimal. To see this, note that

$$v_i^* = c_i(R_i^*) - g^* + \sum_{j \in I} p_{ij}(R_i^*) v_j^*, \quad i \in I \tag{6.4.4}$$

and

$$v_i^* \leq c_i(a) - g^* + \sum_{j \in I} p_{ij}(a) v_j^*, \quad a \in A(i) \text{ and } i \in I. \tag{6.4.5}$$

The equality (6.4.4) and Theorem 6.2.1 imply that $g(R^*) = g^*$. Let $\overline{R}$ be any stationary policy. Taking $a = \overline{R}_i$ in (6.4.5) for all $i \in I$ and applying Theorem 6.2.1, we find $g(\overline{R}) \geq g^*$. In other words, $g(R^*) \leq g(\overline{R})$ for any stationary policy $\overline{R}$. This shows not only that policy R^* is average cost optimal but also shows that the constant g^* in (6.4.3) is uniquely determined as the minimal average cost per time

unit. It is stated without proof that the function v_i^*, $i \in I$, in (6.4.3) is uniquely determined up to an additive constant.

Next the policy-iteration algorithm is applied to compute an average cost optimal policy for the control problem in Example 6.1.1.

Example 6.1.1 (continued) A maintenance problem

It is assumed that the number of possible working conditions equals $N = 5$. The repair costs are given by $C_f = 10$, $C_{p2} = 7$, $C_{p3} = 7$ and $C_{p4} = 5$. The deterioration probabilities q_{ij} are given in Table 6.4.1. The policy-iteration algorithm is initialized with the policy $R^{(1)} = (0, 0, 0, 0, 2, 2)$, which prescribes repair only in the states 5 and 6. In the calculations below, the policy-improvement quantity is abbreviated as

$$T_i(a, R) = c_i(a) - g(R) + \sum_{j \in I} p_{ij}(a) v_j(R)$$

when the current policy is R. Note that always $T_i(a, R) = v_i(R)$ for $a = R_i$.

Iteration 1

Step 1 (value determination). The average cost and the relative values of policy $R^{(1)} = (0, 0, 0, 0, 2, 2)$ are computed by solving the linear equations

$$\begin{aligned}
v_1 &= 0 - g + 0.9v_1 + 0.1v_2 \\
v_2 &= 0 - g + 0.8v_2 + 0.1v_3 + 0.05v_4 + 0.05v_5 \\
v_3 &= 0 - g + 0.7v_3 + 0.1v_4 + 0.2v_5 \\
v_4 &= 0 - 9 + 0.5v_4 + 0.5v_5 \\
v_5 &= 10 - g + v_6 \\
v_6 &= 0 - g + v_1 \\
v_6 &= 0,
\end{aligned}$$

where state $s = 6$ is chosen for the normalizing equation $v_s = 0$. The solution of these linear equations is given by

$$g(R^{(1)}) = 0.5128,\ v_1(R^{(1)}) = 0.5128,\ v_2(R^{(1)}) = 5.6410,\ v_3(R^{(1)}) = 7.4359,$$
$$v_4(R^{(1)}) = 8.4615,\ v_5(R^{(1)}) = 9.4872,\ v_6(R^{(1)}) = 0.$$

Table 6.4.1 The deteriorating probabilities q_{ij}

$i \backslash j$	1	2	3	4	5
1	0.90	0.10	0	0	0
2	0	0.80	0.10	0.05	0.05
3	0	0	0.70	0.10	0.20
4	0	0	0	0.50	0.50

Step 2 (policy improvement). The test quantity $T_i(a, R)$ has the values

$$T_2(0, R^{(1)}) = 5.6410, \quad T_2(1, R^{(1)}) = 7.0000, \quad T_3(0, R^{(1)}) = 7.4359,$$
$$T_3(1, R^{(1)}) = 7.0000, \quad T_4(0, R^{(1)}) = 9.4872, \quad T_4(1, R^{(1)}) = 5.0000.$$

This yields the new policy $R^{(2)} = (0, 0, 1, 1, 2, 2)$ by choosing for each state i the action a that minimizes $T_i(a, R^{(1)})$.
Step 3 (convergence test). The new policy $R^{(2)}$ is different from the previous policy $R^{(1)}$ and hence another iteration is performed.

Iteration 2
Step 1 (value determination). The average cost and the relative values of policy $R^{(2)} = (0, 0, 1, 1, 2, 2)$ are computed by solving the linear equations

$$\begin{aligned} v_1 &= 0 - g + 0.9v_1 + 0.1v_2 \\ v_2 &= 0 - g + 0.8v_2 + 0.1v_3 + 0.05v_4 + 0.05v_5 \\ v_3 &= 7 - g + v_1 \\ v_4 &= 5 - g + v_1 \\ v_5 &= 10 - g + v_6 \\ v_6 &= 0 - g + v_1 \\ v_6 &= 0. \end{aligned}$$

The solution of these linear equations is given by

$$g(R^{(2)}) = 0.4462, \quad v_1(R^{(2)}) = 0.4462, \quad v_2(R^{(2)}) = 4.9077, \quad v_3(R^{(2)}) = 7.000,$$
$$v_4(R^{(2)}) = 5.0000, \quad v_5(R^{(2)}) = 9.5538, \quad v_6(R^{(2)}) = 0.$$

Step 2 (policy improvement). The test quantity $T_i(a, R^{(2)})$ has the values

$$T_2(0, R^{(2)}) = 4.9077, \quad T_2(1, R^{(2)}) = 7.0000, \quad T_3(0, R^{(2)}) = 6.8646,$$
$$T_3(1, R^{(2)}) = 7.0000, \quad T_4(0, R^{(2)}) = 6.8307, \quad T_4(1, R^{(2)}) = 5.0000.$$

This yields the new policy $R^{(3)} = (0, 0, 0, 1, 2, 2)$.
Step 3 (convergence test). The new policy $R^{(3)}$ is different from the previous policy $R^{(2)}$ and hence another iteration is performed.

Iteration 3
Step 1 (value determination). The average cost and the relative values of policy $R^{(3)} = (0, 0, 0, 1, 2, 2)$ are computed by solving the linear equations

$$
\begin{aligned}
v_1 &= 0 - g + 0.9v_1 + 0.1v_2 \\
v_2 &= 0 - g + 0.8v_2 + 0.1v_3 + 0.05v_4 + 0.05v_5 \\
v_3 &= 0 - g + 0.7v_3 + 0.1v_4 + 0.2v_5 \\
v_4 &= 5 - g + v_1 \\
v_5 &= 10 - g + v_6 \\
v_6 &= 0 - g + v_1 \\
v_6 &= 0.
\end{aligned}
$$

The solution of these linear equations is given by

$$
g(R^{(3)}) = 0.4338,\ v_1(R^{(3)}) = 0.4338,\ v_2(R^{(3)}) = 4.7717,\ v_3(R^{(3)}) = 6.5982,
$$
$$
v_4(R^{(3)}) = 5.0000,\ v_5(R^{(3)}) = 9.5662,\ v_6(R^{(3)}) = 0.
$$

Step 2 (policy improvement). The test quantity $T_i(a, R^{(3)})$ has the values

$$
T_2(0, R^{(3)}) = 4.7717,\ T_2(1, R^{(3)}) = 7,\ T_3(0, R^{(3)}) = 6.5987,
$$
$$
T_3(1, R^{(3)}) = 7.0000,\ T_4(0, R^{(3)}) = 6.8493,\ T_4^{(1)}(1, R^{(3)}) = 5.0000.
$$

This yields the new policy $R^{(4)} = (0, 0, 0, 1, 2, 2)$.
Step 3 (convergence test). The new policy $R^{(4)}$ is identical to the previous policy $R^{(3)}$ and is thus average cost optimal. The minimal average cost is 0.4338 per day.

Remark 6.4.2 Deterministic state transitions
For the case of deterministic state transitions the computational burden of policy iteration can be reduced considerably. Instead of solving a system of linear equations at each step, the average cost and relative values can be obtained from recursive calculations. The reason for this is that under each stationary policy the process moves cyclically among the recurrent states. The simplified policy-iteration calculations for deterministic state transitions are as follows:

(a) Determine for the current policy R the cycle of recurrent states among which the process cyclically moves.

(b) The cost rate $g(R)$ equals the sum of one-step costs in the cycle divided by the number of states in the cycle.

(c) The relative values for the recurrent states are calculated recursively, in reverse direction to the natural flow around the cycle, after assigning a value 0 to one recurrent state.

(d) The relative values for transient states are computed first for states which reach the cycle in one step, then for states which reach the cycle in two steps, and so forth.

It is worthwhile pointing out that the simplified policy-iteration algorithm may be an efficient technique to compute a minimum cost-to-time circuit in a deterministic network.

6.5 LINEAR PROGRAMMING APPROACH*

The policy-iteration algorithm solves the average cost optimality equation (6.4.3) in a finite number of steps by generating a sequence of improved policies. Another way of solving the optimality equation is the use of a linear program for the average cost case. The linear programming formulation to be given below allows the unichain assumption in Section 6.4 to be weakened as follows.

Weak unichain assumption *For each average cost optimal stationary policy the associated Markov chain* $\{X_n\}$ *has no two disjoint closed sets.*

This assumption allows non-optimal policies to have multiple disjoint closed sets. The unichain assumption in Section 6.4 may be too strong for some applications; for example, in inventory problems with strictly bounded demands it may be possible to construct stationary policies with disjoint ordering regions such that the levels between which the stock fluctuates remain dependent on the initial level. However, the weak unichain assumption will practically always be satisfied in real-world applications. For the weak unichain case, the minimal average cost per time unit is independent of the initial state and, moreover, the average cost optimality equation (6.4.3) applies and uniquely determines g^* as the minimal average cost per time unit; see Denardo and Fox (1968) for a proof. This reference also gives the following linear programming algorithm for the computation of an average cost optimal policy.

Linear programming algorithm

Step 1. Apply the simplex method to compute an optimal basic solution (x_{ia}^*) to the following linear program:

$$\text{Minimize} \sum_{i \in I} \sum_{a \in A(i)} c_i(a) x_{ia} \tag{6.5.1}$$

subject to

$$\sum_{a \in A(j)} x_{ja} - \sum_{i \in I} \sum_{a \in A(i)} p_{ij}(a) x_{ia} = 0, \quad j \in I,$$

$$\sum_{i \in I} \sum_{a \in A(i)} x_{ia} = 1,$$

$$x_{ia} \geq 0, \quad a \in A(i) \text{ and } i \in I.$$

*This section may be skipped at first reading.

Step 2. Start with the non-empty set $I_0 := \{i \mid \sum_{a \in A(i)} x^*_{ia} > 0\}$. For any state $i \in I_0$, set the decision

$$R^*_i := a \quad \text{for some } a \text{ such that } x^*_{ia} > 0.$$

Step 3. If $I_0 = I$, then the algorithm is stopped with policy R^*. Otherwise, determine some state $i \notin I_0$ and action $a \in A(i)$ such that $p_{ij}(a) > 0$ for some $j \in I_0$. Next set $R^*_i := a$ and $I_0 := I_0 \cup \{i\}$ and repeat step 3.

The linear program (6.5.1) can heuristically be explained by interpreting the variables x_{ia} as

$$x_{ia} = \text{the long-run fraction of decision epochs at which the system is in state } i \text{ and action } a \text{ is made.}$$

The objective of the linear program is the minimization of the long-run average cost per time unit, while the first set of constraints represent the balance equations requiring that for any state $j \in I$ the long-run average number of transitions from state j per time unit must be equal to the long-run average number of transitions into state j per time unit. The last constraint obviously requires that the sum of the fractions x_{ia} must be equal to 1.

Next we sketch a proof that the linear programming algorithm leads to an average cost optimal policy R^* when the weak unichain assumption is satisfied. Our starting point is the average cost optimality equation (6.4.3). Since this equation is solvable, the linear inequalities

$$g + v_i - \sum_{j \in I} p_{ij}(a) v_j \leq c_i(a), \quad a \in A(i) \text{ and } i \in I \tag{6.5.2}$$

must have a solution. It follows from Theorem 6.2.1 that any solution $\{g, v_i\}$ to these inequalities satisfies $g \leq g_i(R)$ for any $i \in I$ and any policy R. Hence we can conclude that for any solution $\{g, v_i\}$ to the linear inequalities (6.5.2) holds that $g \leq g^*$ with g^* being the minimal average cost per time unit. Hence, using the fact that relative values v^*_i, $i \in I$, exist such that $\{g^*, v^*_i\}$ constitutes a solution to (6.5.2), the linear program

$$\text{Maximize } g \tag{6.5.3}$$

subject to

$$g + v_i - \sum_{j \in I} p_{ij}(a) v_j \leq c_i(a), \quad a \in A(i) \text{ and } i \in I,$$
$$g, v_i \text{ unrestricted,}$$

has the minimal average cost g^* as the optimal objective-function value. Next observe that the linear program (6.5.1) is the dual of the primal linear program

(6.5.3). By the dual theorem of linear programming, the primal and dual linear programs have the same optimal objective-function value. Hence the minimal objective-function value of the linear program (6.5.1) equals the minimal average cost g^*. Next we show that an optimal basic solution (x_{ia}^*) to the linear program (6.5.1) induces an average cost optimal policy. To do so, define the set

$$S_0 = \left\{ i \Big| \sum_{a \in A(i)} x_{ia}^* > 0 \right\}.$$

Then the set S_0 is closed under any policy R having the property that action $a = R_i$ satisfies $x_{ia}^* > 0$ for all $i \in S_0$. To see this, suppose that $p_{ij}(R_i) > 0$ for some $i \in S_0$ and $j \notin S_0$. Then the first set of constraints of the linear program (6.5.1) implies that $\sum_a x_{ja}^* > 0$, contradicting $j \notin S_0$. Next consider the set I_0 as constructed in the linear programming algorithm. Let R^* be a policy such that the actions R_i^* for $i \in I_0$ are chosen according to the algorithm. It remains to verify that $I_0 = I$ and that policy R^* is average cost optimal. To do so, let $\{g^*, v_i^*\}$ be the particular optimal basic solution to the primal linear program (6.5.3) such that this basic solution is complementary to the optimal basic solution (x_{ia}^*) of the dual linear program (6.5.1). Then, by the complementary slackness property of linear programming,

$$g^* + v_i^* - \sum_{j \in I} p_{ij}(R_i^*) v_j^* = c_i(R_i^*), \quad i \in S_0.$$

The term $\sum_{j \in I} p_{ij}(R_i^*) v_j^*$ can be replaced by $\sum_{j \in S_0} p_{ij}(R_i^*) v_j^*$ for $i \in S_0$, since the set S_0 is closed under policy R^*. Thus, by Theorem 6.2.1, we can conclude that $g_i(R^*) = g^*$ for all $i \in S_0$. The states in $I_0 \backslash S_0$ are transient under policy R^* and are ultimately leading to a state in S_0. Hence $g_i(R^*) = g^*$ for all $i \in I_0$. To prove that $I_0 = I$, assume to the contrary that $I_0 \neq I$. By the construction of I_0, the set $I \backslash I_0$ is closed under any policy. Let R_0 be any average cost optimal policy. Define the policy R_1 by

$$R_1(i) = \begin{cases} R^*(i), & i \in I_0, \\ R_0(i), & i \in I \backslash I_0. \end{cases}$$

Since $I \backslash I_0$ and I_0 are both closed sets under policy R_1, we have constructed an average cost optimal policy with two disjoint closed sets. This contradicts the weak unichain assumption. Hence $I_0 = I$. This completes the proof.

We illustrate the linear programming formulation of the Markov decision problem from Example 6.1.1. The specification of the basic elements of the Markov decision model for this problem is given in Section 6.1.

Example 6.1.1 (continued) A maintenance problem

The linear programming formulation for this problem is to minimize

$$\sum_{i=2}^{N-1} C_{pi} x_{i1} + C_f x_{N2}$$

subject to

$$x_{10} - \left(q_{11} x_{10} + \sum_{i=2}^{N-1} x_{i1} + x_{N+1,2} \right) = 0,$$

$$x_{j0} + x_{j1} - \sum_{i=1}^{j} q_{ij} x_{i0} = 0, \quad 2 \le j \le N-1,$$

$$x_{N2} - \sum_{i=1}^{N-1} q_{iN} x_{i0} = 0,$$

$$x_{N+1,2} - x_{N2} = 0,$$

$$x_{10} + \sum_{i=2}^{N-1} (x_{i0} + x_{i1}) + x_{N2} + x_{N+1,2} = 1,$$

$$x_{10}, x_{i0}, x_{i1}, x_{N2}, x_{N+1,2} \ge 0.$$

For the numerical data given in Table 6.4.1, this linear program has the minimal objective value 0.4338 and the optimal basic solution

$$x_{10}^* = 0.5479,\ x_{20}^* = 0.2740,\ x_{30}^* = 0.0913,\ x_{41}^* = 0.0228,$$
$$x_{52}^* = 0.0320,\ x_{62}^* = 0.0320 \text{ and the other } x_{ia}^* = 0.$$

This yields the average cost optimal policy $R^* = (0, 0, 0, 1, 2, 2)$ with an average cost of 0.4338, in agreement with the results obtained by policy iteration.

Linear programming and probabilistic constraints

The linear programming formulation may often be a convenient way to handle Markovian decision problems with probabilistic constraints. In many practical applications, constraints are imposed on certain state frequencies. For example, in inventory problems for which shortage costs are difficult to estimate, probabilistic constraints may be placed on the probability of shortage or on the fraction of demand that cannot be met directly from stock on hand. Similarly, in a maintenance

problem involving a randomly changing state, a constraint may be placed on the frequency at which a certain inoperative state occurs.

The following illustrative example taken from Wagner (1975) shows that for control problems with probabilistic constraints it may be optimal to choose the decisions in a random way rather than in a deterministic way. Suppose the daily demand D for some product is described by the probability distribution

$$P\{D = 0\} = P\{D = 1\} = \tfrac{1}{6}, \quad P\{D = 2\} = \tfrac{2}{3}.$$

The demands on the successive days are independent of each other. At the beginning of each day it has to be decided how much to order of the product. The delivery of any order is instantaneous. The variable ordering cost of each unit is $c > 0$. Any unit that is not sold at the end of the day becomes obsolete and must be discarded. The decision problem is to minimize the average ordering cost per day, subject to the constraint that the fraction of the demand to be met is at least $\frac{1}{3}$. This probabilistic constraint is satisfied when using the policy of ordering one unit every day, a policy which has an average cost of c per day. However, this deterministic control rule is not optimal, as can be seen by considering the randomized control rule under which at any given day no unit is ordered with probability $\frac{4}{5}$ and two units are ordered with probability $\frac{1}{5}$. Under this randomized rule the probability that the daily demand is met equals $(\frac{4}{5})(\frac{1}{6}) + (\frac{1}{5})(1) = \frac{1}{3}$ and the average ordering cost per day equals $(\frac{4}{5})(0) + (\frac{1}{5})(2c) = \frac{2}{5}c$. It is readily seen that the randomized rule is optimal.

So far we have considered only stationary policies under which the actions are chosen deterministically. A policy π is called a stationary *randomized* policy when it is described by a probability distribution $\{\pi_a(i), a \in A(i)\}$ for each state $i \in I$. Under policy π action $a \in A(i)$ is chosen with probability $\pi_a(i)$ whenever the process is in state i. If $\pi_a(i)$ is 0 or 1 for every i and a, the stationary randomized policy π reduces to the familiar stationary policy choosing the actions in a deterministic way. For any policy π, let the state-action frequencies $f_{i,a}(\pi)$ be defined by

$$f_{ia}(\pi) = \text{the long-run fraction of decision epochs at which the process is in state } i \text{ and action } a \text{ is chosen when policy } \pi \text{ is used.}$$

Consider now a Markovian decision problem in which the goal is to minimize the long-run average cost per time unit subject to the following linear constraints on the state-action frequencies:

$$\sum_{i \in I} \sum_{a \in A(i)} \alpha_{ia}^{(s)} f_{ia}(\pi) \leq \beta^{(s)}, \quad s = 1, \ldots, L,$$

where $\alpha_{ia}^{(s)}$ and $\beta^{(s)}$ are given constants. It is assumed that the constraints allow for a feasible solution. If the unichain assumption from Section 6.4 holds, it can be shown that an optimal policy may be obtained by solving the following linear

program; see Derman (1970) and Hordijk and Kallenberg (1984):

$$\text{Minimize} \sum_{i \in I} \sum_{a \in A(i)} c_i(a) x_{ia}$$

subject to

$$\sum_{a \in A(j)} x_{ja} - \sum_{i \in I} \sum_{a \in A(i)} p_{ij}(a) x_{ia} = 0, \quad j \in I,$$

$$\sum_{i \in I} \sum_{a \in A(i)} x_{ia} = 1,$$

$$\sum_{i \in I} \sum_{a \in A(i)} \alpha_{ia}^{(s)} x_{ia} \leq \beta^{(s)}, \quad s = 1, \ldots, L,$$

$$x_{ia} \geq 0, \quad a \in A(i) \text{ and } i \in I.$$

Denoting by $\{x_{ia}^*\}$ an optimal basic solution to this linear program and letting the set $S_0 = \{i \mid \sum_a x_{ia}^* > 0\}$, an optimal stationary randomized policy π^* is given by

$$\pi_a^*(i) = \begin{cases} x_{ia}^* / \sum_d x_{id}^*, & a \in A(i) \text{ and } i \in S_0, \\ \text{arbitrary}, & \text{otherwise.} \end{cases}$$

Here the unichain assumption is essential for guaranteeing the existence of an optimal stationary randomized policy.

Example 6.1.1 (continued) A maintenance problem

Suppose that in the maintenance problem a probabilistic constraint is imposed on the fraction of time the system is in repair. It is required that this fraction is no more than 0.08. To handle this constraint, we add to the previous linear program for the maintenance problem the constraint

$$\sum_{i=2}^{N-1} x_{i1} + x_{N2} + x_{N+1,2} \leq 0.08.$$

The new linear program has the optimal solution

$$x_{10}^* = 0.5943, \; x_{20}^* = 0.2971, \; x_{30}^* = 0.0286, \; x_{31}^* = 0.0211,$$
$$x_{41}^* = 0.0177, \; x_{52}^* = x_{62}^* = 0.0206 \text{ and the other } x_{ia}^* = 0.$$

The minimal cost is 0.4423 and the fraction of time the system is in repair is exactly 0.08. The LP solution corresponds to a randomized policy. The actions 0, 0, 1, 2 and 2 are prescribed in the states 1, 2, 4, 5 and 6. In state 3 a biased coin is tossed. The coin shows up heads with probability $0.0286/(0.0286 + 0.0211) = 0.575$. No preventive repair is done if heads comes up, otherwise a preventive repair is done.

A Lagrange-multiplier approach for probabilistic constraints

A heuristic approach for handling probabilistic constraints is the Lagrange-multiplier method. This method produces only stationary non-randomized policies. To describe the method, assume a single probabilistic constraint

$$\sum_{i \in I} \sum_{a \in A(i)} \alpha_{ia} f_{ia}(\pi) \leq \beta$$

on the state-action frequencies. In the Lagrange-multiplier method, the constraint is eliminated by putting it into the criterion function by means of a Lagrange multiplier $\lambda \geq 0$. That is, the goal function is changed from $\sum_{i,a} c_i(a) x_{ia}$ to $\sum_{i,a} c_i(a) x_{ia} + \lambda(\sum_{i,a} \alpha_{ia} x_{ia} - \beta)$. The Lagrange multiplier may be interpreted as the cost to each unit that is used from some resource. The original Markov decision problem without probabilistic constraint is obtained by taking $\lambda = 0$. It is assumed that the probabilistic constraint is not satisfied for the optimal stationary policy in the unconstrained problem; otherwise, this policy is optimal for the constrained problem as well. Thus, for a given value of the Lagrange multiplier $\lambda > 0$, we consider the unconstrained Markov decision problem with one-step costs

$$c_i^{\lambda}(a) = c_i(a) + \lambda \alpha_{ia}$$

and one-step transition probabilities $p_{ij}(a)$ as before. Solving this unconstrained Markov decision problem yields an optimal deterministic policy $R(\lambda)$ that prescribes always a fixed action $R_i(\lambda)$ whenever the system is in state i. Let $\beta(\lambda)$ be the constraint level associated with policy $R(\lambda)$, that is,

$$\beta(\lambda) = \sum_{i \in I} \alpha_{i, R_i(\lambda)} f_{i, R_i(\lambda)}(R(\lambda)).$$

If $\beta(\lambda) > \beta$ one should increase λ, otherwise one should decrease λ. Why? The Lagrange multiplier λ should be adjusted until the *smallest* value of λ is found for which $\beta(\lambda) \leq \beta$. Bisection is a convenient method to adjust λ. How do we calculate $\beta(\lambda)$ for a given value of λ? To do so, observe that $\beta(\lambda)$ can be interpreted as the average cost in a single Markov chain with an appropriate cost structure. Consider the Markov chain describing the state of the system under policy $R(\lambda)$. In this Markov process, the long-run average cost per time unit equals $\beta(\lambda)$ when it is assumed that a direct cost of $\alpha_{i, R_i(\lambda)}$ is incurred each time the process visits state i. An effective method to compute the average cost $\beta(\lambda)$ is to apply value iteration to a single Markov chain; see Example 6.6.1 in the next section.

The average cost of the stationary policy obtained by the Lagrangian approach will in general be larger than the average cost of the stationary randomized policy resulting from the linear programming formulation. Also, it should be pointed out that there is no guarantee that the policy obtained by the Lagrangian approach is the best policy among all stationary policies satisfying the probabilistic constraint, although in most practical situations this may be expected to be the case. In spite

of the possible pitfalls of the Lagrangian approach, this approach may be quite useful in practical applications having a specific structure.

6.6 VALUE-ITERATION ALGORITHM

The policy-iteration algorithm and the linear programming formulation both require that in each iteration a system of linear equations of the same size as the state space is solved. In general, this will be computationally burdensome for a large state space and makes these algorithms computationally unattractive for large-scale Markov decision problems. In this section we discuss an alternative algorithm which avoids solving systems of linear equations but uses instead the recursive solution approach from dynamic programming. This method is the *value-iteration algorithm* which computes recursively a sequence of value functions approximating the minimal average cost per time unit. The value functions provide lower and upper bounds on the minimal average cost and under a certain aperiodicity condition these bounds converge to the minimal average cost. The aperiodicity condition is not restrictive, since it can be forced to hold by a simple data transformation. The value-iteration algorithm endowed with these lower and upper bounds is in general the best computational method for solving large-scale Markov decision problems. This is even true in spite of the fact that the value-iteration algorithm does not have the robustness of the policy-iteration algorithm: the number of iterations is problem dependent and typically increases in the number of states of the problem under consideration. Another important advantage of value iteration is that it is usually easy to write a code for specific applications. By exploiting the structure of the particular application one usually avoids computer memory problems that may be encountered when using policy iteration. Value iteration is not only a powerful method for controlled Markov chains, but it is also a useful tool to compute bounds on performance measures in a single Markov chain; see Example 6.6.1.

In this section the value-iteration algorithm will be analysed under the weak unichain assumption from Section 6.5. Under this assumption the minimal average cost per time unit is independent of the initial state. Let

$$g^* = \text{the minimal long-run average cost per time unit.}$$

The value-iteration algorithm computes recursively for $n = 1, 2, \ldots$ the value function $V_n(i)$ from

$$V_n(i) = \min_{a \in A(i)} \left\{ c_i(a) + \sum_{j \in I} p_{ij}(a) V_{n-1}(j) \right\}, \quad i \in I, \tag{6.6.1}$$

starting with an arbitrarily chosen function $V_0(i)$, $i \in I$. The quantity $V_n(i)$ can be interpreted as the minimal total expected costs with n periods left to the time horizon when the current state is i and a terminal cost of $V_0(j)$ is incurred when the system ends up at state j; see Denardo (1982) and Derman (1970) for a proof.

Intuitively, one might expect that the one-step difference $V_n(i) - V_{n-1}(i)$ will come very close to the minimal average cost per time unit and that the stationary policy whose actions minimize the right-hand side of (6.6.1) for all i will be very close in cost to the minimal average cost. However, these matters appear to be rather subtle for the average cost criterion due to the effect of possible periodicities in the underlying decision processes. Before explaining this in more detail, we investigate an operator which is induced by the recursion equation (6.6.1). The operator T adds to each function $v = (v_i,\ i \in I)$ a function Tv whose ith component $(Tv)_i$ is defined by

$$(Tv)_i = \min_{a \in A(i)} \left\{ c_i(a) + \sum_{j \in I} p_{ij}(a) v_j \right\}, \quad i \in I. \tag{6.6.2}$$

Note that $(Tv)_i = V_n(i)$ if $v_i = V_{n-1}(i)$, $i \in I$. The following theorem plays a key role in the value-iteration algorithm.

Theorem 6.6.1 *Suppose that the weak unichain assumption is satisfied. Let $v = (v_i)$ be given. Define the stationary policy $R(v)$ as a policy which adds to each state $i \in I$ an action $a = R_i(v)$ that minimizes the right-hand side of (6.6.2). Then*

$$\min_{i \in I} \{(Tv)_i - v_i\} \leq g^* \leq g_s(R(v)) \leq \max_{i \in I} \{(Tv)_i - v_i\} \tag{6.6.3}$$

for any $s \in I$, where g^* is the minimal long-run average cost per time unit and $g_s(R(v))$ denotes the long-run average cost per time unit under policy $R(v)$ when the initial state is s.

Proof To prove the first inequality, choose any stationary policy R. By the definition of $(Tv)_i$, we have for any state $i \in I$ that

$$(Tv)_i \leq c_i(a) + \sum_{j \in I} p_{ij}(a) v_j, \quad a \in A(i), \tag{6.6.4}$$

where the equality sign holds for $a = R_i(v)$. Choosing $a = R_i$ in (6.6.4) gives

$$(Tv)_i \leq c_i(R_i) + \sum_{i \in I} p_{ij}(R_i) v_j, \quad i \in I. \tag{6.6.5}$$

Define the lower bound

$$m = \min_{i \in I} \{(Tv)_i - v_i\}.$$

Since $m \leq (Tv)_i - v_i$ for all i, it follows from (6.6.5) that $m + v_i \leq c_i(R_i) + \sum_{j \in I} p_{ij}(R_i) v_j$ for all $i \in I$, and so

$$c_i(R_i) - m + \sum_{j \in I} p_{ij}(R_i) v_j \geq v_i, \quad i \in I.$$

An application of Theorem 6.2.1 now gives that

$$g_i(R) \geq m, \quad i \in I.$$

This inequality holds for each policy R and so $g^* = \min_R g_i(R) \geq m$ proving the first inequality in (6.6.3). The proof of the last inequality in (6.6.3) is very similar. By the definition of policy $R(v)$,

$$(Tv)_i = c_i(R_i(v)) + \sum_{j \in I} p_{ij}(R_i(v))v_j, \quad i \in I. \tag{6.6.6}$$

Define the upper bound

$$M = \max_{i \in I} \{(Tv)_i - v_i\}.$$

Since $M \geq (Tv)_i - v_i$ for all $i \in I$, we obtain from (6.6.6) that

$$c_i(R_i(v)) - M + \sum_{j \in I} p_{ij}(R_i(v))v_j \leq v_i, \quad i \in I.$$

Hence, by Theorem 6.2.1, $g_i(R(v)) \leq M$ for all $i \in I$, proving the last inequality in (6.6.3). This completes the proof.

We now formulate the value-iteration algorithm. In the formulation it is no restriction to assume that

$$c_i(a) > 0 \quad \text{for all } i \in I \text{ and } a \in A(i).$$

Otherwise, add a sufficiently large positive constant to each $c_i(a)$. This affects the average cost of each policy by the same constant.

Value-iteration algorithm

Step 0 (initialization). Choose $V_0(i)$, $i \in I$ with $0 \leq V_0(i) \leq \min_a c_i(a)$. Let $n := 1$.

Step 1 (value-iteration step). For each state $i \in I$, compute

$$V_n(i) = \min_{a \in A(i)} \left\{ c_i(a) + \sum_{j \in I} p_{ij}(a) V_{n-1}(j) \right\}.$$

Let $R(n)$ be any stationary policy such that the action $a = R_i(n)$ minimizes the right-hand side of the equation for $V_n(i)$ for each state i.

Step 2 (bounds on the minimal costs). Compute the bounds

$$m_n = \min_{i \in I} \{V_n(i) - V_{n-1}(i)\}, \quad M_n = \max_{i \in I} \{V_n(i) - V_{n-1}(i)\}.$$

Step 3 (stopping test). If

$$0 \leq M_n - m_n \leq \varepsilon m_n$$

with $\varepsilon > 0$ a prespecified accuracy number (e.g. $\varepsilon = 10^{-3}$), stop with policy $R(n)$. *Step 4 (continuation).* $n := n + 1$ and repeat step 1.

By Theorem 6.6.1, we have

$$0 \leq \frac{g_i(R(n)) - g^*}{g^*} \leq \frac{M_n - m_n}{m_n} \leq \varepsilon, \quad i \in I \tag{6.6.7}$$

when the algorithm is stopped after the nth iteration with policy $R(n)$. In other words, the average cost of policy $R(n)$ cannot deviate more than $100\varepsilon\%$ from the theoretically minimal average cost when the bounds m_n and M_n satisfy $0 \leq M_n - m_n \leq \varepsilon m_n$. In practical applications one is usually satisfied with a policy whose average cost is sufficiently close to the theoretically minimal average cost.

Convergence of the bounds

The remaining question is whether the lower and upper bounds m_n and M_n converge to the same limit so that the algorithm will be stopped after finitely many iterations. The answer is yes only if a certain *aperiodicity* condition is satisfied. In general m_n and M_n need not have the same limit, as the following example demonstrates. Consider the trivial Markov decision problem with two states 1 and 2 and a single action a_0 in each state. The one-step costs and the one-step transition probabilities are given by $c_1(a_0) = 1$, $c_2(a_0) = 0$, $p_{12}(a_0) = p_{21}(a_0) = 1$ and $p_{11}(a_0) = p_{22}(a_0) = 0$. Then the system cycles between the states 1 and 2. It is easily verified that $V_{2k}(1) = V_{2k}(2) = k$, $V_{2k-1}(1) = k$ and $V_{2k-1}(2) = k - 1$ for all $k \geq 1$. Hence $m_n = 0$ and $M_n = 1$ for all n, implying that the sequences $\{m_n\}$ and $\{M_n\}$ have different limits. The reason for the oscillating behaviour of $V_n(i) - V_{n-1}(i)$ is the periodicity of the Markov chain describing the state of the system. The next theorem gives sufficient conditions for the convergence of the value-iteration algorithm.

Theorem 6.6.2 *Suppose that the weak unichain assumption holds and that for each average cost optimal stationary policy the associated Markov chain $\{X_n\}$ is aperiodic. Then there are finite constants $\alpha > 0$ and $0 < \beta < 1$ such that*

$$|M_n - m_n| \leq \alpha\beta^n, \quad n \geq 1.$$

In particular, $\lim_{n\to\infty} M_n = \lim_{n\to\infty} m_n = g^*$.

A proof of this deep theorem will not be given. A special case of the theorem will be proved in Section 6.7. This special case is related to the data transformation by which the periodicity issue can be circumvented. Before discussing this data transformation, we prove the interesting result that the sequences $\{m_n\}$ and $\{M_n\}$ are always *monotone* irrespective of the chain structure of the Markov chains.

Theorem 6.6.3 *In the standard value-iteration algorithm the lower and upper bounds satisfy*

$$m_{k+1} \geq m_k \quad \textit{and} \quad M_{k+1} \leq M_k \quad \textit{for all} \quad k \geq 1.$$

Proof By the definition of policy $R(n)$,

$$V_n(i) = c_i(R_i(n)) + \sum_{j \in I} p_{ij}(R_i(n)) V_{n-1}(j), \quad i \in I. \tag{6.6.8}$$

In the same way as (6.6.5) was obtained, we find for any policy R that

$$c_i(R_i) + \sum_{j \in I} p_{ij}(R_i) V_{n-1}(j) \geq V_n(i), \quad i \in I. \tag{6.6.9}$$

Taking $n = k$ in (6.6.8) and taking $n = k + 1$ and $R = R(k)$ in (6.6.9) gives

$$V_{k+1}(i) - V_k(i) \leq \sum_{j \in I} p_{ij}(R_i(k)) \{V_k(j) - V_{k-1}(j)\}, \quad i \in I. \tag{6.6.10}$$

Similarly, by taking $n = k + 1$ in (6.6.8) and taking $n = k$ and $R = R(k + 1)$ in (6.6.9), we find

$$V_{k+1}(i) - V_k(i) \geq \sum_{j \in I} p_{ij}(R_i(k+1)) \{V_k(j) - V_{k-1}(j)\}, \quad i \in I. \tag{6.6.11}$$

Since $V_k(j) - V_{k-1}(j) \leq M_k$ for all $j \in I$ and $\sum_{j \in I} p_{ij}(R_i(k)) = 1$, it follows from (6.6.10) that $V_{k+1}(i) - V_k(i) \leq M_k$ for all $i \in I$. This gives $M_{k+1} \leq M_k$. Similarly, we obtain from (6.6.11) that $m_{k+1} \geq m_k$.

Data transformation

The periodicity issue can be circumvented by a *perturbation* of the one-step transition probabilities. The perturbation technique is based on the following two observations. First, a recurrent state allowing for a direct transition to itself must be aperiodic. Second, the relative frequencies at which the states of a Markov chain are visited do not change when the state changes are delayed with a constant factor and the probability of a self-transition is accordingly enlarged. In other words, if the one-step transition probabilities p_{ij} of a Markov chain $\{X_n\}$ are perturbed as $\overline{p}_{ij} = \tau p_{ij}$ for $j \neq i$ and $\overline{p}_{ii} = \tau p_{ii} + 1 - \tau$ for some constant τ with $0 < \tau < 1$, the perturbed Markov chain $\{\overline{X}_n\}$ with one-step transition probabilities $\overline{p}_{ij}$ is aperiodic and has the same equilibrium probabilities as the original Markov chain $\{X_n\}$ (verify). Thus a Markov decision model involving periodicities may be perturbed as follows. Choosing some constant τ with $0 < \tau < 1$, the state space, the action sets, the one-step costs and the one-step transition probabilities of the perturbed

Markov decision model are defined by

$$\begin{aligned}
\overline{I} &= I, \\
\overline{A}(i) &= A(i), \quad i \in \overline{I}, \\
\overline{c}_i(a) &= c_i(a), \quad a \in \overline{A}(i) \text{ and } i \in \overline{I}, \\
p_{ij}(a) &= \begin{cases} \tau p_{ij}(a), & j \neq i,\ a \in \overline{A}(i) \text{ and } i \in \overline{I}, \\ \tau p_{ij}(a) + 1 - \tau, & j = i,\ a \in \overline{A}(i) \text{ and } i \in \overline{I}. \end{cases}
\end{aligned}$$

For each stationary policy, the associated Markov chain $\{\overline{X}_n\}$ in the perturbed model is aperiodic. It is not difficult to verify that for each stationary policy the average cost per time unit in the perturbed model is the same as that in the original model. For the unichain case this is an immediate consequence of the representation (6.2.7) for the average cost and the fact that for each stationary policy the Markov chain $\{\overline{X}_n\}$ has the same equilibrium probabilities as the Markov chain $\{X_n\}$ in the original model. For the multichain case, a similar argument can be used to show that the two models are in fact equivalent. Thus the value-iteration algorithm can be applied to the perturbed model in order to solve the original model. In specific problems involving periodicities, the 'optimal' value of τ is usually not clear beforehand; empirical investigations indicate that $\tau = \frac{1}{2}$ is usually a satisfactory choice.

Modified value iteration with a dynamic relaxation factor

Value iteration does not have the fast convergence of policy iteration. The number of iterations required by the value-iteration algorithm is problem dependent and increases when the number of problem states gets larger. Also, the tolerance number ε in the stopping criterion affects the number of iterations required. The stopping criterion should be based on the lower and upper bounds m_n and M_n but not on any repetitive behaviour of the generated policies $R(n)$.

The convergence rate of value iteration can often be accelerated by using a relaxation factor, such as in successive overrelaxation for solving a single system of linear equations. Then at the nth iteration a new approximation to the value function $V_n(i)$ is obtained by using both the previous values $V_{n-1}(i)$ and the residuals $V_n(i) - V_{n-1}(i)$. It is possible to select dynamically a relaxation factor and thus avoid the experimental determination of the best value of a fixed relaxation factor. The following modification of the standard value-iteration algorithm can be formulated. Steps 0, 1, 2 and 3 are as before, while step 4 of the standard value-iteration algorithm is modified as follows.

Step 4(a). Determine the states u and v such that

$$V_n(u) - V_{n-1}(u) = m_n \quad \text{and} \quad V_n(v) - V_{n-1}(v) = M_n$$

and compute the relaxation factor

$$\omega = \frac{M_n - m_n}{M_n - m_n + \sum_{j \in I} \{p_{uj}(R_u) - p_{vj}(R_v)\}\{V_n(j) - V_{n-1}(j)\}},$$

where R_u and R_v are the actions which are prescribed by policy $R(n)$ in the states u and v.

Step 4(b). For each $i \in I$, change $V_n(i)$ according to

$$V_n(i) := V_{n-1}(i) + \omega\{V_n(i) - V_{n-1}(i)\}.$$

Step 4(c). $n := n + 1$ and go to step 1.

In the case of a tie when selecting in step 4(a) the state u for which the minimum in m_n is obtained, it is conventional to choose the minimizing state of the previous iteration when that state is one of the candidates to choose; otherwise, choose the first state achieving the minimum in m_n. The same convention is used for the maximizing action v in M_n.

The choice of the dynamic relaxation factor ω is motivated as follows. We change the estimate $V_n(i)$ as $\overline{V}_n(i) = V_{n-1}(i) + \omega\{V_n(i) - V_{n-1}(i)\}$ for all i in order to accomplish at the $(n+1)$th iteration that

$$c_u(R_u) + \sum_{j \in I} p_{uj}(R_u)\overline{V}_n(j) - \overline{V}_n(u) = c_v(R_v) + \sum_{j \in I} p_{vj}(R_v)\overline{V}_n(j) - \overline{V}_n(v),$$

in the implicit hope that the difference between the new upper and lower bounds M_{n+1} and m_{n+1} will decrease more quickly. Using the relation $m_n = V_n(u) - V_{n-1}(u) = c_u(R_u) + \sum_j p_{uj}(R_u)V_{n-1}(j) - V_{n-1}(u)$ and the similar relation for M_n, it is a matter of simple algebra to verify from the above condition the expression for ω. We omit the easy proof that $\omega > 0$. Numerical experiments indicate that using a dynamic relaxation factor in value iteration often greatly enhances the speed of convergence of the algorithm. The modified value-iteration algorithm is theoretically not guaranteed to converge, but in practice the algorithm will usually work very well. It is important to note that the relaxation factor ω is kept outside the recursion equation in step 1 so that the bounds m_n and M_n in step 2 are not destroyed. Although the bounds apply, it is no longer true that the sequences $\{m_n\}$ and $\{M_n\}$ are monotonic.

To conclude this section, we apply value iteration to two examples. The first example concerns the maintenance problem from Example 6.1.1 and the second example illustrates the usefulness of value iteration for the computation of performance measures for a single Markov chain.

Example 6.1.1 (continued) A maintenance problem

For the maintenance problem the recursion equation (6.6.1) becomes

$$V_n(1) = 0 + \sum_{j=1}^{N} q_{1j} V_{n-1}(j),$$

$$V_n(i) = \min\left\{0 + \sum_{j=i}^{N} q_{ij} V_{n-1}(j),\ C_{pi} + V_{n-1}(1)\right\}, \quad 1 < i < N,$$

$$V_n(N) = C_f + V_{n-1}(N+1),$$
$$V_n(N+1) = 0 + V_{n-1}(1).$$

We have applied the standard value-iteration algorithm to the numerical data from Table 6.4.1. For each stationary policy the associated Markov chain $\{X_n\}$ is aperiodic. Taking $V_0(i) = 0$ for all i and the accuracy number $\varepsilon = 10^{-3}$, the algorithm is stopped after $n = 28$ iterations with the stationary policy $R(n) = (0, 0, 0, 1, 2, 2)$ together with the lower and upper lower bounds $m_n = 0.4336$ and $M_n = 0.4340$. The average cost of policy $R(n)$ is estimated by $\frac{1}{2}(m_n + M_n) = 0.4338$ and this cost cannot deviate more than 0.1% from the theoretically minimal average cost. In fact policy $R(n)$ is optimal as we know from previous results obtained by policy iteration. To get a feeling of how strongly the required number of iterations depends on ε, we applied standard value-iteration for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-4}$ as well. For these choices of the accuracy number ε, standard value-iteration required 21 and 35 iterations respectively.

Example 6.6.1 A finite-capacity queue with deterministic arrivals

Consider a single-server queueing system having a finite waiting room for K customers (including any customer in service). The arrival process of customers is deterministic. Every D time units a customer arrives. A customer finding a full waiting room upon arrival is lost. The service times of the customers are independent random variables having an Erlang (r, μ) distribution. What is the long-run fraction of customers who are lost?

Taking the constant interarrival time as time unit, the fraction of lost customers can be seen as an average cost per time unit when a cost of 1 is incurred each time an arriving customer finds the waiting room full. The queueing process embedded at the arrival epochs can be described by a Markov process by noting that the Erlang (r, μ) distributed service time can be seen as the sum of r independent phases each having an exponential distribution with mean $1/\mu$. A customer is served by serving its phases one at a time. The queueing problem can now be converted into a Markov decision model with a single action in each state. The state of the system is observed at the arrival epochs and the set of possible states of the system is given by

$$I = \{0, 1, \ldots, Kr\}.$$

State i corresponds to the situation that i uncompleted service phases are present just prior to the arrival of a new customer. In each state i there is a single action to be denoted by $a = 0$. The action $a = 0$ in state i corresponds to the acceptance of the newly arriving customer when $i \leq Kr - r$ and corresponds to the rejection of the customer otherwise. The one-step costs $c_i(a)$ are given by

$$c_i(a) = \begin{cases} 0 & \text{if } i \leq Kr - r, \\ 1 & \text{if } i > Kr - r. \end{cases}$$

Denote by $a_\ell = e^{-\mu D}(\mu D)^\ell/\ell!$ the probability of the completion of ℓ service phases during an interarrival time D when the server is continuously busy. Then the recursive value-iteration equation (6.6.1) becomes

$$V_n(i) = \sum_{\ell=0}^{i+r-1} a_\ell V_{n-1}(i+r-\ell) + \left(1 - \sum_{\ell=0}^{i+r-1} a_\ell\right) V_{n-1}(0), \quad 0 \le i \le Kr - r$$

$$V_n(i) = 1 + \sum_{\ell=0}^{i-1} a_\ell V_{n-1}(i-\ell) + \left(1 - \sum_{\ell=0}^{i-1} a_\ell\right) V_{n-1}(0), \quad Kr - r < i \le Kr.$$

The discrete-time Markov chain describing the number of service phases present at the arrival epochs is aperiodic. Hence the lower and upper bounds m_n and M_n from the value-iteration algorithm both converge to the long-run fraction of customers who are lost.

6.7 CONVERGENCE PROOFS

In this section we give convergence proofs for the policy-iteration algorithm and the value-iteration algorithm. The finite convergence of the policy-iteration algorithm is proved for the unichain case. For the standard value-iteration algorithm the convergence of the bounds m_n and M_n to the same limit is proved under the unichain assumption together with the assumption that the one-step transition probability $p_{ii}(a) > 0$ for all $i \in I$ and $a \in A(i)$. The latter aperiodicity assumption is automatically satisfied when the data transformation discussed in Section 6.6 is applied.

Convergence proof for policy iteration

We first establish a lexicographical ordering for the average cost and the relative values associated with the policies that are generated by the algorithm. For that purpose we need to standardize the relative value functions since a relative value function is not uniquely determined. Let us number or renumber the possible states as $i = 1, \ldots, N$. In view of the fact that the relative values of a given policy are unique up to an additive constant, the sequence of policies generated by the algorithm does not depend on the particular choice of the relative value function for a given policy. For each stationary policy Q, we now consider the particular relative value function $w_i(Q)$ defined by (6.3.1), where the regeneration state r is chosen as the *largest* state in $I(Q)$. The set $I(Q)$ is defined by

$$I(Q) = \text{the set of states that are recurrent under policy } Q.$$

Let R and $\overline{R}$ be immediate successors in the sequence of policies generated by the algorithm. Suppose that $\overline{R} \neq R$. We assert that either

(a) $g(\overline{R}) < g(R)$, or

(b) $g(\overline{R}) = g(R)$ and $w_i(\overline{R}) \leq w_i(R)$ for all $i \in I$ with strict inequality for at least one state i.

That is, each iteration either reduces the cost rate or else reduces the relative value of a (transient) state. Since the number of possible stationary policies is finite, this assertion implies that the algorithm converges after finitely many iterations. To prove the assertion, the starting point is the relation

$$c_i(\overline{R}_i) - g(R) + \sum_{j \in I} p_{ij}(\overline{R}_i) w_j(R) \leq w_i(R), \quad i \in I, \tag{6.7.1}$$

with strict inequality only for those states i with $\overline{R}_i \neq R_i$. This relation is an immediate consequence of the construction of policy $\overline{R}$. By Theorem 6.2.1 and (6.7.1), we have $g(\overline{R}) \leq g(R)$. The strict inequality $g(\overline{R}) < g(R)$ holds only if the strict inequality holds in (6.7.1) for some state i that is recurrent under the new policy $\overline{R}$.

Consider now the case of $g(\overline{R}) = g(R)$. Then it is true that the equality sign holds in (6.7.1) for all $i \in I(\overline{R})$. Thus, by the convention made in the policy-improvement step,

$$\overline{R}_i = R_i, \quad i \in I(\overline{R}). \tag{6.7.2}$$

This implies that

$$I(R) = I(\overline{R}), \tag{6.7.3}$$

since the set $I(\overline{R})$ is closed under policy $\overline{R}$ and any two states in $I(\overline{R})$ communicate under policy $\overline{R}$. In its turn (6.7.3) implies that

$$w_j(R) = w_j(\overline{R}), \quad j \in I(\overline{R}). \tag{6.7.4}$$

This can be seen as follows. From the definition (6.3.1) of the relative values and the fact that the set of recurrent states is a closed set, it follows that for any policy Q the relative values for the recurrent states $i \in I(Q)$ do not depend on the actions in the transient states $i \notin I(Q)$. In view of the convention to take the largest state in $I(Q)$ as the reference state for the definition of the relative value function $w_i(Q)$, it follows from (6.7.2) and (6.7.3) that (6.7.4) holds. The remainder of the proof is now easy. Proceeding in the same way as in the derivation of (6.3.3), we find by iterating the inequality (6.7.1) that

$$\begin{aligned} w_i(R) &\geq c_i(\overline{R}_i) - g(R) + \sum_{j \in I} p_{ij}(\overline{R}_i) w_j(R) \qquad (6.7.5) \\ &\geq V_m(i, \overline{R}) - mg(R) + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) w_j(R), \quad i \in I \text{ and } m \geq 1, \end{aligned}$$

where the strict inequality sign holds in the first inequality for each i with $\overline{R}_i \neq R_i$. By (6.3.5) with R replaced by $\overline{R}$ and the fact that $g(\overline{R}) = g(R)$, we have for any $m \geq 1$ that

$$w_i(\overline{R}) = V_m(i, \overline{R}) - mg(R) + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) w_j(\overline{R}), \quad i \in I.$$

Replacing $w_j(\overline{R})$ by $w_j(R) - \{w_j(R) - w_j(\overline{R})\}$, we next find that

$$\begin{aligned} V_m(i, \overline{R}) - mg(R) + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) w_j(R) \\ = w_i(\overline{R}) + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) \{w_j(R) - w_j(\overline{R})\}, \quad i \in I \text{ and } m \geq 1. \end{aligned}$$

Hence (6.7.5) can be rewritten as

$$\begin{aligned} w_i(R) \geq c_i(\overline{R}_i) - g(R) + \sum_{j \in I} p_{ij}(\overline{R}_i) w_j(R) \\ \geq w_i(\overline{R}) + \sum_{j \in I} p_{ij}^{(m)}(\overline{R}) \{w_j(R) - w_j(\overline{R})\}, \quad i \in I \text{ and } m \geq 1, \end{aligned}$$

where the strict inequality sign holds in the first inequality for each i with $R_i \neq \overline{R}_i$. Using (6.7.4) and noting that $p_{ij}^{(m)}(\overline{R}) \to 0$ as $m \to \infty$ for j transient under $\overline{R}$, it follows that $w_i(R) \geq w_i(\overline{R})$ for all $i \in I$ with strict inequality for each i with $R_i \neq \overline{R}_i$. This completes the proof.

Convergence proof for value iteration

The proof of Theorem 6.6.2 is only given for the special case that the following assumption is satisfied.

Strong aperiodicity assumption *(i) for each stationary policy R the associated Markov chain $\{X_n\}$ has no two disjoint closed sets;*
(ii) $p_{ii}(a) > 0$ for all $i \in I$ and $a \in A(i)$.

Note that assumption (ii) automatically holds when the data transformation from Section 6.6 is applied to the original model.

We first establish an important lemma about the chain structure of the product of Markov matrices associated with the stationary policies. In this lemma the notation $P(f)$ is used for the stochastic matrix $(p_{ij}(f(i)))$, $i, j \in I$ associated with the stationary policy f. The (i, j)th element of the matrix product PQ is denoted by $(PQ)_{ij}$.

Lemma 6.7.1 *Suppose that the strong aperiodicity assumption holds. Let N be the number of states of the Markov decision model. Then, for any two N-tuples*

$(f_N, \ldots, f_1)$ and $(g_N, \ldots, g_1)$ of stationary policies and for any two states r and *s, there is some state j such that*

$$[P(f_N)\cdots P(f_1)]_{rj} > 0 \quad \textit{and} \quad [P(g_N)\cdots P(g_1)]_{sj} > 0. \tag{6.7.6}$$

Proof Define for $k = 1, \ldots, N$ the sets $S(k)$ and $T(k)$ by

$$S(k) = \{j \in I \mid [P(f_k)\cdots P(f_1)]_{rj} > 0\},$$
$$T(k) = \{j \in I \mid [P(g_k)\cdots P(g_1)]_{sj} > 0\}.$$

Since $p_{ii}(a) > 0$ for all $j \in I$ and $a \in A(i)$, we have

$$S(k+1) \supseteq S(k) \quad \text{and} \quad T(k+1) \supseteq T(k), \quad k = 1, \ldots, N-1.$$

Assume now to the contrary that (6.7.6) does not hold. Then $S(N) \cap T(N)$ is empty. In other words, $S(N)$ and $T(N)$ are disjoint sets with $S(N) \neq I$ and $T(N) \neq I$. Thus, since the sets $S(k)$ and $T(k)$ are non-decreasing, there are integers v and w with $1 \leq v, w < N$ such that $S(v) = S(v+1)$ and $T(w) = T(w+1)$. This implies that the set $S(v)$ of states is closed under policy f_{v+1} and the set $T(w)$ of states is closed under policy g_{w+1}. Since the sets $S(N)$ and $T(N)$ are disjoint and $S(N) \supseteq S(v)$ and $T(N) \supseteq T(w)$, we have that the sets $S(v)$ and $T(w)$ are disjoint. Construct now a stationary policy R with $R_i = f_{v+1}(i)$ for $i \in S(v)$ and $R_i = g_{w+1}(i)$ for $i \in T(w)$. Then policy R has the two disjoint closed sets $S(v)$ and $T(w)$. This contradicts the first part of the strong aperiodicity assumption. Hence the result (6.7.6) must hold.

Proof of Theorem 6.6.2 (under the strong aperiodicity assumption) We first introduce some notation. Let $R(n)$ be any stationary policy for which the action $R_i(n)$ minimizes the right-hand side of the value-iteration equation (6.6.1) for all $i \in I$. Denote by P_n the stochastic matrix whose (i, j)th element equals $p_{ij}(R_i(n))$ and define the vector V_n by $V_n = (V_n(i),\ i \in I)$. By the proof of Theorem 6.6.3,

$$V_n - V_{n-1} \leq P_{n-1}(V_{n-1} - V_{n-2}) \quad \text{and} \quad V_n - V_{n-1} \geq P_n(V_{n-1} - V_{n-2}). \tag{6.6.7}$$

Fix $n \geq 2$. Since $M_n = V_n(i_1) - V_{n-1}(i_1)$ and $m_n = V_n(i_2) - V_{n-1}(i_2)$ for some states i_1 and i_2, we find

$$M_n - m_n \leq [P_{n-1}(V_{n-1} - V_{n-2})](i_1) - [P_n(V_{n-1} - V_{n-2})](i_2).$$

Applying repeatedly the inequalities (6.6.7), we find for any $1 \leq k < n$

$$M_n - m_n \leq [P_{n-1}P_{n-2}\cdots P_{n-k}(V_{n-k} - V_{n-k-1})](i_1) - [P_nP_{n-1}\cdots P_{n-k+1}(V_{n-k} - V_{n-k-1})](i_2). \tag{6.6.8}$$

The remainder of the proof uses the same ideas as in the proof of Theorem 3.5.12. Fix $n > N$ and choose $k = N$ in (6.6.8), where N is the number of states. This yields

$$M_n - m_n \leq \sum_{j \in I} d_j \{V_{n-N}(j) - V_{n-N-1}(j)\},$$

where d_j is a shorthand notation for

$$d_j = [P_{n-1} \cdots P_{n-N}]_{i_1 j} - [P_n \cdots P_{n-N+1}]_{i_2 j}.$$

Write $d^+ = \max(d, 0)$ and $d^- = -\min(d, 0)$. Then $d = d^+ - d^-$ and $d^+, d^- \geq 0$. Thus

$$\begin{aligned} M_n - m_n &\leq \sum_{j \in I} d_j^+ \{V_{n-N}(j) - V_{n-N-1}(j)\} - \sum_{j \in I} d_j^- \{V_{n-N}(j) - V_{n-N-1}(j)\} \\ &\leq M_{n-N} \sum_{j \in I} d_j^+ - m_{n-N} \sum_{j \in I} d_j^- = (M_{n-N} - m_{n-N}) \sum_{j \in I} d_j^+, \end{aligned}$$

by $\sum_j d_j^+ = \sum_j d_j^-$. This identity is a consequence of $\sum_j d_j = 0$. Next use the relation $(p - q)^+ = p - \min(p, q)$ to conclude that

$$\begin{aligned} M_n - m_n \leq (M_{n-N} - m_{n-N}) \\ \times \left[1 - \sum_{j \in I} \min([P_{n-1} \cdots P_{n-N}]_{i_1 j}, [P_n \cdots P_{n-N+1}]_{i_2 j} \right]. \end{aligned}$$

Now we invoke Lemma 6.7.1. Since the number of states and the number of stationary policies are both finite, there is a *positive* number ρ such that

$$\sum_{j \in I} \min\{[P(f_N) \cdots P(f_1)]_{rj}, \ [P(g_N) \cdots P(g_1)]_{sj}\} \geq \rho$$

for any two N-tuples $(f_N, \ldots, f_1)$ and $(g_N, \ldots, g_1)$ of stationary policies and for any two states r and s. Thus

$$M_n - m_n \leq (1 - \rho)(M_{n-N} - m_{n-N}).$$

In Theorem 6.6.3 it was shown that $\{M_n - m_n, n \geq 1\}$ is non-increasing. Thus we find that

$$M_n - m_n \leq (1 - \rho)^{[n/N]}(M_0 - m_0), \quad n \geq 1,$$

implying the desired result.

EXERCISES

6.1 Consider a periodic review production/inventory problem where the demands for a single product in the successive weeks are independent random variables with a common discrete probability distribution $\{\phi(j),\ j = 0, \ldots, r\}$. Any demand in excess of on-hand inventory is lost. At the beginning of each week it has to be decided whether or not to start a production run. The lot size of each production run consists of a fixed number of Q units. The production lead time is one week so that a batch delivery of the entire lot occurs at the beginning of the next week. Due to capacity restrictions on the inventory, a production run is never started when the on-hand inventory is greater than M. The following costs are involved. A fixed set-up cost of $K > 0$ is incurred for a new production run started after the production facility has been idle for some time. The holding costs incurred during a week are proportional to the on-hand inventory at the end of that week, where $h > 0$ is the proportionality constant. A fixed lost-sales cost of $p > 0$ is incurred for each unit of excess demand. Formulate the problem of finding an average cost optimal production rule as a Markov decision problem.

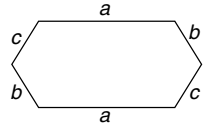
6.2 A piece of electronic equipment having two identical devices is inspected at the beginning of each day. Redundancy has been built into the system so that the system is still operating if only one device works. The system goes down when both devices are no longer working. The failure rate of a device depends both on its age and on the condition of the other device. A device in use for m days will fail on the next day with probability $r_1(m)$ when the other device is currently being overhauled and with probability $r_2(m)$ otherwise. It is assumed that both $r_1(m)$ and $r_2(m)$ are equal to 1 when m is sufficiently large. A device that is found in the failure state upon inspection has to be overhauled. An overhaul of a failed device takes T_0 days. Also a preventive overhaul of a working device is possible. Such an overhaul takes T_1 days. It is assumed that $1 \leq T_1 < T_0$. At each inspection it has to be decided to overhaul one or both of the devices, or let them continue working through the next day. The goal is to minimize the long-run fraction of time the system is down. Formulate this problem as a Markov decision problem. (*Hint:* define the states (i, j), $(i, -k)$ and $(-h-k)$. The first state means that both devices are working for i and j days respectively, the second state means that one device is working for i days and the other is being overhauled with a remaining overhaul time of k days, and the third state means that both devices are being overhauled with remaining overhaul times of h and k days.)

6.3 Two furnaces in a steelworks are used to produce pig iron for working up elsewhere in the factory. Each furnace needs overhauling from time to time because of failure during operation or to prevent such a failure. Assuming an appropriately chosen time unit, an overhaul of a furnace always takes a fixed number of L periods. The overhaul facility is capable of overhauling both furnaces simultaneously. A furnace just overhauled will operate successfully during i periods with probability q_i, $1 \leq i \leq M$. If a furnace has failed, it must be overhauled; otherwise, there is an option of either a preventive overhaul or letting the furnace operate for the next period. Since other parts of the steelworks are affected when not all furnaces are in action, a loss of revenue of $c(j)$ is incurred for each period during which j furnaces are out of action, $j = 1, 2$. No cost is incurred if both furnaces are working. Formulate the problem of finding an average cost optimal overhauling policy as a Markov decision problem. This problem is based on Stengos and Thomas (1980).

6.4 A factory has a tank for temporarily storing chemical waste. The tank has a capacity of 4 m^3. Each week the factory produces k m^3 of chemical waste with probability p_k for $k = 0, \ldots, 3$ with $p_0 = 1/8$, $p_1 = 1/2$, $p_2 = 1/4$ and $p_3 = 1/8$. If the amount of waste produced exceeds the remaining capacity of the tank, the excess is specially handled at a cost of \$30 per cubic metre. At the end of the week a decision has to be made as to whether or not to empty the tank. There is a fixed cost of \$25 to empty the tank and a variable cost

of \$5 for each cubic metre of chemical waste that is removed. Compute an average cost optimal policy by policy iteration or linear programming.

6.5 A stamping machine produces six-cornered plates of the illustrated form.



The machine has three pairs of adjustable knives. In the diagram these pairs are denoted by a, b and c. Each pair of knives can fall from the correct position during the stamping of a plate. The following five situations can occur: (1) all three pairs have the correct position, (2) only pairs b and c have the correct position, (3) only pair b has the correct position, (4) only pair c has the correct position and (5) no pair has the correct position. The probabilities q_{ij} that during a stamping a change from situation i to situation j occurs are given by

$$(q_{ij}) = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

After each stamping it is possible to adjust the machine such that all pairs of knives have the correct position again. The following costs are involved. The cost of bringing all pairs of knives into the correct position is 10. Each plate produced when j pairs of knives have the wrong position involves an adjustment cost of $4j$. Compute a maintenance rule that minimizes the average cost per stamping by policy iteration or linear programming.

6.6 An electricity plant has two generators $j = 1$ and 2 for generating electricity. The required amount of electricity fluctuates during the day. The 24 hours in a day are divided into six consecutive periods of 4 hours each. The amount of electricity required in period k is d_k kWh for $k = 1, \ldots, 6$. Also the generator j has a capacity of generating c_j kWh of electricity per period of 4 hours for $j = 1, 2$. An excess of electricity produced during one period cannot be used for the next period. At the beginning of each period k it has to be decided which generators to use for that period. The following costs are involved. An operating cost of r_j is incurred for each period in which generator j is used. Also, a set-up cost of S_j is incurred each time generator j is turned on after having been idle for some time. Develop a policy-iteration algorithm that exploits the fact that the state transitions are deterministic. Solve for the numerical data $d_1 = 20$, $d_2 = 40$, $d_3 = 60$, $d_4 = 90$, $d_5 = 70$, $d_6 = 30$, $c_1 = 40$, $c_2 = 60$, $r_1 = 1000$, $r_2 = 1100$, $S_1 = 500$ and $S_2 = 300$.

6.7 Every week a repairman travels to customers in five towns on the successive working days of the week. The repairman visits Amsterdam (town 1) on Monday, Rotterdam (town 2) on Tuesday, Brussels (town 3) on Wednesday, Aachen (town 4) on Thursday and Arnhem (town 5) on Friday. In the various towns it may be necessary to replace a certain crucial element in a piece of electronic equipment rented by customers. The probability distribution of the number of replacements required at a visit to town j is given by $\{p_j(k), k \geq 0\}$ for $j = 1, \ldots, 5$. The numbers of required replacements on the successive days are independent of each other. The repairman is able to carry M spare parts. If the number of spare parts the repairman carries is not enough to satisfy the demand in a town, another repairman has to be sent the next day to that town to complete the remaining replacements. The cost of such a special mission to town j is K_j. At the end of each day the repairman may decide to send for a replenishment of the spare parts to the town where the repairman is. The cost of sending such a replenishment to town j is a_j. Develop a value-iteration algorithm for the

computation of an average cost optimal policy and indicate how to formulate converging lower and upper bounds on the minimal costs. Solve for the numerical data $M = 5$, $K_j = 200$ for all j, $a_1 = 60$, $a_2 = 30$, $a_3 = 50$, $a_4 = 25$, $a_5 = 100$, where the probabilities $p_j(k)$ are given in the following table.

$k \backslash j$	1	2	3	4	5
0	0.5	0.25	0.375	0.3	0.5
1	0.3	0.5	0.375	0.5	0.25
2	0.2	0.25	0.25	0.2	0.25

6.8 The slotted ALOHA system is a much used random access protocol in packet communication systems where the time is slotted in intervals of fixed lengths and a transmission of a packet can only be started at the beginning of a time slot. There are N terminals. At the beginning of each time slot, each terminal emits a packet with a certain probability. The terminals act independently of each other in trying to use the transmission channel for sending a packet. If more than one terminal sends a packet in the same time slot, a collision occurs and all transmissions attempted in that time slot are unsuccessful. A successful transmission returns the terminal to its originating mode, whereas an unsuccessful attempt puts it temporarily in retransmission mode. There is a given probability p that a terminal in originating mode attempts to transmit a packet at the beginning of a time slot. This probability is beyond control. However, the probability at which a terminal in retransmission mode is allowed to retransmit its packet at the beginning of a time slot can be controlled. The control rule gives each terminal in retransmission mode permission to retransmit with the same probability. In other words, a control rule is specified by probabilities $\{r_1, \ldots, r_N\}$, where r_n is the permission probability when n terminals are in retransmission mode. Develop a policy-iteration algorithm to compute an optimal control rule when the criterion is to maximize the average throughput per time slot. Also compare the maximal average throughput with the average throughput of the so-called TSO policy, where r_n is chosen as $[1-(N-n+1)p]/(n-Np)$ when $0 < Np < 1$ and r_n is chosen as $1/n$ otherwise. Solve for the numerical data ($N = 15$, $p = 0.05$) and ($N = 25$, $p = 0.05$). (*Hint*: the choice of one-step costs $c_i(a)$ simplifies by noting that maximizing the average throughput is equivalent to minimizing the average number of terminals in retransmission mode at the beginning of a time slot.)

6.9 A motorist has a vehicle insurance which charges reduced premiums when no claims are made over one or more years. When an accident occurs the motorist has the option of either making a claim and thereby perhaps losing a reduction in premium, or paying the costs associated with the accident himself. The premium payment is due at the beginning of each year and the payment depends only on the previous payment and the number of claims made in the past year. There are five possible premiums $\pi(1) = 500$, $\pi(2) = 375$, $\pi(3) = 300$, $\pi(4) = 250$, $\pi(5) = 200$. The premium structure is as shown in the table below. In any given month the motorist will have an accident with a probability of $\lambda = \frac{1}{24}$ and no accident with a probability of $1 - \lambda$. The costs associated with any accident have a lognormal distribution with mean 500 and a squared coefficient of variation of 4.

	Subsequent premium		
Current premium	No claim	One claim	Two or more claims
$\pi(1)$	$\pi(2)$	$\pi(1)$	$\pi(1)$
$\pi(2)$	$\pi(3)$	$\pi(1)$	$\pi(1)$
$\pi(3)$	$\pi(4)$	$\pi(1)$	$\pi(1)$
$\pi(4)$	$\pi(5)$	$\pi(2)$	$\pi(1)$
$\pi(5)$	$\pi(5)$	$\pi(3)$	$\pi(1)$

Develop a value-iteration algorithm to compute an average cost optimal claim rule. (*Hint*: take the beginning of each month as the decision epochs and let the action $a = d$ mean that damage in the coming month will be claimed only if this damage exceeds the level d. Define the state of the system as (t, i) with $t = 0, \ldots, 12$ and $i = 1, \ldots, 6$, where t denotes the number of months until the next premium payment and the indicator variable i refers to the status of the next premium payment. Explain why you need no data transformation to handle periodicities but you can use the bounds $m_n = \min_i\{V_{12n}(0, i) - V_{12(n-1)}(0, i)\}$ and $M_n = \max_i\{V_{12n}(0, i) - V_{12(n-1)}(0, i)\}$, where $V_{12n+t}(t, i)$ is defined as the minimal total expected cost if the motorist still has an insurance contract for $t + 12n$ months and the present state is (t, i).)

6.10 The stock level of a given product is reviewed at the beginning of each week. Upon review a decision has to be made whether or not to replenish the stock level. The stock can be replenished up to level M, the maximum amount that can be held in stock. The lead time of a replenishment order is negligible. The demands for the product in the successive weeks are independent random variables having a Poisson distribution with a given mean μ. Any demand occurring when the system is out of stock is lost. The following costs are involved. For each replenishment order there is a fixed set-up cost of $K > 0$ and a variable ordering cost of $c \geq 0$ for each unit ordered. In each week a holding cost of $h > 0$ is charged against each unit in stock at the end of the week. A penalty cost of $p > 0$ is incurred for each unit of demand that is lost. The problem is to find a stock control rule minimizing the long-run average cost per week.

(a) Use value iteration to solve for the numerical data $M = 100$, $\mu = 25$, $K = 64$, $c = 0$, $h = 1$ and $p = 5$. Also try other numerical examples and verify experimentally that the optimal control rule is always of the (s, S) type when the maximum stock level M is sufficiently large. Under an (s, S) policy with $s \leq S$ the inventory position is ordered up to the level S when at a review the inventory position is below the reorder point s; otherwise, no ordering is done. Using the flexibility in the policy-improvement procedure, Federgruen and Zipkin (1984) developed a tailor-made policy-iteration algorithm that generates a sequence of improved policies within the class of (s, S) policies.

(b) Suppose the probabilistic constraint is imposed that the long-run fraction of demand lost should not exceed $1 - \beta$ for a prespecified service level β (note that this fraction equals the average demand lost per week divided by μ). Use linear programming to find an optimal control minimizing the long-run average cost per week subject to this service level constraint. Solve for the numerical data $\beta = 0.99$, $M = 100$, $\mu = 25$, $K = 64$, $c = 0$, $h = 1$ and $p = 0$. Also compare the average cost and the service level of the optimal randomized policy with the average cost and the service level of the best stationary policy obtained by the Lagrange-multiplier approach.

BIBLIOGRAPHIC NOTES

The policy-iteration method for the discrete-time Markov decision model was developed in Howard (1960). A theoretical foundation to Howard's policy-iteration method was given in Blackwell (1962); see also Denardo and Fox (1968) and Veinott (1966). Linear programming formulations for the Markov decision model were first given by De Ghellinck (1960) and Manne (1960) and streamlined later by Denardo and Fox (1968), Derman (1970) and Hordijk and Kallenberg (1979, 1984). The computational usefulness of the value-iteration algorithm was greatly enlarged by Odoni (1969) and Hastings (1971), who introduced lower and upper

bounds on the minimal average cost and on the average cost of the policies generated by the algorithm. These authors extended the original value-iteration bounds of MacQueen (1966) for the discounted cost case to the average cost case. The modified value-iteration algorithm with a dynamic relaxation factor comes from Popyack *et al*. (1979). The first proof of the geometric convergence of the undiscounted value-iteration algorithm was given by White (1963) under a very strong recurrence condition. The proof in Section 6.7 is along the same lines as the proof given in Van der Wal (1980). General proofs for the geometric convergence of value-iteration can be found in the papers of Bather (1973) and Schweitzer and Federgruen (1979). These papers demonstrate the deepness and the beauty of the mathematics underlying the average cost criterion. In general there is a rich mathematical theory behind the Markov decision model. A good account of this theory can be found in the books of Hernandez-Lerma and Lasserre (1996), Puterman (1994) and Sennott (1999). A recommended reference for constrained Markov decision processes is the book of Altman (1999).

The Markov decision model finds applications in a wide variety of fields. Golabi *et al*. (1982), Kawai (1983), Stengos and Thomas (1980) and Tijms and Van der Duyn Schouten (1985) give applications to replacement and maintenance problems. Norman and Shearn (1980) and Kolderman and Volgenant (1985) discuss applications to insurance and Su and Deininger (1972) give an application to water-resource control. Applications to control problems in telecommunication are mentioned in the next chapter. A survey of real applications of Markov decision models can be found in White (1985).

REFERENCES

Altman, E. (1999) *Constrained Markov Decision Processes*. Chapman and Hall, London.

Bather, J. (1973) Optimal decision procedures for finite Markov chains. *Adv. Appl. Prob.*, **5**, 521–540.

Bellman, R. (1957) *Dynamic Programming*. Princeton University Press, Princeton NJ.

Blackwell, D. (1962) Discrete dynamic programming. *Ann. Math. Statist.*, **33**, 719–726.

De Ghellinck, G. (1960) Les problèmes de décisions séquentielles. *Cahiers Centre Etudes Recherche Opér.*, **2**, 161–179.

Denardo, E.V. (1982) *Dynamic Programming*. Prentice Hall, Englewood Cliffs NJ.

Denardo, E.V. and Fox, B.L. (1968) Multichain Markov renewal programs. *SIAM J. Appl. Math.*, **16**, 468–487.

Derman, C. (1970) *Finite State Markovian Decision Processes*. Academic Press, New York.

Federgruen, A. and Zipkin, P. (1984) An efficient algorithm for computing optimal *(s, S)* policies. *Operat. Res.*, **32**, 1268–1285.

Golabi, K., Kulkarni, R.B. and Way, C.B. (1982) A statewide pavement management system. *Interfaces*, **12**, no. 6, 5–21.

Hastings, N.A.J. (1971) Bounds on the gain of a Markov decision process. *Operat. Res.*, **19**, 240–244.

Hernandez-Lerma, O. and Lasserre, J.B. (1996) *Discrete-Time Markov Control Processes*. Springer-Verlag, Berlin.

Hordijk, A. and Kallenberg, L.C.M. (1979) Linear programming and Markov decision chains. *Management Sci.*, **25**, 352–362.

Hordijk, A. and Kallenberg, L.C.M. (1984) Constrained undiscounted stochastic dynamic programming. *Math. Operat. Res.*, **9**, 276–289.

Howard, R.A. (1960) *Dynamic Programming and Markov Processes*. John Wiley & Sons, Inc., New York.

Kawai, H. (1983) An optimal ordering and replacement policy of a Markovian degradation system under complete observation, part I. *J. Operat. Res. Soc. Japan*, **26**, 279–290.

Kolderman, J. and Volgenant, A. (1985) Optimal claiming in an automobile insurance system with bonus-malus structure. *J. Operat. Res. Soc.*, **36**, 239–247.

MacQueen, J. (1966) A modified dynamic programming method for Markovian decision problems. *J. Math. Appl. Math.*, **14**, 38–43.

Manne, A. (1960) Linear programming and sequential decisions. *Management Sci.*, **6**, 259–267.

Norman, J.M. and Shearn, D.C.S. (1980) Optimal claiming on vehicle insurance revisited. *J. Operat. Res. Soc.*, **31**, 181–186.

Odoni, A. (1969) On finding the maximal gain for Markov decision processes. *Operat. Res.*, **17**, 857–860.

Popyack, J.L., Brown, R.L. and White, C.C. III (1979) Discrete versions of an algorithm due to Varaiya. *IEEE Trans. Automat. Contr.*, **24**, 503–504.

Puterman, M.L. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York.

Schweitzer, P.J. and Federgruen, A. (1979) Geometric convergence of value iteration in multichain Markov decision problems. *Adv. Appl. Prob.*, **11**, 188–217.

Sennott, L.I. (1999) *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley & Sons, Inc., New York.

Stengos, D. and Thomas, L.C. (1980) The blast furnaces problem. *Eur. J. Operat. Res.*, **4**, 330–336.

Su, Y. and Deininger, R. (1972) Generalization of White's method of successive approximations to periodic Markovian decision processes. *Operat. Res.*, **20**, 318–326.

Tijms, H.C. and Van der Duyn Schouten, F.A. (1985) A Markov decision algorithm for optimal inspections and revisions in a maintenance system with partial information. *Eur. J. Operat. Res.*, **21**, 245–253.

Van der Wal, J. (1980) The method of value oriented successive approximations for the average reward Markov decision process. *OR Spektrum*, **1**, 233–242.

Veinott, A.F. Jr (1966) On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.*, **37**, 1284–1294.

Wagner, H.M. (1975) *Principles of Operations Research*, 2nd edn. Prentice Hall, Englewood Cliffs NJ.

White, D.J. (1963) Dynamic programming, Markov chains and the method of successive approximations. *J. Math. Anal. Appl.*, **6**, 373–376.

White, D.J. (1985) Real applications of Markov decision processes. *Interfaces*, **15**, no. 6, 73–78.

CHAPTER 7

Semi-Markov Decision Processes

7.0 INTRODUCTION

The previous chapter dealt with the discrete-time Markov decision model. In this model, decisions can be made only at fixed epochs $t = 0, 1, \ldots$. However, in many stochastic control problems the times between the decision epochs are not constant but random. A possible tool for analysing such problems is the semi-Markov decision model. In Section 7.1 we discuss the basic elements of this model. Also, for the optimality criterion of the long-run average cost per time unit, we give a data-transformation method by which the semi-Markov decision model can be converted into an equivalent discrete-time Markov decision model. The data-transformation method enables us to apply the recursive method of value-iteration to the semi-Markov decision model. Section 7.2 summarizes various algorithms for the computation of an average cost optimal policy.

In Section 7.3 we discuss the value-iteration algorithm for a semi-Markov decision model in which the times between the decision epochs are exponentially distributed. For this particular case the computational effort of the value-iteration algorithm can considerably be reduced by introducing fictitious decision epochs. This simple trick creates sparse transition matrices leading to a much more effective value-iteration algorithm. Section 7.4 illustrates how value iteration in combination with an embedding idea can be used in the optimization of queues. The semi-Markov decision model is a very useful tool for optimal control in queueing systems. In Section 7.5 we will exploit a remarkable feature of the policy-iteration algorithm, namely that the algorithm typically achieves its largest improvements in costs in the first few iterations. This finding is sometimes useful to attack the curse of dimensionality in applications with a multidimensional state space. The idea is to determine first the relative values for a reasonable starting policy and to apply next a single policy-improvement step. This heuristic approach will be illustrated to a dynamic routing problem.

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

7.1 THE SEMI-MARKOV DECISION MODEL

Consider a dynamic system whose state is reviewed at random epochs. At those epochs a decision has to be made and costs are incurred as a consequence of the decision made. The set of possible states is denoted by I. For each state $i \in I$, a set $A(i)$ of possible actions is available. It is assumed that the state space I and the action sets $A(i)$, $i \in I$ are finite. This controlled dynamic system is called a *semi-Markov decision process* when the following Markovian properties are satisfied: if at a decision epoch the action a is chosen in state i, then the time until the next decision epoch and the state at that epoch depend only on the present state i and the subsequently chosen action a and are thus independent of the past history of the system. Also, the costs incurred until the next decision epoch depend only on the present state and the action chosen in that state. We note that in specific problems the state occurring at the next transition will often depend on the time until that transition. Also, the costs usually consist of lump costs incurred at the decision epochs and rate costs incurred continuously in time. As an example, consider a single-product inventory system in which the demand process is described by a Poisson process and the inventory position can be replenished at any time. In this example the decision epochs are the demand epochs and they occur randomly in time. The decision is whether or not to raise the inventory position after a demand has occurred. The costs typically consist of fixed replenishment costs and holding costs that are incurred continuously in time.

The long-run average cost per time unit

The long-run average cost per time unit is taken as the optimality criterion. For this criterion the semi-Markov decision model is in fact determined by the following characteristics:

$p_{ij}(a)$ = the probability that at the next decision epoch the system will be in state j if action a is chosen in the present state i,

$\tau_i(a)$ = the expected time until the next decision epoch if action a is chosen in the present state i,

$c_i(a)$ = the expected costs incurred until the next decision epoch if action a is chosen in the present state i.

It is assumed that $\tau_i(a) > 0$ for all $i \in I$ and $a \in A(i)$. As before, a stationary policy R is a rule which adds to each state i a single action $R_i \in A(i)$ and always prescribes to take this action whenever the system is observed in state i at a decision epoch. Since the state space is finite, it can be shown that under each stationary policy the number of decisions made in any finite time interval is finite with probability 1. We omit the proof of this result. Let

$$X_n = \text{the state of the system at the } n\text{th decision epoch.}$$

Then it follows that under a stationary policy R the embedded stochastic process $\{X_n\}$ is a discrete-time Markov chain with one-step transition probabilities $p_{ij}(R_i)$.

Define the random variable $Z(t)$ by

$$Z(t) = \text{the total costs incurred up to time } t, \qquad t \geq 0.$$

Fix now a stationary policy R. Denote by $E_{i,R}$ the expectation operator when the initial state $X_0 = i$ and policy R is used. Then the limit

$$g_i(R) = \lim_{t\to\infty} \frac{1}{t} E_{i,R}[Z(t)]$$

exists for all $i \in I$. This result can be proved by using the renewal-reward theorem in Section 2.2. The details are omitted. Just as in the discrete-time model, we can give a stronger interpretation to the average cost $g_i(R)$. If the initial state i is recurrent under policy R, then the long-run *actual* average cost per time unit equals $g_i(R)$ with probability 1. If the Markov chain $\{X_n\}$ associated with policy R has no two disjoint closed sets, the average cost $g_i(R)$ does not depend on the initial state $X_0 = i$.

Theorem 7.1.1 *Suppose that the embedded Markov chain $\{X_n\}$ associated with policy R has no two disjoint closed sets. Then*

$$\lim_{t\to\infty} \frac{Z(t)}{t} = g(R) \quad \textit{with probability } 1 \tag{7.1.1}$$

for each initial state $X_0 = i$, where the constant $g(R)$ is given by

$$g(R) = \sum_{j\in I} c_j(R_j)\pi_j(R) \Big/ \sum_{j\in I} \tau_j(R_j)\pi_j(R)$$

with $\{\pi_j(R)\}$ denoting the equilibrium distribution of the Markov chain $\{X_n\}$.

Proof We give only a sketch of the proof of (7.1.1). The key to the proof of (7.1.1) is that

$$\lim_{t\to\infty} \frac{Z(t)}{t} = \lim_{m\to\infty} \frac{E(\text{costs over the first } m \text{ decision epochs})}{E(\text{time over the first } m \text{ decision epochs})} \tag{7.1.2}$$

with probability 1. To verify this relation, fix a recurrent state r and suppose that $X_0 = r$. Let a cycle be defined as the time elapsed between two consecutive transitions into state r. By the renewal-reward theorem in Section 2.2,

$$\lim_{t\to\infty} \frac{Z(t)}{t} = \frac{E(\text{costs induring one cycle})}{E(\text{length of one cycle})}$$

with probability 1. By the expected-value version of the renewal-reward theorem,

$$\lim_{m\to\infty} \frac{1}{m} E(\text{costs over the first } m \text{ decision epochs}) = \frac{E(\text{costs incurred during one cycle})}{E(\text{number of transitions in one cycle})}$$

and

$$\lim_{m\to\infty}\frac{1}{m}E(\text{time over the first } m \text{ decision epochs}) = \frac{E(\text{length of one cycle})}{E(\text{number of transitions in one cycle})}.$$

Together the above three relations yield (7.1.2). The remainder of the proof is simple. Obviously, we have

$$E(\text{costs over the first } m \text{ decision epochs}) = \sum_{t=0}^{m-1}\sum_{j\in I} c_j(R_j)p_{rj}^{(t)}(R)$$

and

$$E(\text{time over the first } m \text{ decision epochs}) = \sum_{t=0}^{m-1}\sum_{j\in I} \tau_j(R_j)p_{rj}^{(t)}(R).$$

Dividing the numerator and the denominator of the right-hand side of (7.1.2) by m, letting $m \to \infty$ and using $\lim_{m\to\infty}(1/m)\sum_{t=0}^{m-1} p_{rj}^{(t)}(R) = \pi_j(R)$, the result (7.1.1) follows when the initial state $X_0 = r$. For initial state $X_0 = i$ the result next follows by mimicking the proof of Theorem 3.5.11 and noting that state r will be reached from state i with probability 1 after finitely many transitions.

A stationary policy R^* is said to be *average cost optimal* if $g_i(R^*) \le g_i(R)$ for all $i \in I$ and all stationary policies R. The algorithms for computing an average cost optimal policy in the discrete-time Markov decision model can be extended to the semi-Markov decision model. This will be done in the next section. However, before doing this, we discuss a data-transformation method that converts the semi-Markov decision model into a discrete-time Markov decision model such that for each stationary policy the average cost per time unit in the discrete-time Markov model is the same as in the semi-Markov model. This is a very useful result. The data-transformation method is an extension of the uniformization technique for continuous-time Markov chains discussed in Section 4.5.

The data-transformation method

First choose a number τ with

$$0 < \tau \le \min_{i,a} \tau_i(a).$$

Consider now the discrete-time Markov decision model whose basic elements are given by

$$\overline{I} = I, \quad \overline{A}(i) = A(i), \qquad i \in \overline{I},$$
$$\overline{c}_i(a) = c_i(a)/\tau_i(a), \qquad a \in \overline{A}(i) \text{ and } i \in \overline{I},$$

$$\overline{p}_{ij}(a) = \begin{cases} (\tau/\tau_i(a))p_{ij}(a), & j \neq i,\ a \in \overline{A}(i) \text{ and } i \in \overline{I}, \\ (\tau/\tau_i(a))p_{ij}(a) + [1 - (\tau/\tau_i(a))], & j = i,\ a \in \overline{A}(i) \text{ and } i \in \overline{I}. \end{cases}$$

This discrete-time Markov decision model has the same class of stationary policies as the original semi-Markov decision model. For each stationary policy R, let $\overline{g}_i(R)$ denote the long-run average cost per time unit in the discrete-time model when policy R is used and the initial state is i. Then it holds for each stationary policy R that

$$g_i(R) = \overline{g}_i(R), \qquad i \in I. \tag{7.1.3}$$

This result does not require any assumption about the chain structure of the Markov chains associated with the stationary policies. However, we prove the result (7.1.3) only for the unichain case. Fix a stationary policy R and assume that the embedded Markov chain $\{X_n\}$ in the semi-Markov model has no two disjoint closed sets. Denote by $\overline{X}_n$ the state at the nth decision epoch in the transformed discrete-time model. It is directly seen that the Markov chain $\{\overline{X}_n\}$ is also unichain under policy R. The equilibrium probabilities $\overline{\pi}_j(R)$ of the Markov chain $\{\overline{X}_n\}$ satisfy the equilibrium equations

$$\begin{aligned} \overline{\pi}_j(R) &= \sum_{i \in I} \overline{\pi}_i(R)\overline{p}_{ij}(R_i) \\ &= \sum_{i \in I} \overline{\pi}_i(R)\frac{\tau}{\tau_i(R_i)}p_{ij}(R_i) + \left[1 - \frac{\tau}{\tau_j(R_j)}\right]\overline{\pi}_j(R), \qquad j \in I. \end{aligned}$$

Hence, letting $u_j = \overline{\pi}_j(R)/\tau_j(R_j)$ and dividing by τ, we find that

$$u_j = \sum_{i \in I} u_i\, p_{ij}(R_i), \qquad j \in I.$$

These equations are precisely the equilibrium equations for the equilibrium probabilities $\pi_j(R)$ of the embedded Markov chain $\{X_n\}$ in the semi-Markov model. The equations determine the $\pi_j(R)$ uniquely up to a multiplicative constant. Thus, for some constant $\gamma > 0$,

$$\pi_j(R) = \gamma \frac{\overline{\pi}_j(R)}{\tau_j(R_j)}, \qquad j \in I.$$

Since $\sum_{j \in I} \overline{\pi}_j(R) = 1$, it follows that $\gamma = \sum_{j \in I} \tau_j(R_j)\pi_j(R)$. The desired result (7.1.3) now follows easily. We have

$$\begin{aligned} \overline{g}(R) &= \sum_{j \in I} \overline{c}_j(R_j)\overline{\pi}_j(R) = \frac{1}{\gamma}\sum_{j \in I} \frac{c_j(R_j)}{\tau_j(R_j)}\pi_j(R)\tau_j(R_j) \\ &= \sum_{j \in I} c_j(R_j)\pi_j(R) / \sum_{j \in I} \tau_j(R_j)\pi_j(R) \end{aligned}$$

and so, by Theorem 7.1.1, $\overline{g}(R) = g(R)$. Thus we can conclude that an average cost optimal policy in the semi-Markov model can be obtained by solving an appropriate discrete-time Markov decision model. This conclusion is particularly useful with respect to the value-iteration algorithm. In applying value iteration to the transformed model, it is no restriction to assume that for each stationary policy the associated Markov chain $\{\overline{X}_n\}$ is aperiodic. By choosing the constant τ strictly less than $\min_{i,a} \tau_i(a)$, we always have $p_{ii}(a) > 0$ for all i, a and thus the required aperiodicity.

7.2 ALGORITHMS FOR AN OPTIMAL POLICY

In this section we outline how the algorithms for the discrete-time Markov decision model can be extended to the semi-Markov decision model.

Policy-iteration algorithm

The policy-iteration algorithm will be described under the unichain assumption. This assumption requires that for each stationary policy the embedded Markov chain $\{X_n\}$ has no two disjoint closed sets. By data transformation, it is directly verified that the value-determination equations (6.3.2) for a given stationary policy R remain valid provided that we replace g by $g\tau_i(R_i)$. The policy-improvement procedure from Theorem 6.2.1 also remains valid when we replace g by $g\tau_i(\overline{R}_i)$. Suppose that $g(R)$ and $v_i(R)$, $i \in I$, are the average cost and the relative values of a stationary policy R. If a stationary policy $\overline{R}$ is constructed such that, for each state $i \in I$,

$$c_i(\overline{R}_i) - g(R)\tau_i(\overline{R}_i) + \sum_{j \in I} p_{ij}(\overline{R}_i) v_j(R) \leq v_i(R), \tag{7.2.1}$$

then $g(\overline{R}) \leq g(R)$. Moreover, $g(\overline{R}) < g(R)$ if the strict inequality sign holds in (7.2.1) for some state i which is recurrent under $\overline{R}$. These statements can be verified by the same arguments as used in the second part of the proof of Theorem 6.2.1.

Under the unichain assumption, we can now formulate the following policy-iteration algorithm:

Step 0 (initialization). Choose a stationary policy R.

Step 1 (value-determination step). For the current rule R, compute the average cost $g(R)$ and the relative values $v_i(R)$, $i \in I$, as the unique solution to the linear equations

$$v_i = c_i(R_i) - g\tau_i(R_i) + \sum_{j \in I} p_{ij}(R_i) v_j, \qquad i \in I,$$

$$v_s = 0,$$

where s is an arbitrarily chosen state.

Step 2 (policy-improvement step). For each state $i \in I$, determine an action a_i yielding the minimum in

$$\min_{a \in A(i)} \left\{ c_i(a) - g(R)\tau_i(a) + \sum_{j \in I} p_{ij}(a) v_j(R) \right\}.$$

The new stationary policy $\overline{R}$ is obtained by choosing $\overline{R}_i = a_i$ for all $i \in I$ with the convention that $\overline{R}_i$ is chosen equal to the old action R_i when this action minimizes the policy-improvement quantity.
Step 3 (convergence test). If the new policy $\overline{R} = R$, then the algorithm is stopped with policy R. Otherwise, go to step 1 with R replaced by $\overline{R}$.

In the same way as for the discrete-time Markov decision model, it can be shown that the algorithm converges in a finite number of iterations to an average cost optimal policy. Also, as a consequence of the convergence of the algorithm, there exist numbers g^* and v_i^* satisfying the average cost optimality equation

$$v_i^* = \min_{a \in A(i)} \left\{ c_i(a) - g^* \tau_i(a) + \sum_{j \in I} p_{ij}(a) v_j^* \right\}, \qquad i \in I. \tag{7.2.2}$$

The constant g^* is uniquely determined as the minimal average cost per time unit. Moreover, each stationary policy whose actions minimize the right-hand side of (7.2.2) for all $i \in I$ is average cost optimal. The proof of these statements is left as an exercise for the reader.

Value-iteration algorithm

For the semi-Markov decision model the formulation of a value-iteration algorithm is not straightforward. A recursion relation for the minimal expected costs over the first n decision epochs does not take into account the non-identical transition times and thus these costs cannot be related to the minimal average cost per time unit. However, by the data transformation method from Section 7.1, we can convert the semi-Markov decision model into a discrete-time Markov decision model such that both models have the same average cost for each stationary policy. A value-iteration algorithm for the original semi-Markov decision model is then implied by the value-iteration algorithm for the transformed discrete-time Markov decision model. In the discrete-time model it is no restriction to assume that all $\overline{c}_i(a) = c_i(a)/\tau_i(a)$ are positive; otherwise, add a sufficiently large positive constant to each $\overline{c}_i(a)$. The following recursion method results for the semi-Markov decision model:

Step 0. Choose $V_0(i)$ such that $0 \leq V_0(i) \leq \min_a \{c_i(a)/\tau_i(a)\}$ for all i. Choose a number τ with $0 < \tau \leq \min_{i,a} \tau_i(a)$. Let $n := 1$.

Step 1. Compute the function $V_n(i)$, $i \in I$, from

$$V_n(i) = \min_{a \in A(i)} \left[\frac{c_i(a)}{\tau_i(a)} + \frac{\tau}{\tau_i(a)} \sum_{j \in I} p_{ij}(a) V_{n-1}(j) + \left(1 - \frac{\tau}{\tau_i(a)}\right) V_{n-1}(i) \right]. \tag{7.2.3}$$

Let $R(n)$ be a stationary policy whose actions minimize the right-hand side of (7.2.3).

Step 2. Compute the bounds

$$m_n = \min_{j \in I} \{V_n(j) - V_{n-1}(j)\}, \qquad M_n = \max_{j \in I} \{V_n(j) - V_{n-1}(j)\}.$$

The algorithm is stopped with policy $R(n)$ when $0 \leq (M_n - m_n) \leq \varepsilon m_n$, where ε is a prespecified accuracy number. Otherwise, go to step 3.

Step 3. $n := n + 1$ and go to step 1.

Let us assume that the weak unichain assumption from Section 6.5 is satisfied for the embedded Markov chains $\{X_n\}$ associated with the stationary policies. It is no restriction to assume that the Markov chains $\{\overline{X}_n\}$ in the transformed model are aperiodic. Then the algorithm stops after finitely many iterations with a policy $R(n)$ whose average cost function $g_i(R(n))$ satisfies

$$0 \leq \frac{g_i(R(n)) - g^*}{g^*} \leq \varepsilon, \qquad i \in I,$$

where g^* denotes the minimal average cost per time unit. Regarding the choice of τ in the algorithm, it is recommended to take $\tau = \min_{i,a} \tau_i(a)$ when the embedded Markov chains $\{X_n\}$ in the semi-Markov model are aperiodic; otherwise, $\tau = \frac{1}{2} \min_{i,a} \tau_i(a)$ is a reasonable choice.

Linear programming formulation

The linear program for the semi-Markov decision model is given under the weak unichain assumption for the embedded Markov chains $\{X_n\}$. By the data transformation and the change of variable $u_{ia} = x_{ia}/\tau_i(a)$, the linear program (6.3.1) in Section 6.5 becomes:

$$\text{Minimize} \sum_{i \in I} \sum_{a \in A(i)} c_i(a) u_{ia}$$

subject to

$$\sum_{a \in A(j)} u_{ja} - \sum_{i \in I} \sum_{a \in A(i)} p_{ij}(a) u_{ia} = 0, \qquad a \in A(i) \text{ and } i \in I,$$

$$\sum_{i \in I} \sum_{a \in A(i)} \tau_i(a) u_{ia} = 1 \quad \text{and} \quad u_{ia} \geq 0, \qquad a \in A(i) \text{ and } i \in I.$$

The algorithm for deriving an optimal stationary policy from the LP solution is the same as in Section 6.5. In the same way as in Section 6.5 the linear programming formulation can be extended to cover probabilistic constraints such as the fraction of time that the system is in some subset I_0 of states should not exceed α. In the situation of probabilistic constraints, the average cost optimal policy usually involves randomized decisions.

7.3 VALUE ITERATION AND FICTITIOUS DECISIONS

The value-iteration method is often the most preferred method to compute a (nearly) average cost optimal policy. In each iteration of the method the lower and upper bounds indicate how much the average cost of the current policy deviates from the minimal average cost. The computational burden of the value-iteration algorithm depends not only on the number of states, but also on the density of the non-zero transition probabilities $p_{ij}(a)$. By the very nature of the value-iteration algorithm, it is computationally burdensome to have many non-zero $p_{ij}(a)$. In applications with exponentially distributed times between the decision epochs, the computational effort of the value-iteration algorithm can often be considerably reduced by including so-called fictitious decision epochs. The state of the system is left unchanged at the fictitious decision epochs. The inclusion of fictitious decision epochs does not change the Markovian nature of the decision process, since the times between state transitions are exponentially distributed and thus have the memoryless property. The trick of fictitious decision epochs reduces not only the computational effort, but also simplifies the formulation of the value-iteration algorithm. The inclusion of fictitious decision epochs has as a consequence that the state space must be enlarged with an indicator variable to distinguish between the fictitious decision epochs and the real decision epochs. However, the greater simplicity in formulation and the reduction in computing times outweigh the enlarged state space.

Example 7.3.1 Optimal allocation of servers to competing customers

In communication networks an important problem is the allocation of servers to competing customer classes. Suppose messages of the types 1 and 2 arrive at a communication system according to independent Poisson processes with respective rates λ_1 and λ_2. The communication system has c identical transmission channels for handling the messages, where each channel can handle only one message at a time. The system has no buffer for storing temporarily messages that find all channels occupied upon arrival. Such messages have to be rejected anyway. However, a newly arriving message may also be rejected when there is a free channel. The goal is to find a control rule that minimizes the average rejection rate or, equivalently, maximizes the average throughput of accepted messages. In Example 5.4.2 the best control rule was determined within the subclass of L-policies. Markov decision theory enables us to compute an overall optimal policy. To do so, we

make the assumption that the transmission times of the messages are exponentially distributed with mean $1/\mu_1$ for type 1 messages and with mean $1/\mu_2$ for type 2 messages.

Formulation with fictitious decision epochs

A straightforward formulation of the problem as a semi-Markov decision problem uses the arrival epochs as the only decision epochs. In such a formulation the vectors $(p_{ij}(a), j \in I)$ of one-step transition probabilities have many non-zero entries. In our specific problem this difficulty can be circumvented by including the service completion epochs as fictitious decision epochs in addition to the real decision epochs, being the arrival epochs of messages. By doing so, a transition from any state is always to one of at most four neighbouring states. In the approach with fictitious decision epochs, we take as state space

$$I = \{(i_1, i_2, k) \mid i_1, i_2 = 0, 1, \ldots, c;\ i_1 + i_2 \leq c;\ k = 0, 1, 2\}.$$

State (i_1, i_2, k) with $k = 1$ or 2 corresponds to the situation in which a type k message arrives and finds i_1 messages of type 1 and i_2 messages of type 2 being transmitted. The auxiliary state $(i_1, i_2, 0)$ corresponds to the situation in which the transmission of a message is just completed and i_1 messages of type 1 and i_2 messages of type 2 are left behind in the system. Note that the type of the transmitted message is not relevant. For the states (i_1, i_2, k) with $k = 1$ or 2 the possible actions are denoted by

$$a = \begin{cases} 0, & \text{reject the arriving message,} \\ 1, & \text{accept the arriving message,} \end{cases}$$

with the stipulation that $a = 0$ is the only feasible decision when $i_1 + i_2 = c$. The fictitious decision of leaving the system alone in the state $s = (i_1, i_2, 0)$ is also denoted by $a = 0$. Thanks to the fictitious decision epochs, each transition from a given state is to one of at most four neighbouring states. In other words, most of the one-step transition probabilities are zero. Further, the transition probabilities are extremely easy to specify, because of the fact that $\min(X_1, X_2)$ is exponentially distributed with mean $1/(\alpha_1 + \alpha_2)$ and $P\{X_1 < X_2\} = \alpha_1/(\alpha_1 + \alpha_2)$ when X_1 and X_2 are independent random variables having exponential distributions with respective means $1/\alpha_1$ and $1/\alpha_2$. Put for abbreviation

$$\nu(i_1, i_2) = \lambda_1 + \lambda_2 + i_1\mu_1 + i_2\mu_2.$$

Then, for action $a = 0$ in state $s = (i_1, i_2, k)$,

$$p_{sv}(0) = \begin{cases} \lambda_1/\nu(i_1, i_2), & v = (i_1, i_2, 1), \\ \lambda_2/\nu(i_1, i_2), & v = (i_1, i_2, 2), \\ i_1\mu_1/\nu(i_1, i_2), & v = (i_1 - 1, i_2, 0), \\ i_2\mu_2/\nu(i_1, i_2), & v = (i_1, i_2 - 1, 0). \end{cases}$$

and $\tau_s(0) = 1/\nu(i_1, i_2)$. For action $a = 1$ in state $s = (i_1, i_2, 1)$,

$$p_{sv}(1) = \begin{cases} \lambda_1/\nu(i_1+1, i_2), & v = (i_1+1, i_2, 1), \\ \lambda_2/\nu(i_1+1, i_2), & v = (i_1+1, i_2, 2), \\ (i_1+1)\mu_1/\nu(i_1+1, i_2), & v = (i_1, i_2, 0), \\ i_2\mu_2/\nu(i_1+1, i_2), & v = (i_1+1, i_2-1, 0). \end{cases}$$

and $\tau_s(1) = 1/\nu(i_1+1, i_2)$. Similarly, for action $a = 1$ in state $(i_1, i_2, 2)$. Finally, the one-step expected costs $c_s(a)$ are simply given by

$$c_s(a) = \begin{cases} 1, & s = (i_1, i_2, 1) \text{ and } a = 0, \\ 1, & s = (i_1, i_2, 2) \text{ and } a = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Value-iteration algorithm

Now, having specified the basic elements of the semi-Markov decision model, we are in a position to formulate the value-iteration algorithm for the computation of a (nearly) optimal acceptance rule. In the data transformation, we take

$$\tau = \frac{1}{\lambda_1 + \lambda_2 + c_1\mu_1 + c_2\mu_2}.$$

Using the above specifications, the value-iteration scheme becomes quite simple for the allocation problem. Note that the expressions for the one-step transition times $\tau_s(a)$ and the one-step transition probabilities $p_{st}(a)$ have a common denominator and so the ratio $p_{st}(a)/\tau_s(a)$ has a very simple form. In specifying the value-iteration scheme (7.2.3), we distinguish between the auxiliary states $(i_1, i_2, 0)$ and the other states. In the states $(i_1, i_2, 0)$ the only possible decision is to leave the system alone. Thus

$$\begin{aligned} V_n(i_1, i_2, 0) = {} & \tau\lambda_1 V_{n-1}(i_1, i_2, 1) + \tau\lambda_2 V_{n-1}(i_1, i_2, 2) + \tau i_1\mu_1 V_{n-1}(i_1-1, i_2, 0) \\ & + \tau i_2\mu_2 V_{n-1}(i_1, i_2-1, 0) + \{1 - \tau\nu(i_1, i_2)\} V_{n-1}(i_1, i_2, 0), \end{aligned}$$

where $V_{n-1}(i_1, i_2, 1) = 0$ when $i_1 < 0$ or $i_2 < 0$. For the states $(i_1, i_2, 1)$,

$$\begin{aligned} V_n(i_1, i_2, 1) = \min \Big[& \nu(i_1, i_2) + \tau\lambda_1 V_{n-1}(i_1, i_2, 1) + \tau\lambda_2 V_{n-1}(i_1, i_2, 2) \\ & + \tau i_1\mu_1 V_{n-1}(i_1-1, i_2, 0) + \tau i_2\mu_2 V_{n-1}(i_1, i_2-1, 0) \\ & + \{1 - \tau\nu(i_1, i_2)\} V_{n-1}(i_1, i_2, 1), \\ & \tau\lambda_1 V_{n-1}(i_1+1, i_2, 1) + \tau\lambda_2 V_{n-1}(i_1+1, i_2, 2) \\ & + \tau(i_1+1)\mu_1 V_{n-1}(i_1, i_2, 0) + \tau i_2\mu_2 V_{n-1}(i_1+1, i_2-1, 0) \\ & + \{1 - \tau\nu(i_1+1, i_2)\} V_{n-1}(i_1, i_2, 1) \Big], \end{aligned}$$

provided we put $V_{n-1}(i_1, i_2, 1) = V_{n-1}(i_1, i_2, 2) = \infty$ when $i_1 + i_2 = c+1$ in order to exclude the unfeasible decision $a = 1$ in the states $(i_1, i_2, 1)$ with $i_1 + i_2 = c$. A similar expression applies to $V_n(i_1, i_2, 2)$. This completes the specification of the recursion step of the value-iteration algorithm. The other steps of the algorithm go without saying.

The value-iteration algorithm for the semi-Markov decision formulation with fictitious decision epochs requires the extra states $(i_1, i_2, 0)$. However, the number of additions and multiplications per iteration is of the order c^2 rather than of the order c^4 as in a straightforward semi-Markov decision formulation. It appears from numerical experiments that there is a considerable overall reduction in computational effort when using the formulation with fictitious decision epochs. A further reduction in the computing times can be achieved by applying modified value iteration rather than standard value iteration; see Section 6.6.

Numerical investigations indicate that the overall optimal control rule has an intuitively appealing structure. It is characterized by integers $L_0, L_1, \ldots, L_{c-1}$ with $L_0 \geq L_1 \geq \cdots \geq L_{c-1}$. One type of message (call it the priority type) is always accepted as long as not all transmission channels are occupied. An arriving message of the non-priority type finding i priority type messages present upon arrival is only accepted when less than L_i non-priority type messages are present and not all channels are occupied. In the numerical example with $c = 10$, $\lambda_1 = 10$, $\lambda_2 = 7$, $\mu_1 = 10$ and $\mu_2 = 1$, the optimal L_i-values are given by

$$L_0 = L_1 = 8,\ L_2 = L_3 = 7,\ L_4 = 6,\ L_5 = 5,\ L_6 = 4,$$
$$L_7 = 3,\ L_8 = 2,\ L_9 = 1.$$

The minimal average loss rate is 1.767. A challenging open problem is to find a theoretical proof that an overall optimal policy has the L_i-structure. Another empirical result that deserves further investigation is the finding that the average loss rate under an L_i-policy is nearly insensitive to the form of the probability distributions of the transmission times; see also the discussion in Example 5.4.2.

7.4 OPTIMIZATION OF QUEUES

The semi-Markov model is a natural and powerful tool for the optimization of queues. Many queueing problems in telecommunication ask for the computation of an optimal control rule for a given performance measure. If the control rule is determined by one or two parameters, one might first use Markov chain analysis to calculate the performance measure for given values of the control parameters and next use a standard optimization procedure to find the optimal values of the control parameters. However, this is not always the most effective approach. Below we give an example of a controlled queueing system for which the semi-Markov decision approach is not only more elegant, but is also more effective than a direct search procedure. In this application the number of states is *unbounded*. However, by exploiting the structure of the problem, we are able to cast the problem into

a Markov decision model with a finite state space. Using a simple but generally useful *embedding* idea, we avoid brute-force truncation of the infinite set of states.

Example 7.4.1 Optimal control of a stochastic service system

A stochastic service system has s identical channels available for providing service, where the number of channels in operation can be controlled by turning channels on or off. For example, the service channels could be checkouts in a supermarket or production machines in a factory. Requests for service are sent to the service facility according to a Poisson process with rate λ. Each arriving request for service is allowed to enter the system and waits in line until an operating channel is provided. The service time of each request is exponentially distributed with mean $1/\mu$. It is assumed that the average arrival rate λ is less than the maximum service rate $s\mu$. A channel that is turned on can handle only one request at a time. At any time, channels can be turned on or off depending on the number of service requests in the system. A non-negative switching cost $K(a, b)$ is incurred when adjusting the number of channels turned on from a to b. For each channel turned on there is an operating cost at a rate of $r > 0$ per unit of time. Also, for each request a holding cost of $h > 0$ is incurred for each unit of time the message is in the system until its service is completed. The objective is to find a rule for controlling the number of channels turned on such that the long-run average cost per unit of time is minimal.

Since the Poisson process and the exponential distribution are memoryless, the state of the system at any time is described by the pair (i, t), where

$$i = \text{the number of service requests present,}$$

$$t = \text{the number of channels being turned on.}$$

The decision epochs are the epochs at which a new request for service arrives or the service of a request is completed. In this example the number of possible states is unbounded since the state variable i has the possible values $0, 1, \ldots$. A brute-force approach would result in a semi-Markov decision formulation in which the state variable i is bounded by a sufficiently large chosen integer U such that the probability of having more than U requests in the system is negligible under any reasonable control rule. This approach would lead to a very large state space when the arrival rate λ is close to the maximum service rate $s\mu$. A more efficient Markov decision formulation is obtained by restricting the class of control rules rather than truncating the state space. It is intuitively obvious that under each reasonable control rule all of the s channels will be turned on when the number of requests in the system is sufficiently large. In other words, choosing a sufficiently large integer M with $M \geq s$, it is from a practical point of view no restriction to assume that in the states (i, t) with $i \geq M$ the only feasible action is to turn on all of the s channels. However, this implies that we can restrict the control of the system only to those arrival epochs and service completion epochs at which

no more than M service requests remain in the system. By doing so, we obtain a semi-Markov decision formulation with the state space

$$I = \{(i, t) \mid 0 \le i \le M,\ 0 \le t \le s\},$$

and the action sets

$$A(i,t) = \begin{cases} \{a \mid a = 0, \ldots, s\}, & 0 \le i \le M-1,\ 0 \le t \le s, \\ \{s\}, & i = M,\ 0 \le t \le s. \end{cases}$$

Here action a in state (i, t) means that the number of channels turned on is adjusted from t to a. This semi-Markov decision formulation involves the following stipulation: if action $a = s$ is taken in state (M, t), then the next decision epoch is defined as the first service completion epoch at which either M or $M - 1$ service requests are left behind. Also, if action $a = s$ is taken in state (M, t), the 'one-step' costs incurred until the next decision epoch are defined as the sum of the switching cost $K(t, s)$ and the holding and operating costs made during the time until the next decision epoch. Denote by the random variable $T_M(s)$ the time until the next decision epoch when action $a = s$ is taken in state (M, t). The random variable $T_M(s)$ is the sum of two components. The first component is the time until the next service completion or the next arrival, whichever occurs first. The second component is zero if a service completion occurs first; otherwise, it is distributed as the time needed to reduce the number of service requests present from $M + 1$ to M. The semi-Markov decision formulation with an embedded state space makes sense only when it is feasible to calculate the one-step expected transition times $\tau_{(M,t)}(s)$ and the one-step expected costs $c_{(M,t)}(s)$.

The calculation of these quantities is easy, since service completions occur according to a Poisson process with rate $s\mu$ as long as all of the s channels are occupied. In other words, whenever M or more requests are in the system, we can equivalently imagine that a single 'superchannel' is servicing requests one at a time at an exponential rate of $s\mu$. This analogy enables us to invoke the formulas (2.6.2) and (2.6.3). Taking $n = 1$ and replacing the mean μ by $1/(s\mu)$ in these formulas, we find that the expected time needed to reduce the number of requests present from $M + 1$ to M, given that all channels are on, is

$$\frac{1/(s\mu)}{1 - \lambda/(s\mu)} = \frac{1}{s\mu - \lambda}$$

and the expected holding and operating costs incurred during the time needed to reduce the number of requests present from $M + 1$ to M, given that all channels are on, is

$$\frac{hM}{s\mu - \lambda} + \frac{hs\mu}{s\mu - \lambda}\left\{\frac{1}{s\mu} + \frac{\lambda}{s\mu(s\mu - \lambda)}\right\} + \frac{rs}{s\mu - \lambda} = \frac{h(M+1) + rs}{s\mu - \lambda} + \frac{h\lambda}{(s\mu - \lambda)^2}.$$

Here the term $hM/(s\mu - \lambda)$ represents the expected holding costs for the M service requests which are continuously present during the time needed to reduce

the number in system from $M+1$ to M. If all of the s channels are busy, then the time until the next event (service completion or new arrival) is exponentially distributed with mean $1/(\lambda+s\mu)$ and the next event is generated by an arrival with probability $\lambda/(\lambda+s\mu)$. Putting the pieces together, we find

$$\tau_{(M,t)}(s)=\frac{1}{\lambda+s\mu}+\frac{\lambda}{\lambda+s\mu}\left(\frac{1}{s\mu-\lambda}\right)=\frac{s\mu}{(\lambda+s\mu)(s\mu-\lambda)}$$

and

$$c_{(M,t)}(s)=K(t,s)+\frac{hM+rs}{\lambda+s\mu}+\frac{\lambda}{\lambda+s\mu}\left\{\frac{h(M+1)+rs}{s\mu-\lambda}+\frac{h\lambda}{(s\mu-\lambda)^2}\right\}.$$

Also, by the last argument above,

$$p_{(M,t)(M-1,s)}(s)=\frac{s\mu}{\lambda+s\mu}\qquad\text{and}\qquad p_{(M,t)(M,s)}(s)=\frac{\lambda}{\lambda+s\mu}.$$

For the other states of the embedded state space I, the basic elements of the semi-Markov decision model are easily specified. We have

$$\tau_{(i,t)}(a)=\frac{1}{\lambda+\min(i,a)\mu},\qquad 0\le i\le M-1,\ 0\le a\le s,$$

and

$$c_{(i,t)}(a)=K(t,a)+\frac{hi+ra}{\lambda+\min(i,a)\mu},\qquad 0\le i\le M-1,\ 0\le a\le s.$$

The one-step transition probabilities are left to the reader. Next we formulate the value-iteration algorithm. In the data transformation we take $\tau=1/(\lambda+s\mu)$. Then the recurrence relation (7.2.3) becomes

$$\begin{aligned}V_n((i,t))=\min_{0\le a\le s}\Bigg[&\{\lambda+\min(i,a)\mu\}K(t,a)+hi+ra\\&+\frac{\lambda}{\lambda+s\mu}V_{n-1}((i+1,a))+\frac{\min(i,a)\mu}{\lambda+s\mu}V_{n-1}((i-1,a))\\&+\left\{1-\frac{\lambda+\min(i,a)\mu}{\lambda+s\mu}\right\}V_{n-1}((i,t))\Bigg]\end{aligned}$$

for the states (i,t) with $0\le i\le M-1$, $0\le t\le s$. For the states (M,t),

$$\begin{aligned}V_n((M,t))=&\frac{1}{s\mu}(\lambda+s\mu)(s\mu-\lambda)K(t,s)+\frac{h\lambda}{s\mu-\lambda}+hM+rs\\&+\frac{s\mu-\lambda}{\lambda+s\mu}V_{n-1}((M-1,s))+\frac{\lambda(s\mu-\lambda)}{s\mu(\lambda+s\mu)}V_{n-1}((M,s))\\&+\left\{1-\frac{s\mu-\lambda}{s\mu}\right\}V_{n-1}((M,t))\end{aligned}$$

with the convention $V_{n-1}((-1,t))=0$.

Numerical results

We consider the switching cost function $K(a, b) = \kappa\,|a - b|$ and assume the numerical data

$$s = 10, \qquad \mu = 1, \qquad r = 30 \qquad \text{and} \qquad h = 10.$$

The arrival rate λ is 7 and 8, while the proportionality constant κ for the switching cost has the two values 10 and 25. In each example, we take the bound $M = 20$ for the states (i, t) with $i \geq M$ in which all of the s channels are always turned on. The value-iteration algorithm is started with $V_0((i, t)) = 0$ for all states (i, t) and uses the tolerance number $\varepsilon = 10^{-3}$ for its stopping criterion. Our numerical calculations indicate that for the case of linear switching costs, the average cost optimal control rule is characterized by parameters $s(i)$ and $t(i)$: the number of channels turned on is raised up to the level $s(i)$ in the states (i, t) with $t < s(i)$, the number of channels turned on is left unchanged in the states (i, t) with $s(i) \leq t \leq t(i)$ and the number of channels turned on is reduced to $t(i)$ in the states (i, t) with $t > t(i)$. Table 7.4.1 gives the (nearly) optimal values of $s(i)$ and $t(i)$ for each of the four examples considered. In each of these examples we applied both standard value iteration and modified value iteration; see Section 6.6. It was found that modified value iteration with a dynamic relaxation factor required considerably fewer iterations than standard value iteration. In the four examples, standard value iteration required

Table 7.4.1 Numerical results obtained by value iteration

	$\lambda = 7, \kappa = 10$		$\lambda = 8, \kappa = 10$		$\lambda = 7, \kappa = 25$		$\lambda = 8, \kappa = 25$	
i	$s(i)$	$t(i)$	$s(i)$	$t(i)$	$s(i)$	$t(i)$	$s(i)$	$t(i)$
0	0	3	0	4	0	6	0	6
1	1	4	1	4	1	6	1	7
2	2	4	2	5	2	6	2	7
3	2	5	3	5	3	6	3	7
4	3	6	3	6	3	7	3	8
5	4	6	4	7	4	7	4	8
6	5	7	5	8	5	8	5	8
7	5	8	5	8	5	8	6	9
8	6	9	6	9	6	9	6	9
9	6	9	7	10	6	9	7	10
10	7	10	7	10	7	10	7	10
11	8	10	8	10	7	10	7	10
12	8	10	9	10	7	10	8	10
13	9	10	9	10	8	10	8	10
14	9	10	10	10	8	10	9	10
15	10	10	10	10	8	10	9	10
16	10	10	10	10	9	10	10	10
17	10	10	10	10	9	10	10	10
18	10	10	10	10	9	10	10	10
19	10	10	10	10	10	10	10	10
≥ 20	10	10	10	10	10	10	10	10

174, 311, 226 and 250 iterations. Modified value iteration required 59, 82, 87 and 71 iterations and ended up with the respective bounds $(m_n, M_n) = (319.3, 319.5)$, $(367.1, 367.4)$, $(331.5, 331.8)$ and $(378.0, 378.3)$ on the minimal average cost.

7.5 ONE-STEP POLICY IMPROVEMENT

The policy-iteration algorithm has the remarkable feature that it achieves the largest improvements in costs in the first few iterations. These findings underlie a heuristic approach for Markov decision problems with a multidimensional state space. In such decision problems it is usually not feasible to solve the value-determination equations. However, a policy-improvement step offers in general no computational difficulties. This suggests a heuristic approach that determines first a good estimate for the relative values and next applies a single policy-improvement step. By the nature of the policy-iteration algorithm one might expect to obtain a good decision rule by the heuristic approach. How to compute the relative values to be used in the policy-improvement step typically depends on the specific application. The heuristic approach is illustrated in the next example.

Example 7.5.1 Dynamic routing of customers to parallel queues

An important queueing model arising in various practical situations is one in which arriving customers (messages or jobs) have to be assigned to one of several different groups of servers. Problems of this type occur in telecommunication networks and flexible manufacturing. The queueing system consists of n multi-server groups working in parallel, where each group has its own queue. There are s_k servers in group k $(k = 1, \ldots, n)$. Customers arrive according to a Poisson process with rate λ. Upon arrival each customer has to be assigned to one of the n server groups. The assignment is irrevocable. The customer waits in the assigned queue until a server becomes available. Each server can handle only one customer at a time.

The problem is to find an assignment rule that (nearly) minimizes the average sojourn time per customer. This problem will be analysed under the assumption that the service times of the customers are independent and exponentially distributed. The mean service time of a customer assigned to queue k is $1/\mu_k$ $(k = 1, \ldots, n)$. It is assumed that $\lambda < \sum_{k=1}^{n} s_k \mu_k$. In what follows we consider the minimization of the overall average number of customers in the system. In view of Little's formula, the minimization of the average sojourn time per customer is equivalent to the minimization of the average number of customers in the system.

Bernoulli-splitting rule

An intuitively appealing control rule is the shortest-queue rule. Under this rule each arriving customer is assigned to the shortest queue. Except for the special case of $s_1 = \cdots = s_n$ and $\mu_1 = \cdots = \mu_n$, this rule is in general not optimal. In particular, the shortest-queue rule may perform quite unsatisfactorily in the situation

of a few fast servers and many slow servers. Another simple rule is the Bernoulli-splitting rule. Under this rule each arrival is assigned with a given probability p_k to queue k $(k = 1, \ldots, n)$ irrespective of the queue lengths. This assignment rule produces independent Poisson streams at the various queues, where queue k receives a Poisson stream at rate λp_k. The probabilities p_k must satisfy $\sum_k p_k = 1$ and $\lambda p_k < s_k \mu_k$ for $k = 1, \ldots, n$. This condition guarantees that no infinitely long queues can build up. Under the Bernoulli-splitting rule it is easy to give an explicit expression for the overall average number of customers in the system. The separate queues act as independent queues of the $M/M/s$ type. This basic queueing model is discussed in Chapter 5. In the $M/M/s$ queue with arrival rate α and s exponential servers each with service rate μ, the long-run average number of customers in the system equals

$$L(s, \alpha, \mu) = \frac{\rho(s\rho)^s}{s!(1-\rho)^2} \left\{ \sum_{k=0}^{s-1} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!(1-\rho)} \right\}^{-1} + s\rho,$$

where $\rho = \alpha/(s\mu)$. Under the Bernoulli-splitting rule the overall average number of customers in the system equals

$$\sum_{k=1}^{n} L(s_k, \lambda p_k, \mu_k). \tag{7.5.1}$$

The best Bernoulli-splitting rule is found by minimizing this expression with respect to $p_1, \ldots, p_n$ subject to the condition $\sum_k p_k = 1$ and $0 \leq \lambda p_k < s_k \mu_k$ for $k = 1, \ldots, n$. This minimization problem must be numerically solved by some search procedure (for $n = 2$, bisection can be used to find the minimum of a unimodal function in a single variable).

Policy-improvement step

The problem of assigning the arrivals to one of the server groups is a Markov decision problem with a multidimensional state space. The decision epochs are the arrival epochs of new customers. The state of the system at a decision epoch is an n-dimensional vector $x = (i_1, \ldots, i_n)$, where i_j denotes the number of customers present in queue j. This description uses the memoryless property of the exponential service times. The action $a = k$ in state x means that the new arrival is assigned to queue k. To deal with the optimality criterion of the long-run average number of customers in the system, we impose the following cost structure on the system. A cost at rate j is incurred whenever there are j customers in the system. Then the long-run average cost per time unit gives the long-run overall average number of customers in the system.

Denote by policy $R^{(0)}$ the best Bernoulli-splitting rule and let $p_k^{(0)}$, $k = 1, \ldots, n$ be the splitting probabilities associated with policy $R^{(0)}$. We already pointed out that the average cost for rule $R^{(0)}$ is easy to compute. Below it will be shown that the relative values are also easy to obtain for rule $R^{(0)}$. Let us first explain how to

derive an improved policy from the Bernoulli-splitting rule $R^{(0)}$. This derivation is based on first principles discussed in Section 6.2. The basic idea of the policy-improvement step is to minimize for each state x the difference $\Delta(x, a, R^{(0)})$ defined by

$$\Delta(x, a, R^{(0)}) = \text{the difference in total expected costs over an infinitely long period of time by taking first action } a \text{ and next using policy } R^{(0)} \text{ rather than using policy } R^{(0)} \text{ from scratch when the initial state is } x.$$

The difference is well defined since the Markov chain associated with policy $R^{(0)}$ is aperiodic. Under the Bernoulli-splitting rule the n queues act as independent $M/M/s$ queues. Define for each separate queue j,

$$D_j(i) = \text{the difference in total expected costs in queue } j \text{ over an infinitely long period of time by starting with } i+1 \text{ customers in queue } j \text{ rather than with } i \text{ customers.}$$

Then, for each state $x = (i_1, \ldots, i_n)$ and action $a = k$,

$$\Delta(x, a, R^{(0)}) = \sum_{\substack{j=1 \\ j \neq k}}^{n} p_j^{(0)}[-D_j(i_j) + D_k(i_k)] + p_k^{(0)} \times 0$$

$$= -\sum_{j=1}^{n} p_j^{(0)} D_j(i_j) + D_k(i_k).$$

Since the term $\sum_j p_j^{(0)} D_j(i_j)$ does not depend on the action $a = k$, the step of minimizing $\Delta(x, k, R^{(0)})$ over k reduces to the computation of

$$\min_{1 \leq k \leq n} \{D_k(i_k)\}.$$

Hence a remarkably simple expression is evaluated in the policy-improvement step applied to the Bernoulli-splitting rule. The suboptimal rule resulting from the single application of the policy-improvement step is called the *separable rule*. The performance of this rule will be discussed below.

It remains to specify the function $D_k(i)$, $i = 0, 1, \ldots$, for each queue k. To do so, consider an $M/M/s$ queue in isolation, where customers arrive according to a Poisson process with rate α and there are s exponential servers each with service rate μ. Each arrival is admitted to the queue. The state of the system describes the number of customers present. A cost at rate j is incurred when there are j customers present. The long-run average cost per time unit is given by

$$g = L(s, \alpha, \mu).$$

The $M/M/s$ queueing process can be seen as a Markov decision process with a single decision in each state. The decision is to leave the system alone. In this Markov decision formulation it is convenient to consider the state of the system both at the arrival epochs and the service completion epochs. In the $M/M/s$ queue the situation of i customers present just after a service completion is probabilistically the same as the situation of i customers present just after an arrival. In accordance with (6.3.1), we define the relative cost function $w(i)$ by

$$w(i) = \begin{cases} K(i) - gT(i), & i = 1, 2, \dots, \\ 0, & i = 0, \end{cases} \tag{7.5.2}$$

where

$T(i) =$ the expected time until the first return to an empty system starting with i customers present,

$K(i) =$ the total expected cost incurred until the first return to an empty system starting with i customers present.

Then, by the economic interpretation of the relative values given in Section 6.3, we have for any $i = 0, 1, \dots$ that

$w(i+1) - w(i) =$ the difference in total expected costs over an infinitely long period of time by starting in state $i+1$ rather than in state i.

The desired function $D_k(i)$ for queue k follows by taking

$$D_k(i) = w_k(i+1) - w_k(i) \qquad \text{with } \alpha = \lambda p_k,\ s = s_k \text{ and } \mu = \mu_k.$$

The basic functions $K(i)$ and $T(i)$ are easy to compute. By conditioning,

$$T_i = \frac{1}{\alpha + i\mu} + \frac{\alpha}{\alpha + i\mu} T_{i+1} + \frac{i\mu}{\alpha + i\mu} T_{i-1}, \qquad 1 \le i \le s. \tag{7.5.3}$$

$$K_i = \frac{i}{\alpha + i\mu} + \frac{\alpha}{\alpha + i\mu} K_{i+1} + \frac{i\mu}{\alpha + i\mu} K_{i-1}, \qquad 1 \le i \le s. \tag{7.5.4}$$

where $T_0 = K_0 = 0$. Further, we have

$$T_i = \frac{i - s}{s\mu - \alpha} + T_s, \qquad i > s,$$

$$K_i = \frac{1}{s\mu - \alpha} \left\{ \frac{1}{2}(i-s)(i-s-1) + i - s + \frac{\alpha(i-s)}{s\mu - \alpha} \right\} + \frac{s(i-s)}{s\mu - \alpha}, \qquad i > s.$$

To see the latter relations, note that the time to reach an empty system from state $i > s$ is the sum of the time to reach state s and the time to reach an empty system from state s. By the memoryless property of the exponential distribution, the multi-server $M/M/s$ queue operates as a single-server $M/M/1$ queue with service rate $s\mu$ when s or more customers are present. Next, by applying the formulas (2.6.2) and (2.6.3), we find the formulas for T_i and K_i when $i > s$. Substituting the expressions

for T_{s+1} and K_{s+1} into (7.5.3) and (7.5.4) with $i = s$, we get two systems of linear equations for T_i, $1 \leq i \leq s$ and K_i, $1 \leq i \leq s$. Once these systems of linear equations have been solved, we can next compute T_i and K_i for any desired $i > s$.

Summarizing, the heuristic algorithm proceeds as follows.

Heuristic algorithm

Step 1. Compute the best values $p_k^{(0)}$, $k = 1, \ldots, n$, of the Bernoulli-splitting probabilities by minimizing the expression (7.5.1) subject to $\sum_{k=1}^{n} p_k = 1$ and $0 \leq \lambda p_k < s_k \mu_k$ for $k = 1, \ldots, n$.

Step 2. For each queue $k = 1, \ldots, n$, solve the system of linear equations (7.5.3) and (7.5.4) with $\alpha = \lambda p_k^{(0)}$, $s = s_k$ and $\mu = \mu_k$. Next compute for each queue k the function $w_k(i)$ from (7.5.2) with $\alpha = \lambda p_k^{(0)}$, $s = s_k$ and $\mu = \mu_k$.

Step 3. For each state $x = (i_1, \ldots, i_n)$, determine an index k_0 achieving the minimum in

$$\min_{1 \leq k \leq n} \{w_k(i_k + 1) - w_k(i_k)\}.$$

The separable rule assigns a new arrival in state $x = (i_1, \ldots, i_n)$ to queue k_0.

Numerical results

Let us consider the numerical data

$$s_1 = 10, \qquad s_2 = 1, \qquad \mu_1 = 1 \qquad \text{and} \qquad \mu_2 = 9.$$

The traffic load ρ, which is defined by

$$\rho = \lambda/(s_1\mu_1 + s_2\mu_2),$$

is varied as $\rho = 0.2$, 0.5, 0.7, 0.8 and 0.9. In addition to the theoretically minimal average sojourn time, Table 7.5.1 gives the average sojourn time per customer for the Bernoulli-splitting rule (B-split) and for the heuristic separable rule. The table also gives the average sojourn time per customer under the shortest expected delay (SED) rule. Under this rule an arriving customer is assigned to the queue in which its expected individual delay is smallest (if there is a tie, the customer is sent to queue 1). The results in the table show that this intuitively appealing control policy performs unsatisfactorily for the case of heterogeneous services. However, the heuristic separable rule shows an excellent performance for all values of ρ.

Table 7.5.1 The average sojourn times

ρ	SED	B-split	Separable	Optimal
0.2	0.192	0.192	0.191	0.191
0.5	0.647	0.579	0.453	0.436
0.7	0.883	0.737	0.578	0.575
0.8	0.982	0.897	0.674	0.671
0.9	1.235	1.404	0.941	0.931

EXERCISES

7.1 Consider a production facility that operates only intermittently to manufacture a single product. The production will be stopped if the inventory is sufficiently high, whereas the production will be restarted when the inventory has dropped sufficiently low. Customers asking for the product arrive according to a Poisson process with rate λ. The demand of each customer is for one unit. Demand which cannot be satisfied directly from stock on hand is lost. Also, a finite capacity C for the inventory is assumed. In a production run, any desired lot size can be produced. The production time of a lot size of Q units is a random variable T_Q having a probability density $f_Q(t)$. The lot size is added to the inventory at the end of the production run. After the completion of a production run, a new production run is started or the facility is closed down. At each point of time the production can be restarted. The production costs for a lot size of $Q \geq 1$ units consist of a fixed set-up cost $K > 0$ and a variable cost c per unit produced. Also, there is a holding cost of $h > 0$ per unit kept in stock per time unit, and a lost-sales cost of $p > 0$ is incurred for each lost demand. The goal is to minimize the long-run average cost per time unit. Formulate the problem as a semi-Markov decision model.

7.2 Consider the maintenance problem from Example 6.1.1 again. The numerical data are given in Table 6.4.1. Assume now that a repair upon failure takes either 1, 2 or 3 days, each with probability 1/3. Use the semi-Markov model to compute by policy iteration or linear programming an average cost optimal policy. Can you explain why you get the same optimal policy as in Example 6.1.1?

7.3 A cargo liner operates between the five harbours $A_1, \ldots, A_5$. A cargo shipment from harbour A_i to harbour A_j $(j \neq i)$ takes a random number τ_{ij} of days (including load and discharge) and yields a random pay-off of ξ_{ij}. The shipment times τ_{ij} and the pay-offs ξ_{ij} are normally distributed with means $\mu(\tau_{ij})$ and $\mu(\xi_{ij})$ and standard deviations $\sigma(\tau_{ij})$ and $\sigma(\xi_{ij})$. We assume the numerical data:

	$\mu(\tau_{ij})[\sigma(\tau_{ij})]$				
$i \backslash j$	1	2	3	4	5
1	-	$3\left[\frac{1}{2}\right]$	6 [1]	$3\left[\frac{1}{2}\right]$	$2\left[\frac{1}{2}\right]$
2	4 [1]	-	$1\left[\frac{1}{4}\right]$	7 [1]	5 [1]
3	5 [1]	$1\left[\frac{1}{4}\right]$	-	6 [1]	8 [1]
4	$3\left[\frac{1}{2}\right]$	8 [1]	5 [1]	-	$2\left[\frac{1}{2}\right]$
5	$2\left[\frac{1}{2}\right]$	5 [1]	9 [1]	$2\left[\frac{1}{2}\right]$	-

	$\mu(\xi_{ij})[\sigma(\xi_{ij})]$				
$i \backslash j$	1	2	3	4	5
1	-	8 [1]	12 [2]	6 [1]	6 [1]
2	20 [3]	-	$2\left[\frac{1}{2}\right]$	14 [3]	16 [2]
3	16 [3]	$2\left[\frac{1}{2}\right]$	-	18 [3]	16 [1]
4	6 [1]	10 [2]	20 [2]	-	$6\left[\frac{1}{2}\right]$
5	8 [2]	16 [3]	20 [2]	8 [1]	-

Compute by policy iteration or linear programming a sailing route for which the long-run average reward per day is maximal.

7.4 Consider Exercise 2.20 again. Assume that the assignment types $j = 1, \ldots, n$ are numbered or renumbered according to $E(\xi)/E(\tau_j) \geq E(\xi_{j+1})/E(\tau_{j+1})$ for all j. Use the optimality equation (7.2.2) to verify that the long-run average reward per time unit is maximal by accepting only assignments of the types $j = 1, \ldots, r$, where r is the smallest integer such that

$$\sum_{j=1}^{r} \lambda_j E(\xi_j) \Bigg/ \left[1 + \sum_{j=1}^{r} \lambda_j E(\tau_j) \right] > E(\xi_{r+1})/E(\tau_{r+1})$$

with $E(\xi_{n+1})/E(\tau_{n+1}) = 0$ by convention.

7.5 Adjust the value-iteration algorithm for the control problem from Example 7.3.1 when finite-source input is assumed rather than Poisson input. Solve for the numerical data $c = 10$, $M_1 = M_2 = 10$, $\delta_1 = 3$, $\delta_2 = 1$, $\mu_1 = 4$, $\mu_2 = 1$, where M_i is the number of customers from source i and δ_i is the exponential rate at which a customer from source i generates new service requests when the customer has no other request in service. Try other numerical examples and investigate the structure of an optimal control rule.

7.6 Consider a flexible manufacturing facility producing parts, one at a time, for two assembly lines. The time needed to produce one part for assembly line k is exponentially distributed with mean $1/\mu_k$, $k = 1, 2$. Each part produced for line k is put into the buffer for line k. This buffer has space for only N_k parts, including the part (if any) in assembly. Each line takes parts one at a time from its buffer as long as the buffer is not empty. At line k, the assembly time for one part is exponentially distributed with mean $1/\lambda_k$, $k = 1, 2$. The production times at the flexible manufacturing facility and the assembly times at the lines are independent of each other. A real-time control for the flexible manufacturing facility is exercised. After each production at this facility, it must be decided what type of part is to be produced next. The system cannot produce for a line whose buffer is full. Also, the system cannot remain idle if not all the buffers are full. The control is based on the full knowledge of the buffer status at both lines. The system incurs a lost-opportunity cost at a rate of γ_k per time unit when line k is idle. The goal is to control the production at the flexible manufacturing facility in such a way that the long-run average cost per time unit is minimal. Develop a value-iteration algorithm for this control problem. Solve for the numerical data $\mu_1 = 5$, $\mu_2 = 10$, $\lambda_1 = 4$, $\lambda_2 = 8$, $N_1 = N_2 = 5$, $\gamma_1 = \gamma_2 = 1$. This problem is based on Seidman and Schweitzer (1984).

7.7 Consider a tandem network with two assembly facilities in series. The output of the first station is the input for the second station. Raw material is processed at station 1, and half-finished goods at station 2. Each of the stations 1 and 2 has a finite buffer for temporarily storing raw material and half-finished goods. The buffer size is M at station 1 and N at station 2 (excluding any unit in processing). Units of raw material arrive at station 1 according to a Poisson process with rate λ. A unit of raw material finding the buffer full at station 1 upon arrival is rejected and is brought elsewhere. Station 1 is a single-server station and station 2 is a multiple-server station with c servers. Each server can handle only one unit at a time and the processing times are exponentially distributed with mean $1/\mu_1$ at station 1 and mean $1/\mu_2$ at station 2. If the assembly of a unit is finished at station 1, it is forwarded to station 2 provided the buffer is not full at station 2; otherwise, the unit remains at station 1 and blocks this station until room becomes available at station 2. Station 1 cannot start a new assembly as long as it is blocked. The control problem is as follows. Upon arrival of a new unit at station 1, a decision has to be made to accept this unit or to reject it. The cost of rejecting a unit at station 1 is $R > 0$. Also, there is a blocking cost at rate $b > 0$ per time unit that station 1 is blocked. The goal is to find a control rule minimizing the long-run average cost per time unit. Develop a value-iteration algorithm. Solve for the numerical data $\lambda = 20$, $\mu_1 = 15$, $\mu_2 = 3$, $c = 5$, $M = 10$, $N = 3$, $R = 3.5$

and $b = 20$. Try other numerical examples and investigate whether the optimal control rule is characterized by integers $L_0, \ldots, L_M$ so that an arriving unit of raw material finding i units present at station 1 is only accepted when less than L_i units are present at station 2.

7.8 Consider the situation that two groups of servers share a common waiting room. The first group consists of c_1 servers and the second group consists of c_2 servers. Customers for the first group arrive according to a Poisson process with rate λ_1 and, independently of this process, customers for the second group arrive according to a Poisson process with rate λ_2. Upon arrival of a new customer, a decision has to be made to accept or reject them. An accepted customer keeps one place in the waiting room occupied until their service is completed. The service times of the customers are exponentially distributed with a mean $1/\mu_1$ for a customer going to the first group and mean $1/\mu_2$ for a customer going to the second group. Each server can handle only one customer at a time and serves only customers for the group to which the server belongs. The goal is to find a control rule minimizing the total average rejection rate. Develop a value-iteration algorithm for this control problem. Solve for the numerical data $c_1 = c_2 = 1$, $\lambda_1 = 1.2$, $\lambda_2 = 1$, $\mu_1 = \mu_2 = 1$ and $M = 15$, where M denotes the number of places in the waiting room. Try other numerical examples and verify experimentally that the optimal control rule is characterized by two sequences $\{a_1^{(r)}, 0 \leq r \leq M\}$ and $\{a_2^{(r)}, 0 \leq r \leq M\}$ so that an arriving customer of type k finding r customers of the other type present upon arrival is accepted only if less than $a_r^{(k)}$ customers of the same type k are present and the waiting room is not full. This problem is based on Tijms and Eikeboom (1986).

7.9 Consider the problem of designing an optimal buffer management policy in a shared-memory switch with the feature that packets already accepted in the switch can be dropped (pushed out). The system has two output ports and a finite buffer shared by the two output ports. Packets of types 1 and 2 arrive according to independent Poisson processes with rates λ_1 and λ_2. Packets of type i are destined for output port i for $i = 1, 2$. At each of the two output ports there is a single transmission channel. Each channel can transmit only one packet at a time and the transmission time at output port i is exponentially distributed with mean $1/\mu_i$ for $i = 1, 2$. Upon arrival of a new packet, the system has to decide whether to accept the packet, to reject it, or to accept it and drop a packet of the other type. A packet that is rejected or dropped is called a lost packet and has no further influence on the system. The total buffer size is B and it is assumed that an accepted packet occupies a buffer place until its transmission is completed. The goal is to find a control rule minimizing the overall fraction of packets that are lost. Develop a value-iteration algorithm. Solve for the numerical data $\lambda_1 = 1$, $\lambda_2 = 10$, $\mu_1 = 2$, $\mu_2 = 20$ and $B = 12$. Try other numerical examples and investigate whether the optimal control rule has a specific structure. This problem is based on Cidon *et al.* (1995).

7.10 Consider a two-server facility with heterogeneous servers. The faster server is always available and the slower server is activated for assistance when too many customers are waiting. Customers arrive according to a Poisson process with rate λ. The service facility has ample waiting room. The service times of the customers are independent of each other and have an exponential distribution. The mean service time is $1/\mu_1$ when service is provided by the faster server and is $1/\mu_2$ for the slower server, where $1/\mu_1 < 1/\mu_2$. It is assumed that the load factor $\lambda/(\mu_1 + \mu_2)$ is less than 1. The slower server can only be turned off when it has completed a service. The slower server cannot be on when the system is empty. If the slower server is kept on while customers are waiting for service, it cannot remain idle. Each server can handle only one customer at a time. Service is non-pre-emptive; that is, the faster server cannot take over a customer from the slower server. A fixed cost of $K \geq 0$ is incurred each time the slower server is turned on and there is an operating cost of $r > 0$ per time unit the slower server is on. Also, a holding cost of $h > 0$ per time unit is incurred for each customer in the system. The goal is to find a switching rule that minimizes the

long-run average cost per time unit. Using the embedding idea from Section 7.4, develop a value-iteration algorithm for the control problem. Solve for the numerical data $\lambda = 3$, $\mu_1 = 2.8$, $\mu_2 = 2.2$, $h = 2$, $r = 4$ and $K = 10$. Try other numerical examples and verify experimentally that the optimal control rule is a so-called hysteretic (m, M) rule under which the slower server is turned on when the number of customers present is M or more and the slower server is switched off when this server completes a service and the number of customers left behind in the system is below m. This problem is based on Nobel and Tijms (2000), who developed a tailor-made policy-iteration algorithm for this problem.

7.11 Consider again the heterogeneous server problem from Exercise 7.10. Assume now that there are two slower servers in addition to the faster server, where the two slower servers may have different speeds. The faster server is always available for service, while the slower servers are activated for service when too many customers are present. The service time of a customer is exponentially distributed with mean $1/\mu_i$ when service is provided by server i. Server 1 is the faster server and servers 2 and 3 are the slower servers. It is assumed that $\lambda/(\mu_1 + \mu_2 + \mu_3) < 1$ and $\mu_1 > \max(\mu_2, \mu_3)$. There is an operating cost of $r_i > 0$ per time unit when the slower server i is on for $i = 2, 3$. A holding cost of $h > 0$ per time unit is incurred for each customer in the system. There is no switching cost for turning either of the slower servers on. Develop a value-iteration algorithm for this problem. Assuming that the slower servers are numbered such that $r_2/\mu_2 < r_3/\mu_3$, verify experimentally that the optimal control rule is characterized by critical numbers $1 \leq m_1 < m_2$ and prescribes using the slower server 2 when the number of customers present is more than m_1, and using both slow servers when the number of customers present is more than m_2.

7.12 Messages arrive at a transmission channel according to a Poisson process with a controllable arrival rate. The two possible arrival rates are λ_1 and λ_2 with $0 \leq \lambda_2 < \lambda_1$. The buffer at the transmission channel has ample space for temporarily storing arriving messages. The channel can only transmit one message at a time. The transmission time of each message is exponentially distributed with mean $1/\mu$. It is assumed that $\lambda_2/\mu < 1$. At any point in time it can be decided to change the arrival intensity from one rate to the other. There is a fixed cost of $K \geq 0$ for changing the arrival rate. An operating cost of $r_i > 0$ per time unit is incurred when the prevailing arrival rate is λ_i, $i = 1, 2$. Also, there is a holding cost of $h > 0$ per time unit for each message awaiting service. The goal is to find a control rule that minimizes the long-run average cost per time unit. Using the embedding idea from Example 7.4.1, develop a value-iteration algorithm for this control problem. Solve for the numerical data $\lambda_1 = 4$, $\lambda_2 = 2$, $\mu = 5$, $K = 5$, $r_1 = 1$, $r_2 = 10$ and $h = 2$. Try other numerical examples and investigate whether the optimal control rule has a specific structure.

7.13 Customers of types 1 and 2 arrive at a shared resource according to independent Poisson processes with respective rates λ_1 and λ_2. The resource has c service units. An arriving customer of type i requires b_i service units. The customer is rejected when less than b_i units are available upon arrival. An accepted customer of type i immediately enters service and has an exponentially distributed residency time with mean $1/\mu_i$. During this residency time the customer keeps all of the b_i assigned service units occupied. These units are released simultaneously when the customer departs. Develop a value-iteration algorithm for the computation of a control rule that minimizes the total average rejection rate. Solve for the numerical data $c = 30$, $b_1 = 2$, $b_2 = 5$, $\lambda_1 = 6$, $\lambda_2 = 8$, $\mu_1 = 1$ and $\mu_2 = 0.5$. Try other numerical examples and verify experimentally that the optimal control rule can be characterized by two monotone sequences $\{a_1^{(r)}\}$ and $\{a_2^{(r)}\}$. Under this control rule an arriving customer of type i finding r customers of the other type present upon arrival is accepted only when less than $a_i^{(r)}$ customers of the same type i are present and at least b_i service units are free.

7.14 Consider Exercise 7.13 again, but assume now that the residency times have a Coxian-2 distribution. Develop a value-iteration algorithm to compute the total average rejection rate under a *fixed* reservation policy. A reservation policy is characterized by two integers r_1 and r_2 with $r_1 \geq b_1$ and $r_2 \geq b_2$. Under the reservation policy an arriving customer of type i is accepted only if r_i or more service units are available. Verify experimentally that the total average rejection rate for a fixed reservation policy is nearly insensitive to the second and higher moments of the residency times. For the case of exponentially distributed residency times, take the average rejection rate of the best reservation policy and verify how close it is to the theoretically minimal average rejection rate.

7.15 Consider a production/inventory system with N inventory points that share a common production unit. At the beginning of each period, the production unit can produce for any number of inventory points, with the stipulation that the total production size is restricted by the capacity C of the production unit. The production time is negligible for any production scheme. The demands at the various inventory points are independent of each other. In each period the demand at inventory point j is Poisson distributed with mean μ_j for $j = 1, \ldots, N$. Excess demand is lost at any inventory point. The following costs are involved. The cost of producing z_j units for inventory point j equals $K_j + c_j z_j$ for $z_j > 0$ regardless of how much is produced for each of the other inventory points. In any period there is a holding cost of $h_j > 0$ for each unit in stock at inventory point j at the end of the period. A stockout cost of p_j is incurred for each unit of lost demand at inventory point j. Can you think of a heuristic approach based on solving N one-dimensional problems and performing a single policy-improvement step? This problem is based on Wijngaard (1979).

BIBLIOGRAPHIC NOTES

Semi-Markov decision processes were introduced in De Cani (1964), Howard (1964), Jewell (1963) and Schweitzer (1965). The semi-Markov decision model has many applications, especially in queueing control. The data-transformation method converting a semi-Markov decision model into an equivalent discrete-time Markov decision model was introduced in Schweitzer (1971). This uniformization method was used in the paper of Lippman (1975) to establish the structure of optimal control rules in queueing applications of continuous-time Markov decision processes with exponentially distributed times between the decision epochs. The idea of using fictitious decision epochs is also contained in this paper. The embedding idea used in Section 7.4 is adapted from De Leve *et al.* (1977); see also Tijms (1980). Embedding is especially useful for developing a tailor-made policy-iteration algorithm that operates on a subclass of structured policies. The heuristic approach of attacking a multidimensional Markov decision problem through decomposition and a single-improvement step goes back to Norman (1970) and has been successfully applied in Krishnan and Ott (1986,1987) and Wijngaard (1979), among others. The heuristic solution for the dynamic routing problem from Example 7.5.1 comes from Krishnan and Ott (1987) and has been extended in Sassen *et al.* (1997) to the case of general service times. Other heuristic approaches to handle large-scale Markov decision processes are discussed in Cooper *et al.* (2003) and Schweitzer and Seidman (1985).

REFERENCES

Cidon, I., Georgiadis, L., Gúerin, R. and Khamisy, A. (1995) Optimal buffer sharing. *IEEE J. Selected Areas in Commun.*, **13**, 1229–1240.

Cooper, W.L., Henderson, S.G. and Lewis, M.E. (2003) Convergence of simulation-based policy iteration. *Prob. Engng and Inform. Sci.*, **17**, 000–000.

De Cani, J.S. (1964) A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity. *Management Sci.*, **10**, 716–733.

De Leve, G., Federgruen, A. and Tijms, H.C. (1977) A general Markov decision method, I: model and method, II: applications. *Adv. Appl. Prob.*, **9**, 296–335.

Howard, R.A. (1964) Research in semi-Markovian decision structures. *J. Operat. Res. Soc. Japan*, **6**, 163–199.

Jewell, W.S. (1963) Markov renewal programming: I and II. *Operat. Res.*, **11**, 938–971.

Krishnan, K.R. and Ott, T.J. (1986) State-dependent routing for telephone traffic: theory and results. In: *Proce 25th IEEE Conference on Decision and Control*, Athens, Greece, pp. 2124–2128 IEEE, New York.

Krishnan, K.R. and Ott, T.J. (1987) Joining the right queue: a Markov decision rule. In: *Proce 26th IEEE Conference on Decision and Control*, Los Angeles, pp. 1863–1868 IEEE, New York.

Lippman, S.A. (1975) Applying a new device in the optimization of exponential queueing systems. *Operat. Res.*, **23**, 687–710.

Nobel, R.D. and Tijms, H.C. (2000) Optimal control of a queueing system with heterogeneous servers and setup costs. *IEEE Trans. Automat. Contr.*, **45**, 780–784.

Norman, J.M. (1972) *Heuristic Procedures in Dynamic Programming*. Manchester University Press, Manchester.

Sassen, S.A.E., Tijms, H.C. and Nobel, R.D. (1997) A heuristic rule for routing customers to parallel servers. *Statistica Neerlandica*, **51**, 107–121.

Schweitzer, P.J. (1965) *Perturbation Theory and Markovian Decision Processes.* PhD dissertation, Massachusetts Institute of Technology.

Schweitzer, P.J. (1971) Iterative solution of the functional equations of undiscounted Markov renewal programming. *J. Math. Anal. Appl.*, **34**, 495–501.

Schweitzer, P.J. and Seidman, A. (1985) Generalized polynomial approximations in Markovian decision processes. *J. Math. Anal. Appl.*, **110**, 568–582.

Seidman, A. and Schweitzer, P.J. (1984) Part selection policy of a flexible manufacturing cell feeding several production lines. *AIEE Trans.*, **16**, 355–362.

Tijms, H.C. (1980) An algorithm for average cost denumerable state semi-Markov decision problems with applications to controlled production and queueing systems. In: *Recent Developments in Markov Decision Processes*, edited by R. Hartley, L.C. Thomas and D.J. White, pp. 143–179, Academic Press, New York.

Tijms, H.C. and Eikeboom, A.M. (1986) A simple technique in Markovian control with applications to resource allocation in communication networks. *Operat. Res. Lett.*, **5**, 25–32.

Wijngaard, J. (1979) Decomposition for dynamic programming in production and inventory control. *Engng and Process Econom.*, **4**, 385–388.

CHAPTER 8

Advanced Renewal Theory

8.0 INTRODUCTION

A renewal process is a counting process that generalizes the Poisson process. In the Poisson process the interoccurrence times between the events are independent random variables with an *exponential* distribution, whereas in a renewal process the interoccurrence times have a *general* distribution. A first introduction to renewal theory has been already given in Section 2.1. In that section several limit theorems were given without proof. These limit theorems will be proved in Section 8.2 after having discussed the renewal function in more detail in Section 8.1. A key tool in proving the limit theorems is the so-called key renewal theorem. Section 8.3 deals with the alternating renewal model and gives an application of this model to a reliability problem. In queueing and insurance problems it is often important to have asymptotic estimates for the waiting-time probability and the ruin probability. In Section 8.4 such estimates are derived by using renewal-theoretic methods. This derivation illustrates the simplicity of analysis to be achieved by a general renewal-theoretic approach to hard individual problems.

8.1 THE RENEWAL FUNCTION

Let us first repeat some definitions and results that were given earlier in Section 2.1. The starting point is a sequence $X_1, X_2, \ldots$ of non-negative independent random variables having a common probability distribution function

$$F(x) = P\{X_k \leq x\}, \quad x \geq 0$$

for $k = 1, 2, \ldots$. Letting $\mu_1 = E(X_k)$, it is assumed that $0 < \mu_1 < \infty$. The random variable X_k denotes the interoccurrence time between the $(k-1)$th and kth events in some specific probability problem; see Section 2.1 for examples. Letting

$$S_0 = 0 \quad \text{and} \quad S_n = \sum_{i=1}^{n} X_i, \quad n = 1, 2, \ldots ,$$

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

we have that S_n is the epoch at which the nth event occurs. For each $t \geq 0$, let

$$N(t) = \text{the largest integer } n \geq 0 \text{ for which } S_n \leq t.$$

Then the random variable $N(t)$ represents the number of events up to time t. The counting process $\{N(t), t \geq 0\}$ is called the *renewal process* generated by the interoccurrence times $X_1, X_2, \ldots$. It is said that a renewal occurs at time t if $S_n = t$ for some n. Since $F(0) < 1$ the number of renewals up to time t is finite with probability 1 for any $t \geq 0$. The *renewal function* $M(t)$ is defined by

$$M(t) = E[N(t)], \quad t \geq 0.$$

For $n = 1, 2, \ldots$, define the probability distribution function $F_n(t)$ by

$$F_n(t) = P\{S_n \leq t\}, \quad t \geq 0.$$

The function $F_n(t)$ is the n-fold convolution of $F(t)$ with itself. Using the important observation that $N(t) \geq n$ if and only if $S_n \leq t$, it was shown in Section 2.1 that

$$E[N(t)] = \sum_{n=1}^{\infty} F_n(t), \quad t \geq 0. \tag{8.1.1}$$

Moreover, it was established in Section 2.1 that $M(t) < \infty$ for all $t \geq 0$. Another important quantity introduced in Section 2.1 is the *excess* or *residual* life at time t. This random variable is defined by

$$\gamma_t = S_{N(t)+1} - t$$

and denotes the waiting time from time t onwards until the first occurrence of an event after time t. Using Wald's equation, it was shown in Section 2.1 that

$$E(\gamma_t) = \mu_1\{1 + M(t)\} - t. \tag{8.1.2}$$

The following bounds apply to the renewal function:

$$\frac{t}{\mu_1} - 1 \leq M(t) \leq \frac{t}{\mu_1} + \frac{\mu_2}{\mu_1^2},$$

where $\mu_2 = E(X_1^2)$. The left inequality is an immediate consequence of (8.1.2) and the fact that $\gamma_t \geq 0$. The proof of the other inequality is demanding and lengthy. The interested reader is referred to Lorden (1970).

8.1.1 The Renewal Equation

A useful characterization of the renewal function is provided by the so-called *renewal equation.*

Theorem 8.1.1 *Assume that the probability distribution function $F(x)$ of the interoccurrence times has a probability density $f(x)$. Then the renewal function $M(t)$ satisfies the integral equation*

$$M(t) = F(t) + \int_0^t M(t-x) f(x)\, dx, \quad t \geq 0. \tag{8.1.3}$$

This integral equation has a unique solution that is bounded on finite intervals.

Proof The proof of (8.1.3) is instructive. Fix $t > 0$. To compute $E[N(t)]$, we condition on the time of the first renewal and use that the process probabilistically starts over after each renewal. Under the condition that $X_1 = x$, the random variable $N(t)$ is distributed as $1 + N(t-x)$ when $0 \leq x \leq t$ and $N(t)$ is 0 otherwise. Hence, by conditioning upon X_1, we find

$$E[N(t)] = \int_0^\infty E[N(t) \mid X_1 = x] f(x)\, dx = \int_0^t E[1 + N(t-x)] f(x)\, dx,$$

which gives (8.1.3). To prove that the equation (8.1.3) has a unique solution, suppose that $H(t) = F(t) + \int_0^t H(t-x) f(x)\, dx$, $t \geq 0$ for a function $H(t)$ that is bounded on finite intervals. We substitute this equation repeatedly into itself and use the convolution formula

$$F_n(t) = \int_0^t F(t-x) f_{n-1}(x)\, dx,$$

where $f_k(x)$ denotes the probability density of $F_k(x)$. This gives

$$H(t) = \sum_{k=1}^{n} F_k(t) + \int_0^t H(t-x) f_n(x)\, dx, \quad n = 1, 2, \ldots . \tag{8.1.4}$$

Fix now $t > 0$. Since $H(x)$ is bounded on $[0, t]$, the second term on the right-hand side of (8.1.4) is bounded by $cF_n(t)$ for some $c > 0$. Since $M(t) < \infty$, we have $F_n(t) \to 0$ as $n \to \infty$. By letting $n \to \infty$ in (8.1.4), we find $H(t) = \sum_{k=1}^{\infty} F_k(t)$ showing that $H(t) = M(t)$.

Theorem 8.1.1 allows for the following important generalization.

Theorem 8.1.2 *Assume that $F(x)$ has a probability density $f(x)$. Let $a(x)$ be a given, integrable function that is bounded on finite intervals. Suppose the function $Z(t)$, $t \geq 0$, is defined by the integral equation*

$$Z(t) = a(t) + \int_0^t Z(t-x) f(x)\, dx, \quad t \geq 0. \tag{8.1.5}$$

Then this equation has a unique solution that is bounded on finite intervals. The solution is given by

$$Z(t) = a(t) + \int_0^t a(t-x)m(x)\,dx, \quad t \geq 0, \tag{8.1.6}$$

where the renewal density $m(x)$ denotes the derivative of $M(x)$.

Proof We give only a sketch of the proof. The proof is similar to the proof of the second part of Theorem 8.1.1. Substituting the equation (8.1.5) repeatedly into itself yields

$$Z(t) = a(t) + \sum_{k=1}^{n} \int_0^t a(t-x) f_k(x)\,dx + \int_0^t Z(t-x) f_{n+1}(x)\,dx.$$

Next, by letting $n \to \infty$, the desired result readily follows. It is left to the reader to verify that the various mathematical operations are allowed.

The integral equation (8.1.5) is called the *renewal equation*. This important equation arises in many applied probability problems. As an application of Theorem 8.1.2, we derive an expression for the second moment of the excess life at time t.

Lemma 8.1.3 *Assuming that $\mu_2 = E(X_1^2)$ is finite,*

$$E(\gamma_t^2) = \mu_2[1 + M(t)] - 2\mu_1 \left[t + \int_0^t M(x)\,dx \right] + t^2, \quad t \geq 0. \tag{8.1.7}$$

Proof Fix $t \geq 0$. Given that the epoch of the first renewal is x, the random variable γ_t is distributed as γ_{t-x} when $x \leq t$ and γ_t equals $x - t$ otherwise. Thus

$$\begin{aligned} E(\gamma_t^2) &= \int_0^\infty E(\gamma_t^2 \mid X_1 = x) f(x)\,dx \\ &= \int_0^t E(\gamma_{t-x}^2) f(x)\,dx + \int_t^\infty (x-t)^2 f(x)\,dx. \end{aligned}$$

Hence, by letting $Z(t) = E(\gamma_t^2)$ and $a(t) = \int_t^\infty (x-t)^2 f(x)\,dx$, we obtain a renewal equation of the form (8.1.5). Next it is a question of tedious algebra to derive (8.1.7) from (8.1.6). The details of the derivation are omitted.

8.1.2 Computation of the Renewal Function

The following tools are available for the numerical computation of the renewal function:

(a) the series representation,

(b) numerical Laplace inversion,

(c) discretization of the renewal equation.

In Section 2.1.1 we have already seen that the renewal function can be directly computed from the series representation (8.1.1) when the interoccurrence times have a gamma distribution. If the interoccurrence times have a Coxian-2 distribution an explicit expression can be given for the renewal function; see Exercise 8.1. In general the renewal function $M(x)$ can be computed by numerical inversion of its Laplace transform. The Laplace transform $M^*(s) = \int_0^\infty e^{-sx} M(x)\,dx$ is given by

$$M^*(s) = \frac{f^*(s)}{s[1 - f^*(s)]},$$

where $f^*(s) = \int_0^\infty e^{-sx} f(x)\,dx$ denotes the Laplace transform of the probability density of the interoccurrence times; see Appendix E. How to proceed with numerical Laplace inversion is discussed in Appendix F. In this appendix it is also discussed how to proceed when the Laplace transform $f^*(s)$ is analytically intractable.

Next we discuss a simple but useful discretization method. The renewal equation (8.1.3) for $M(t)$ is a special case of an integral equation which is known in numerical analysis as a Volterra integral equation of the second kind. Many numerical methods have been proposed to solve such equations. Unfortunately, these methods typically suffer from the accumulation of round-off errors when t gets larger. However, using basic concepts from the theory of Riemann–Stieltjes integration, a simple and direct solution method with good convergence properties can be given for the renewal equation (8.1.3). This method discretizes the time and computes recursively the renewal function on a grid of points. For fixed $t > 0$, let $[0, t]$ be partitioned according to $0 = t_0 < t_1, < \ldots < t_n = t$, where $t_i = ih$ for a given grid size $h > 0$. Put for abbreviation

$$M_i = M(ih), \quad F_i = F((i - 0.5)h) \quad \text{and} \quad A_i = F(ih), \quad 1 \le i \le n.$$

The recursion scheme for computing the M_i is as follows:

$$M_i = \frac{1}{1 - F_1}\left[A_i + \sum_{j=1}^{i-1} (M_j - M_{j-1}) F_{i-j+1} - M_{i-1} F_1 \right], \quad 1 \le i \le n,$$

starting with $M_0 = 0$. This recursion scheme is a minor modification of the Riemann–Stieltjes method proposed in Xie (1989) (the original method uses F_i instead of A_i). The recursion scheme is easy to program and gives surprisingly accurate results. It is remarkable how well the recursion scheme is able to resist the accumulation of round-off errors as t gets larger. How to choose the grid size $h > 0$ depends not only on the desired accuracy in the answers, but also on the shape of the distribution function $F(x)$ and the length of the interval $[0, t]$. The usual way to find out whether the answers are accurate enough is to do the computations for both a grid size h and a grid size $h/2$. In many cases of practical interest a four-digit accuracy is obtained for a grid size h in the range $0.05 - 0.01$. In Table 8.1.1 some results are given for the renewal function of the Weibull distribution, where

Table 8.1.1 Renewal function for the Weibull distribution

$c_X^2 = 0.25$			$c_X^2 = 2$		
t	exact	asymp	t	exact	asymp
0.1	0.0061	−0.275	0.2	0.3841	0.700
0.2	0.0261	−0.175	0.5	0.7785	1.000
0.4	0.1087	0.025	1.0	1.357	1.500
0.6	0.2422	0.225	1.5	1.901	2.000
0.8	0.4141	0.425	2.0	2.428	2.500
1.0	0.6091	0.625	2.5	2.947	3.000
1.2	0.8143	0.825	3.0	3.460	3.500
1.5	1.124	1.125	3.5	3.969	4.000
2.0	1.627	1.625	5.0	5.485	5.500
2.5	2.125	2.125	7.5	7.995	8.000

a grid size $h = 0.02$ is used for the case $c_X^2 = 0.25$ and a grid size $h = 0.01$ for the case $c_X^2 = 2$. In both cases the normalization $\mu_1 = 1$ is used for the mean interoccurrence time. The table also gives the values of the asymptotic expansion of $M(x)$ that will be discussed in Section 8.2.

The discretization algorithm can also be used to solve an integral equation of the type (8.1.5). The only change is to replace $A_i = F(ih)$ by $A_i = a(ih)+a(0)F(ih)$. A more sophisticated discretization method for the renewal equation (8.1.5) is discussed in Den Iseger *et al.* (1997).

Computation of the distribution of $N(t)$

Numerical Laplace inversion can also be used to calculate the probability distribution of $N(t)$. Since the events $\{N(t) \geq n\}$ and $\{S_n \leq t\}$ are equivalent, we have $P\{N(t) \geq n\} = F_n(t)$ and so

$$P\{N(t) = n\} = F_n(t) - F_{n+1}(t), \quad n = 0, 1, \ldots,$$

where $F_0(t) = 1$ and $F_n(t) = P\{S_n \leq t\}$ for $n \geq 1$. Assuming that the probability distribution function of the interoccurrence times $X_1, X_2, \ldots$ has a probability density $f(t)$, the probability distribution function $F_n(t)$ of the sum $X_1 + \cdots + X_n$ has a probability density $f_n(t)$. The Laplace transform of this probability density is given by

$$\int_0^\infty e^{-st} f_n(t)\, dt = E\left(e^{-s(X_1+\cdots+X_n)}\right) = \left[f^*(s)\right]^n,$$

where $f^*(s) = \int_0^\infty e^{-sx} f(x)\, dx$ denotes the Laplace transform of $f(x)$. Using the relation (E.4) in Appendix E, we thus find

$$\int_0^\infty e^{-st} P\{N(t) = n\}\, dt = \frac{\left[f^*(s)\right]^n - \left[f^*(s)\right]^{n+1}}{s}, \quad n = 0, 1, \ldots$$

Hence for fixed n the probability $P\{N(t) = n\}$ can be calculated by numerical Laplace inversion. Another interesting question is how to compute the probability

$$\lim_{t\to\infty} P\{N(t + D) - N(t) = n\}, \quad n = 0, 1, \ldots$$

for a given constant D. Denote this probability by $a_n(D)$. Using the limiting distribution $P\{\gamma_t \le x\}$ from Theorem 8.2.5 in the next subsection and the relations (E.4) and (E.6) in Appendix E, it is not difficult for the reader to verify that

$$\int_0^\infty e^{-sx} a_0(x)\, dx = \frac{1}{s} - \frac{1 - f^*(s)}{\mu_1 s^2} \tag{8.1.8}$$

$$\int_0^\infty e^{-sx} a_n(x)\, dx = \left(\frac{1 - f^*(s)}{\mu_1 s}\right)\left(\frac{[f^*(s)]^{n-1} - [f^*(s)]^n}{s}\right), \quad n \ge 1. \tag{8.1.9}$$

This is a useful result. For example, the probability distribution $\{a_n(D)\}$ gives the limiting distribution of the number of busy servers in the infinite-server queue with renewal input and deterministic service times ($GI/D/\infty$ queue). This result is easily proved. Since each customer gets immediately assigned a free server upon arrival and the service time of each customer equals the constant D, the only customers present at time $t + D$ are those who have arrived in $(t, t + D]$.

8.2 ASYMPTOTIC EXPANSIONS

In Section 2.2 we proved a law of large numbers for the process $\{N(t)\}$:

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mu_1} \quad \text{with probability 1.} \tag{8.2.1}$$

The proof was elementary. It is tempting to conclude from (8.2.1) that $M(t)/t \to 1/\mu_1$ as $t \to \infty$. Although this result is correct, it cannot be directly concluded from (8.2.1). The reason is that the random variable $N(t)/t$ need not be bounded in t. For a sequence of unbounded random variables Y_n it is not necessarily true that $\lim_{n\to\infty} E(Y_n) = E(Y)$ when Y_n converges to Y with probability 1 as $n \to \infty$. Consider the counterexample in which $Y_n = 0$ with probability $1 - 1/n$ and $Y_n = n$ with probability $1/n$. Then $E(Y_n) = 1$ for all n, whereas Y_n converges to 0 with probability 1.

Theorem 8.2.1 (elementary renewal theorem)

$$\lim_{t\to\infty} \frac{M(t)}{t} = \frac{1}{\mu_1}.$$

Proof The proof will be based on the relation (8.1.2) for the excess variable. By the relation (8.1.2), we have $\mu_1[1+M(t)]-t \geq 0$ and so we obtain the inequality

$$\frac{M(t)}{t} \geq \frac{1}{\mu_1} - \frac{1}{t} \quad \text{for all } t > 0. \tag{8.2.2}$$

Next we prove that for any constant $c > 0$,

$$\frac{M(t)}{t} \leq \frac{1}{\mu(c)} + \frac{1}{t}\left(\frac{c}{\mu(c)} - 1\right) \quad \text{for all } t > 0. \tag{8.2.3}$$

where $\mu(c) = \int_0^c [1 - F(x)]\,dx$. To prove this inequality, fix $c > 0$ and consider the renewal process $\{\overline{N}(t)\}$ associated with the sequence $\{\overline{X}_n\}$, where

$$\overline{X}_n = \begin{cases} X_n & \text{if } X_n \leq c, \\ c & \text{if } X_n > c. \end{cases}$$

Since $N(t) \leq \overline{N}(t)$, we have $M(t) \leq \overline{M}(t)$ for all $t \geq 0$. For the renewal process $\{\overline{N}(t)\}$, the excess life $\overline{\gamma}_t$ satisfies $\overline{\gamma}_t \leq c$ for all t. Since $E(\overline{X}_1) = \int_0^\infty P\{\overline{X}_1 > x\}\,dx$, we have

$$E(\overline{X}_1) = \int_0^c \{1 - F(x)\}\,dx = \mu(c).$$

Thus, by (8.1.2),

$$\mu(c)\left[\overline{M}(t) + 1\right] - t \leq c, \quad t \geq 0.$$

This inequality in conjunction with $M(t) \leq \overline{M}(t)$ yields the inequality (8.2.3). The remainder of the proof is simple. Letting $t \to \infty$ in (8.2.2) and (8.2.3) gives

$$\frac{1}{\mu(c)} \geq \limsup_{t\to\infty} \frac{M(t)}{t} \geq \liminf_{t\to\infty} \frac{M(t)}{t} \geq \frac{1}{\mu_1}$$

for any constant $c > 0$. Next, by letting $c \to \infty$ and noting that $\mu(c) \to \mu_1$ as $c \to \infty$, we obtain the desired result.

So far our results have not required any assumption about the distribution function $F(x)$ of the interoccurrence times. However, in order to characterize the asymptotic behaviour of the solution to the renewal equation it is required that the distribution function $F(x)$ is non-arithmetic. The distribution function F is called *non-arithmetic* if the mass of F is not concentrated on a discrete set of points $0, \lambda, 2\lambda, \ldots$ for some $\lambda > 0$. A distribution function that has a positive density on some interval is non-arithmetic. In the discussion below we make for convenience the even stronger assumption that $F(x)$ has a probability density. To establish the limiting behaviour of the solution to the renewal equation (8.1.5), we need also to impose on the function $a(x)$ a stronger condition than integrability. It

must be required that the function $a(x)$ is directly Riemann integrable. Direct Riemann integrability can be characterized in several ways. A convenient definition is the following one. A function $a(x)$ defined on $[0, \infty)$ is said to be *directly Riemann integrable* when $a(x)$ is almost everywhere continuous and $\sum_{n=1}^{\infty} a_n < \infty$, where a_n is the supremum of $|a(x)|$ on the interval $[n-1, n)$. A sufficient condition for a function $a(x)$ to be directly Riemann integrable is that it can be written as a finite sum of monotone, integrable functions. This condition suffices for most applications.

Theorem 8.2.2 (key renewal theorem) *Assume $F(x)$ has a probability density $f(x)$. For a given function $a(t)$ that is bounded on finite intervals, let the function $Z(t)$ be defined by the renewal equation*

$$Z(t) = a(t) + \int_0^t Z(t-x) f(x)\, dx, \quad t \geq 0.$$

Suppose that $a(t)$ is directly Riemann integrable. Then

$$\lim_{t\to\infty} Z(t) = \frac{1}{\mu_1} \int_0^{\infty} a(x)\, dx.$$

The proof of this theorem is demanding and will not be given. The interested reader is referred to Feller (1971). Next we derive a number of useful results from the key renewal theorem.

Theorem 8.2.3 *Suppose $F(x)$ is non-arithmetic with $\mu_2 = E(X_1^2) < \infty$. Then*

$$\lim_{t\to\infty} \left[M(t) - \frac{t}{\mu_1} \right] = \frac{\mu_2}{2\mu_1^2} - 1, \tag{8.2.4}$$

$$\lim_{t\to\infty} \left[\int_0^t M(x)\, dx - \left\{ \frac{t^2}{2\mu_1} + \left(\frac{\mu_2}{2\mu_1^2} - 1 \right) t \right\} \right] = \frac{\mu_2^2}{4\mu_1^3} - \frac{\mu_3}{6\mu_1^2}, \tag{8.2.5}$$

provided that $\mu_3 = E(X_1^3) < \infty$.

Proof The asymptotic result $M(t)/t \to 1/\mu_1$ as $t \to \infty$ suggests that, for some constant c, $M(t) \approx t/\mu_1 + c$ for t large. Let us therefore define the function $Z_0(t)$ by

$$Z_0(t) = M(t) - \frac{t}{\mu_1}, \quad t \geq 0.$$

Assuming for ease that $F(x)$ has a density $f(x)$, we easily deduce from (8.1.3) that

$$Z_0(t) = a(t) + \int_0^t Z_0(t-x) f(x)\, dx, \quad t \geq 0, \tag{8.2.6}$$

where

$$a(t) = F(t) - \frac{t}{\mu_1} + \frac{1}{\mu_1}\int_0^t (t-x)f(x)\,dx, \quad t \geq 0.$$

Writing $\int_0^t (t-x)f(x)\,dx = \int_0^\infty (t-x)f(x)\,dx - \int_t^\infty (t-x)f(x)\,dx$, we find

$$a(t) = -[1 - F(t)] + \frac{1}{\mu_1}\int_t^\infty (x-t)f(x)\,dx.$$

This shows that $a(t)$ is the sum of two monotone, integrable functions. We have

$$\begin{aligned}\int_0^\infty a(t)\,dt &= -\int_0^\infty [1-F(t)]\,dt + \frac{1}{\mu_1}\int_0^\infty dt \int_t^\infty (x-t)f(x)\,dx \\ &= -\mu_1 + \frac{1}{\mu_1}\int_0^\infty f(x)\,dx \int_0^x (x-t)\,dt \\ &= -\mu_1 + \frac{1}{\mu_1}\int_0^\infty \frac{1}{2}x^2 f(x)\,dx \\ &= -\mu_1 + \frac{\mu_2}{2\mu_1}.\end{aligned}$$

By applying the key renewal theorem to (8.2.6), the result (8.2.4) follows. The proof of (8.2.5) proceeds along the same lines. The relation (8.2.4) suggests that, for some constant c,

$$\int_0^t M(x)\,dx \approx \frac{t^2}{2\mu_1} + t\left(\frac{\mu_2}{2\mu_1^2} - 1\right) + c \quad \text{for } t \text{ large.}$$

To determine the constant c, define the function

$$Z_1(t) = \int_0^t M(x)\,dx - \left[\frac{t^2}{2\mu_1} + t\left(\frac{\mu_2}{2\mu_1^2} - 1\right)\right], \quad t \geq 0.$$

By integrating both sides of the equation (8.1.3) over t and interchanging the order of integration, we get the following renewal equation for the function $U(x) = \int_0^x M(t)\,dt$:

$$U(t) = \int_0^t F(x)\,dx + \int_0^t U(t-x)f(x)\,dx, \quad t \geq 0.$$

From this renewal equation, we obtain after some algebra

$$Z_1(t) = a(t) + \int_0^t Z_1(t-x)f(x)\,dx, \quad t \geq 0, \tag{8.2.7}$$

where

$$a(t) = \frac{\mu_2}{2\mu_1^2} \int_t^\infty \{1 - F(x)\}\, dx + \frac{1}{\mu_1} \left[t \int_t^\infty \{1 - F(x)\}\, dx - \int_t^\infty x\{1 - F(x)\}\, dx \right].$$

The function $a(t)$ is the sum of two monotone functions. Each of the two terms is integrable. Using formula (A.8) in Appendix A, we find after some algebra

$$\int_0^\infty a(t)\, dt = \frac{\mu_2^2}{4\mu_1^2} - \frac{\mu_3}{6\mu_1}.$$

Next, by applying the key renewal theorem to (8.2.7), we obtain (8.2.5).

The asymptotic expansions in Theorem 8.2.3 are very useful. They are accurate for practical purposes already for moderate values of t. Asymptotic expansions for the second moment of $N(t)$ are discussed in Exercise 8.3. An immediate consequence of the relations (8.1.2) and (8.1.7) and Theorem 8.2.3 is the following result for the excess life γ_t.

Corollary 8.2.4 *Suppose $F(x)$ is non-arithmetic. Then*

$$\lim_{t\to\infty} E(\gamma_t) = \frac{\mu_2}{2\mu_1} \quad \textit{and} \quad \lim_{t\to\infty} E(\gamma_t^2) = \frac{\mu_3}{3\mu_1}.$$

Next we discuss the limiting distribution of the excess life γ_t for $t \to \infty$.

Theorem 8.2.5 *Suppose $F(x)$ is non-arithmetic. Then*

$$\lim_{t\to\infty} P\{\gamma_t \le x\} = \frac{1}{\mu_1} \int_0^x \{1 - F(y)\}\, dy, \quad x \ge 0. \tag{8.2.8}$$

Proof For fixed $u \ge 0$, define $Z(t) = P\{\gamma_t > u\}$, $t \ge 0$. By conditioning on the time of the first renewal, we derive a renewal equation for $Z(t)$. Since after each renewal the renewal process probabilistically starts over, it follows that

$$P\{\gamma_t > u \mid X_1 = x\} = \begin{cases} P\{\gamma_{t-x} > u\} & \text{if } x \le t, \\ 0 & \text{if } t < x \le t + u, \\ 1 & \text{if } x > t + u. \end{cases}$$

By the law of total probability,

$$P\{\gamma_t > u\} = \int_0^\infty P\{\gamma_t > u \mid X_1 = x\} f(x)\, dx.$$

This yields the renewal equation

$$Z(t) = 1 - F(t + u) + \int_0^t Z(t - x) f(x)\, dx, \quad t \ge 0.$$

The function $a(t) = 1 - F(t+u)$, $t \geq 0$ is monotone and integrable. By applying the key renewal theorem it now follows that

$$\lim_{t\to\infty} Z(t) = \frac{1}{\mu_1}\int_0^\infty \{1 - F(y+u)\}\,dy = \frac{1}{\mu_1}\int_u^\infty \{1 - F(y)\}\,dy,$$

yielding the desired result by using the fact that $\int_0^\infty \{1 - F(y)\}\,dy = \mu_1$.

In many practical applications the asymptotic expansion (8.2.8) gives a useful approximation to the distribution of γ_t already for moderate values of t. The limiting distribution of the excess life is called the *equilibrium excess distribution* and has applications in a wide variety of contexts. The equilibrium excess distribution can be given the following interpretation. Suppose that an outside person observes the state of the process at an arbitrarily chosen point in time when the process has been in operation for a very long time. Assuming that the outside person has *no* information about the past history of the process, the best prediction the person can give about the residual life of the item in use is according to the equilibrium excess distribution.

The asymptotic expansions in Theorem 8.2.3 will be illustrated by the next example.

Example 8.2.1 The D-policy for controlling the workload

Batches of fluid material arrive at a processing plant according to a Poisson process with rate λ. The batch amounts are independent random variables having a continuous probability distribution with finite first two moments μ_1 and μ_2. It is assumed that $\lambda\mu_1 < 1$. The unprocessed material is temporarily stored in an infinite-capacity buffer. If the processing plant is open, the material is processed at a unity rate. The plant is controlled by the so-called D-policy. If the inventory of unprocessed material becomes zero, the plant is temporarily closed down. The plant is reopened as soon as the buffer content exceeds the threshold value D. The set-up time to restart the processing is zero. The following costs are incurred. A holding cost at rate hx is incurred when the buffer content is x. A fixed set-up cost of $K > 0$ is incurred each time the plant is reopened. What value of the control parameter D minimizes the long-run average cost per time unit?

Preliminary analysis

To answer the above question, we first derive some preliminary results for the $M/G/1$ queue. Note that the control problem can be seen as an $M/G/1$ queue in which the workload is controlled. The workload is defined as the remaining amount of work for the server. Define the basic functions

$$t(x) = \text{the expected amount of time until the workload is zero when the current workload is } x \text{ and the server is working}$$

and

$$h(x) = \text{the expected holding costs incurred until the workload is zero when the current workload is } x \text{ and the server is working.}$$

The functions $t(x)$ and $h(x)$ are given by

$$t(x) = \frac{x}{1-\lambda\mu_1} \quad \text{and} \quad h(x) = \frac{h}{2(1-\lambda\mu_1)}\left(x^2 + \frac{\lambda x \mu_2}{1-\lambda\mu_1}\right) \tag{8.2.9}$$

for $x \geq 0$. The proof is as follows. By conditioning on the number of arrivals during a time x, it follows that

$$t(x) = x + \sum_{n=1}^{\infty} e^{-\lambda x}\frac{(\lambda x)^n}{n!} t_n, \quad x \geq 0,$$

where t_n is defined as the expected amount of time needed to empty the system when service is begun with n batches (= customers) present. Let us also define h_n as the expected holding cost incurred during the time needed to empty the system when service is begun with n batches present. Then, using relation (1.1.8),

$$\begin{aligned} h(x) &= \frac{h}{2}x^2 + \sum_{n=1}^{\infty} e^{-\lambda x}\frac{(\lambda x)^n}{n!}\left[h\sum_{k=1}^{n}\left(x - \frac{kx}{n+1}\right)\mu_1 + h_n\right] \\ &= \frac{h}{2}x^2 + h\frac{\lambda}{2}x^2\mu_1 + \sum_{n=1}^{\infty} e^{-\lambda x}\frac{(\lambda x)^n}{n!}h_n, \quad x \geq 0. \end{aligned}$$

The formula $t_n = n\mu_1/(1-\lambda\mu_1)$ was obtained in Section 2.6. Substituting this into the above relation for $t(x)$ gives the first relation in (8.2.9). By the same arguments as used in Section 2.6 to obtain t_n, we find

$$h_n = \sum_{k=1}^{n}\{h_1 + (n-k)t_1 h\mu_1\} = nh_1 + \frac{1}{2}h\mu_1 n(n-1)t_1.$$

Substituting this into the relation for $h(x)$ gives

$$h(x) = \frac{h}{2}x^2 + h\frac{\lambda}{2}x^2\mu_1 + \lambda x h_1 + \frac{1}{2}h\mu_1(\lambda x)^2 t_1, \quad x \geq 0.$$

Integrating both sides of this equation over the probability density $f(x)$ of the batch size and noting that $h_1 = \int_0^\infty h(x)f(x)\,dx$, we find an explicit expression for h_1 and next we obtain the second relation in (8.2.9). The details are left to the reader.

Analysis of the D-policy

For a given D-policy the stochastic process describing jointly the inventory of unprocessed material and the status of the plant (on or off) is regenerative. The epochs at which the plant is closed down are regeneration epochs. Define a cycle as the time elapsed between two consecutive shutdowns. The long-run average cost per time unit equals the value of E(cost incurred during one cycle) divided by the value of E(expected length of one cycle). To find these two expected values, we define the functions

$$\alpha(x) = E(\text{time until the buffer content exceeds the level } D \text{ when the current buffer content is } D - x \text{ and there is no processing}),$$

$$\beta(x) = E(\text{holding costs incurred until the buffer content exceeds the level } D \text{ when the current buffer content is } D - x \text{ and there is no processing})$$

for $0 \le x \le D$. In particular, $\alpha(D)$ and $\beta(D)$ denote the expected length of the idle period in a cycle and the expected holding costs during that idle period. Also, define the random variable γ_D as the excess of the inventory over the level D when the plant is reopened. Then $E\left[t(D+\gamma_D)\right]$ and $E\left[h(D+\gamma_D)\right]$ represent the expected length of the busy period in a cycle and the expected holding cost incurred during that busy period. Thus, under a given D-policy,

$$\text{the long-run average cost per time unit} = \frac{\beta(D) + K + E[h(D+\gamma_D)]}{\alpha(D) + E[t(D+\gamma_D)]}$$

with probability 1. It remains to find $\alpha(D)$ and $\beta(D)$. By conditioning on the batch size,

$$\alpha(x) = \frac{1}{\lambda} + \int_0^x \alpha(x-y) f(y)\, dy, \quad 0 \le x \le D,$$

$$\beta(x) = \frac{(D-x)h}{\lambda} + \int_0^x \beta(x-y) f(y)\, dy, \quad 0 \le x \le D,$$

where $f(y)$ denotes the density of the batch size. Let $M(x)$ denote the renewal function in the renewal process in which the interoccurrence times have the batch-size density $f(x)$. Denote by $m(x)$ the density of $M(x)$. Then it follows from Theorem 8.1.2 that

$$\alpha(x) = \frac{1}{\lambda} + \int_0^x \frac{1}{\lambda} m(y)\, dy = \frac{1}{\lambda}\{1 + M(x)\}, \quad 0 \le x \le D$$

$$\beta(x) = \frac{(D-x)h}{\lambda} + \frac{h}{\lambda}\int_0^x (D - x + y) m(y)\, dy$$

$$= \frac{(D-x)h}{\lambda} + \frac{h}{\lambda} D M(x) - \frac{h}{\lambda}\int_0^x M(y)\, dy, \quad 0 \le x \le D.$$

Table 8.2.1 Approximate and exact values for D

	$\lambda = 0.5$			$\lambda = 0.8$		
c_B^2	D_{app}	D_{opt}	error (%)	D_{app}	D_{opt}	error (%)
$\frac{1}{3}$	2.911	2.911	0.00	2.214	2.214	0.00
$\frac{1}{2}$	2.847	2.847	0.00	2.155	2.155	0.00
$1\frac{1}{2}$	2.259	2.298	0.13	1.545	1.629	0.24
3	1.142	1.588	11.9	0.318	1.049	15.6

Using this result and the formulas (8.2.9), (8.1.7) and (8.1.2), the above expression for the long-run average cost can be worked out as

$$g(D) = \frac{K\lambda(1-\lambda\mu_1) - h\left[D + \int_0^D M(y)\,dy\right]}{1 + M(D)} + hD + \frac{h\lambda\mu_2}{2(1-\lambda\mu_1)}.$$

The function $g(D)$ is minimal for the unique solution of the equation

$$D + \int_0^D M(y)\,dy = \frac{K\lambda(1-\lambda\mu_1)}{h}. \tag{8.2.10}$$

In general it is computationally demanding to find an exact solution of this equation. Except for special cases, one needs numerical Laplace inversion to compute $\int_0^x M(y)\,dy$; see Appendix F. However, an approximate solution to (8.2.10) is easily calculated when it is assumed that the optimal value of D is sufficiently large compared to μ_1. Then, by Theorem 8.2.3,

$$\int_0^D M(y)\,dy \approx \frac{D^2}{2\mu_1} + D\left(\frac{\mu_2}{2\mu_1^2} - 1\right) + \frac{\mu_2^2}{4\mu_1^3} - \frac{\mu_3}{6\mu_1^2}.$$

Table 8.2.1 gives for several examples the optimal value D_{opt} and the approximate value D_{app} together with the relative error $100\times\left[g(D_{app}) - g(D_{opt})/g(D_{opt})\right]$. In all examples we take $\mu_1 = 1$, $h = 1$ and $K = 25$. The arrival rate λ is 0.5 and 0.8. The squared coefficient of variation c_B^2 of the batch size is $\frac{1}{3}$, $\frac{1}{2}$, $1\frac{1}{2}$ and 3, where the first two values correspond to an Erlang distribution and the latter two values to an H_2 distribution with balanced means. Can you give a heuristic explanation why the optimal value of D decreases when the coefficient of variation of the batch size increases?

8.3 ALTERNATING RENEWAL PROCESSES

An alternating renewal process is a two-state process alternating between an on-state and an off-state. The on-times and the off-times are independent and identically distributed random variables. The two sequences of on-times and off-times are

mutually independent. For any $s > 0$, let

$$P_{on}(s) = P\{\text{the process is in the on-state at time } s\}$$

and

$$U(s) = P\{\text{the amount of time the process is in the on-state during } [0, s]\}.$$

Theorem 8.3.1 *Suppose that the on-times and off-times have exponential distributions with respective means $1/\alpha$ and $1/\beta$. Then, assuming that an on-time starts at epoch 0,*

$$P_{on}(s) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta} e^{-(\alpha+\beta)s}, \quad s \geq 0 \tag{8.3.1}$$

and

$$P\{U(s) \leq x\} = \sum_{n=0}^{\infty} e^{-\beta(s-x)} \left[\frac{\beta(s-x)}{n!}\right]^n \left[1 - \sum_{k=0}^{n} e^{-\alpha x} \frac{(\alpha x)^k}{k!}\right], \quad 0 \leq x < s. \tag{8.3.2}$$

The distribution function $P\{U(s) \leq x\}$ has a mass of $e^{-\alpha s}$ at $x = s$.

Proof Let $P_{off}(s) = P\{\text{the process is in the off-state at time } s\}$. By considering what may happen in the time interval $(s, s+\Delta s]$ with Δs small, it is straightforward to derive the linear differential equation

$$P'_{on}(s) = \beta P_{off}(s) - \alpha P_{on}(s), \quad s > 0.$$

Since $P_{off}(s) = 1 - P_{on}(s)$, we find $P'_{on}(s) = \beta - (\alpha+\beta)P_{on}(s)$, $s > 0$. The solution of this equation is given by (8.3.1). The proof of (8.3.2) is more complicated. The random variable $U(s)$ is equal to s only if the first on-time exceeds s. Hence $P\{U(s) \leq x\}$ has mass $e^{-\alpha s}$ at $x = s$. Now fix $0 \leq x < s$. By conditioning on the lengths of the first on-time and the first off-time, we obtain

$$P\{U(s) \leq x\} = \int_0^x \alpha e^{-\alpha y}\, dy \int_0^{\infty} P\{U(s-y-u) \leq x-y\} \beta e^{-\beta u}\, du.$$

Noting that $P\{U(s-y-u) \leq x-y\} = 1$ if $s-y-u \leq x-y$, we next obtain

$$P\{U(s) \leq x\} = e^{-\beta(s-x)}(1 - e^{-\alpha x}) + \int_0^x \alpha e^{-\alpha y}\, dy \int_0^{s-x} P\{U(s-y-u) \leq x-y\} \beta e^{-\beta u}\, du.$$

Substituting this equation repeatedly into itself leads to the desired result (8.3.2).

Corollary 8.3.2 *Suppose that the on-times and off-times have exponential distributions with respective means* $1/\alpha$ *and* $1/\beta$. *Then, for any* $t_0 > 0$ *and* $0 \le x < t_0$,

$$\lim_{t\to\infty} P\{U(t+t_0) - U(t) \le x\}$$
$$= \frac{\beta}{\alpha+\beta} \sum_{n=0}^{\infty} e^{-\beta(t_0-x)} \frac{[\beta(t_0-x)]^n}{n!} \left[1 - \sum_{k=0}^{n} e^{-\alpha x} \frac{(\alpha x)^k}{k!}\right]$$
$$+ \frac{\alpha}{\alpha+\beta} \sum_{n=0}^{\infty} e^{-\beta(t_0-x)} \frac{[\beta(t_0-x)]^n}{n!} \left[1 - \sum_{k=0}^{n-1} e^{-\alpha x} \frac{(\alpha x)^k}{k!}\right]. \quad (8.3.3)$$

Proof Since $\lim_{t\to\infty} P_{on}(t) = \beta/(\alpha+\beta)$, it follows that

$$\lim_{t\to\infty} P\{U(t+t_0) - U(t) \le x\}$$
$$= \frac{\beta}{\alpha+\beta} P\{U(t_0) \le x\}$$
$$+ \frac{\alpha}{\alpha+\beta} \left[\int_0^{t_0-x} P\{U(t_0-y) \le x\} \beta e^{-\beta y}\, dy + \int_{t_0-x}^{\infty} \beta e^{-\beta y}\, dy\right].$$

Next it is a matter of algebra to obtain the desired result from (8.3.2).

Exercises 8.4 to 8.8 give results for the alternating renewal process with non-exponential on- and off-times. The alternating renewal process is particularly useful in reliability applications. This is illustrated by the next example.

Example 8.3.1 The 1-out-of-2 reliability model with repair

The 1-out-of-2 reliability model deals with a repairable system that has one operating unit and one cold standby unit as protection against failures. The lifetime of an operating unit has a general probability distribution function $F_L(x)$ having density $f_L(x)$ with mean μ_L. If the operating unit fails, it is replaced immediately by the standby unit if available. The failed unit is sent to a repair facility and immediately enters repair if the facility is idle. Only one unit can be in repair at a time. The repair time of a failed unit has a general probability distribution function $G_R(x)$ with mean μ_R. It is assumed that $\mu_R << \mu_L$. The operating times and repair times are mutually independent. The system is down when both units are broken down and is up otherwise.

We are interested in the probability distribution function

$$A(x, t_0) = \lim_{t\to\infty} P\{\text{the total uptime in } (t, t+t_0] \text{ is } \le x\}$$

for an interval of length t_0. In other words, the performance measure is the probability distribution function of the total amount of time the system is available during a time interval of given length t_0 when the system has reached statistical

equilibrium. An approximate analysis will be given. The analysis is based on the following ideas:

1. Compute the means of the up- and down-periods.
2. Approximate the stochastic process of the up- and down-periods by an alternating renewal process in which both the up-periods and the down-periods are independent, exponential random variables and the up-periods are independent of the down-periods.

In view of the assumption $\mu_R << \mu_L$, the occurrence of a system failure is a rare event. This justifies the approximate step of assuming an exponential distribution for the up-period; see also the discussion on rare events at the end of Section 2.2. A similar justification for approximating the distribution of the downtime by an exponential distribution cannot be given. However, in view of the fact that the uptime dominates the downtime, it is reasonable to expect that the distributional form of the downtime has only a minor effect on the accuracy of the approximation. The process alternates between the up-state and the down-state. With the possible exception of the first up-period, the up-periods start when a unit is put into operation while the other unit enters repair. The system regenerates itself at the beginning of those up-periods. We assume that epoch 0 is such a regeneration epoch. Let the random variables τ_{up} and τ_{down} denote the lengths of an up-period and a down-period. Denote by the sequences $\{L_i\}$ and $\{R_i\}$ the successive operating times and the successive repair times. Then

$$E(\tau_{up}) = E\left[\sum_{i=1}^{N} L_i\right],$$

where $N = \min\{n \geq 1 \mid R_n > L_n\}$. The event $\{N = n\}$ is independent of $L_{n+1}, L_{n+2}, \ldots$ for any $n \geq 1$. Thus, by Wald's equation, $E(\tau_{up}) = E(N)\mu_L$. Let

$$q = P\{R > L\}$$

where the random variables L and R denote the operating time and the repair time of a unit. Since $P\{N = n\} = (1 - q)^{n-1}q$ for $n \geq 1$, we find

$$E(\tau_{up}) = \frac{\mu_L}{q}.$$

By conditioning on the lifetime, we have

$$q = \int_0^\infty \{1 - G_R(x)\} f_L(x)\, dx.$$

To find $E(\tau_{down})$, note that $E(\tau_{down}) = E(R - L \mid R > L)$. Using the formula

(A.7) in Appendix A, we find

$$E(\tau_{down}) = \int_0^\infty P\{R - L > t \mid R > L\}\, dt$$
$$= \frac{1}{q} \int_0^\infty \left[\int_0^\infty \{1 - G_R(x+t)\} f_L(x)\, dx \right] dt$$
$$= \frac{1}{q} \int_0^\infty f_L(x) \left[\int_x^\infty \{1 - G_R(u)\}\, du \right] dx,$$

where the latter equality uses an interchange of the order of integration. Interchanging again the order of integration, we next find that

$$E(\tau_{down}) = \frac{1}{q} \int_0^\infty \{1 - G_R(u)\} F_L(u)\, du.$$

We are now in a position to calculate an approximation for the probability distribution function of the total uptime in a time interval of given length t_0 when the system has reached statistical equilibrium. An approximation to the desired probability $A(x, t_0)$ is obtained by applying formula (8.3.3) in which $1/\alpha$ and $1/\beta$ are replaced by $E(\tau_{up})$ and $E(\tau_{down})$ respectively. The numerical evaluation of the right-hand side of (8.3.3) is easy, since the infinite series converges rapidly and involves only Poisson probabilities. Numerical integration is required to calculate the integrals for $E(\tau_{up})$ and $E(\tau_{down})$. It remains to investigate the quality of the approximation for the probabilities $A(x, t_0)$. Several assumptions have been made to get the approximation. The most serious weakness of the approximation is the assumption that the off-time is approximately exponentially distributed. Nevertheless it turns out that the approximation performs very well for practical purposes. Denoting by D_x the probability that the fraction of time the system is unavailable in the time interval of length t_0 is more than $x\%$, Table 8.3.1 gives the approximate and exact values of D_x for several values of x. Note that $D_x = A(1 - t_0 x/100, t_0)$.

Table 8.3.1 The unavailability probabilities

		$c_L^2 = 0.5$				$c_L^2 = 1$			
		D_0	D_2	D_5	D_{10}	D_0	D_2	D_5	D_{10}
$c_R^2 = 0$	app	0.044	0.030	0.016	0.006	0.117	0.086	0.054	0.024
	sim	0.043	0.033	0.020	0.005	0.108	0.091	0.066	0.027
$c_R^2 = 0.5$	app	0.051	0.040	0.028	0.015	0.117	0.095	0.068	0.040
	sim	0.050	0.040	0.029	0.016	0.109	0.092	0.070	0.042
$c_R^2 = 1$	app	0.056	0.047	0.036	0.024	0.117	0.099	0.077	0.050
	sim	0.055	0.047	0.036	0.024	0.110	0.094	0.074	0.050
$c_R^2 = 4$	app	0.076	0.071	0.063	0.053	0.117	0.108	0.096	0.079
	sim	0.075	0.069	0.061	0.050	0.112	0.101	0.089	0.072

The exact values of D_x are obtained by computer simulation. The length of the simulation run has been taken long enough to ensure that the half-width of the 95% confidence interval for the simulated probability is no more than 0.001. The lifetime L of a unit has a Weibull distribution with mean $E(L) = 1$ and the repair time R of a unit has a gamma distribution with mean $E(R) = 0.125$. The squared coefficients of variation of the lifetime and the repair time are $c_L^2 = 0.5$, 1 and $c_R^2 = 0$, 0.5, 1, 4. For the length of the interval we have taken $t_0 = 1$.

8.4 RUIN PROBABILITIES

In many applied probability problems asymptotic expansions provide a simple alternative to computationally intractable solutions. A nice example is the ruin probability in risk theory. Suppose claims arrive at an insurance company according to a Poisson process $\{N(t)\}$ with rate λ. The successive claim amounts $X_1, X_2, \ldots$ are positive, independent random variables having a common probability distribution function $B(x)$ with finite mean μ. The claim amounts are independent of the arrival process. In the absence of claims, the company's reserve increases at a constant rate of $\sigma > 0$ per time unit. It is assumed that $\sigma > \lambda\mu$, i.e. the average premium received per time unit is larger than the average claim rate. Denote by the compound Poisson variable

$$X(t) = \sum_{k=1}^{N(t)} X_k$$

the total amount claimed up to time t. If the company's initial reserve is $x > 0$, then the company's total reserve at time t is $x + \sigma t - X(t)$. We say that a ruin occurs at time t if $x + \sigma t - X(t) < 0$ and $x + \sigma u - X(u) \geq 0$ for $u < t$. Let

$$Q(x) = P\{X(t) > x + \sigma t \text{ for some } t \geq 0\}.$$

Then $Q(x)$ is the probability that a ruin will ever occur when the initial capital is x. Since a ruin can occur only at the claim epochs, we can equivalently write

$$Q(x) = P\left\{\sum_{j=1}^{k} X_j - \sigma T_k > x \text{ for some } k \geq 1\right\}, \tag{8.4.1}$$

where T_k is the epoch at which the kth claim occurs for $k = 1, 2, \ldots$. We are interested in the asymptotic behaviour of $Q(x)$ for large x.

The ruin probability $Q(x)$ arises in a variety of contexts. As another example consider a production/inventory situation in which demands for a given product arrive according to a Poisson process. The successive demands are independent and identically distributed random variables. On the other hand, inventory replenishments of the product occur at a constant rate of $\sigma > 0$ per time unit. In this context, the ruin probability $Q(x)$ represents the probability that a shortage will ever occur when the initial inventory is x.

The ruin probability as waiting-time probability

A less obvious context in which the ruin probability appears is the $M/G/1$ queue. Customers arrive at a single server station according to a Poisson process with rate λ. The service or work requirements of the successive customers are independent random variables having a common probability distribution function $B(x)$ with finite mean μ. The server works at a rate of $\sigma > 0$. It is assumed that $\sigma > \lambda\mu$. For $n = 1, 2, \ldots$ define the random variable D_n by

$$D_n = \text{the delay in queue of the } n\text{th customer (excluding service time).}$$

Assuming that service is in order of arrival, $\lim_{n\to\infty} P\{D_n \le x\}$ exists for all x. Moreover, letting

$$W_q(x) = \lim_{n\to\infty} P\{D_n \le x\},$$

it holds that

$$W_q(x) = 1 - Q(\sigma x), \quad x \ge 0. \tag{8.4.2}$$

A proof of these statements goes as follows. Let τ_n denote the time between the arrival of the nth and $(n+1)$th customers for $n = 1, 2, \ldots$ with the convention that the 0th customer arrives at epoch 0. Then

$$D_{n+1} = \begin{cases} D_n + X_n/\sigma - \tau_n & \text{if } D_n + X_n/\sigma - \tau_n \ge 0, \\ 0 & \text{if } D_n + X_n/\sigma - \tau_n < 0. \end{cases}$$

Hence, letting $U_n = X_n/\sigma - \tau_n$ for $n \ge 1$, we have

$$D_{n+1} = \max(0, D_n + U_n).$$

Substituting this equation in itself, it follows that

$$\begin{aligned} D_{n+1} &= \max\{0, U_n + \max(0, D_{n-1} + U_{n-1})\} \\ &= \max(0, U_n, U_n + U_{n-1} + D_{n-1}), \quad n \ge 1. \end{aligned}$$

By a repeated application of this equation and by $D_1 = 0$, we find

$$\begin{aligned} &\max(0, U_n, U_n + U_{n-1} + D_{n-1}) \\ &\quad = \max(0, U_n, U_n + U_{n-1}, \ldots, U_n + U_{n-1} + \cdots + U_1), \quad n \ge 1. \end{aligned}$$

Since the random variables $U_1, U_2, \ldots$ are independent and identically distributed, $(U_n, \ldots, U_1)$ has the same joint distribution as $(U_1, \ldots, U_n)$. Thus

$$D_{n+1} = \max(0, U_1, U_1 + U_2, \ldots, U_1 + \cdots + U_n), \quad n \ge 1.$$

This implies that

$$P\{D_{n+1} > x\} = P\left\{\sum_{j=1}^{k} U_j > x \text{ for some } 1 \le k \le n\right\}, \quad x \ge 0.$$

Since $\lim_{n\to\infty} P\{E_n\} = P\{\lim_{n\to\infty} E_n\}$ for any monotone sequence $\{E_n\}$ of events, it follows that $\lim_{n\to\infty} P\{D_n > x\}$ exists for all $x \ge 0$. Moreover,

$$\lim_{n\to\infty} P\{D_n > x\} = P\left\{\sum_{j=1}^{k} X_j - \sigma \sum_{j=1}^{k} \tau_j > \sigma x \text{ for some } k \ge 1\right\}, \quad x \ge 0.$$

Together this relation and (8.4.1) prove the result (8.4.2).

A renewal equation for the ruin probability

We now turn to the determination of the ruin probability $Q(x)$. For that purpose, we derive first an integro-differential equation for $Q(x)$. For ease of presentation we assume that the probability distribution function $B(x)$ of the claim sizes has a probability density $b(x)$. Fix $x > 0$. To compute $Q(x - \Delta x)$ with Δx small, we condition on what may happen in the first $\Delta t = \Delta x/\sigma$ time units. In the absence of claims, the company's capital grows from $x - \Delta x$ to x. However, since claims arrive according to a Poisson process with rate λ, a claim occurs in the first $\Delta x/\sigma$ time units with probability $\lambda \Delta x/\sigma + o(\Delta x)$, in which case the company's capital becomes $x - S$ if S is the size of that claim. A ruin occurs if $S > x$. Thus, by conditioning, we get for fixed $x > 0$,

$$\begin{aligned} Q(x - \Delta x) = \left(1 - \frac{\lambda \Delta x}{\sigma}\right) Q(x) + \frac{\lambda \Delta x}{\sigma} \int_x^\infty b(y)\, dy \\ + \frac{\lambda \Delta x}{\sigma} \int_0^x Q(x - y) b(y)\, dy + o(\Delta x). \end{aligned}$$

Subtracting $Q(x)$ from both sides of this equation, dividing by $h = -\Delta x$ and letting $\Delta x \to 0$, we obtain the integro-differential equation

$$Q'(x) = -\frac{\lambda}{\sigma}\{1 - B(x)\} + \frac{\lambda}{\sigma} Q(x) - \frac{\lambda}{\sigma} \int_0^x Q(x - y) b(y)\, dy, \quad x > 0. \quad (8.4.3)$$

Equation (8.4.3) can be converted into an integral equation of the renewal type. To do so, note that

$$\begin{aligned} &\frac{d}{dx} \int_0^x Q(x - y)\{1 - B(y)\}\, dy \\ &\quad = Q(0)\{1 - B(x)\} + \int_0^x Q'(x - y)\{1 - B(y)\}\, dy \end{aligned}$$

$$= Q(0)\{1 - B(x)\} - Q(x - y)\{1 - B(y)\}\Big|_0^x - \int_0^x Q(x - y)b(y)\,dy$$

$$= Q(x) - \int_0^x Q(x - y)b(y)\,dy.$$

Hence (8.4.3) can be rewritten as

$$Q'(x) = -\frac{\lambda}{\sigma}\{1 - B(x)\} + \frac{\lambda}{\sigma}\frac{d}{dx}\int_0^x Q(x - y)\{1 - B(y)\}\,dy \tag{8.4.4}$$

for $x > 0$. Integrating both sides of this equation gives

$$Q(x) = Q(0) - \frac{\lambda}{\sigma}\int_0^x \{1 - B(y)\}\,dy + \frac{\lambda}{\sigma}\int_0^x Q(x - y)\{1 - B(y)\}\,dy \tag{8.4.5}$$

for all $x \geq 0$. The unknown constant $Q(0)$ is easily determined by taking the Laplace transforms of both sides of (8.4.5). Using the relations (E.5), (E.6) and (E.7) in Appendix E and noting $\lim_{x\to\infty} Q(x) = 0$, it is readily verified that

$$Q(0) = \lambda\mu/\sigma,$$

where $\mu = E(X)$ is the mean claim size. The details are left to the reader. Hence the integro-differential equation (8.4.3) is equivalent to

$$Q(x) = a(x) + \int_0^x Q(x - y)h(y)\,dy, \quad x \geq 0, \tag{8.4.6}$$

where the functions $a(x)$ and $h(x)$ are given by

$$a(x) = Q(0) - \frac{\lambda}{\sigma}\int_0^x \{1 - B(y)\}\,dy \quad \text{and} \quad h(x) = \frac{\lambda}{\sigma}\{1 - B(x)\}, \quad x \geq 0.$$

The equation (8.4.6) has the form of a standard renewal equation except that the function $h(x)$, $x \geq 0$, is not a proper probability density. It is true that the function h is non-negative, but

$$\int_0^\infty h(x)\,dx = \frac{\lambda}{\sigma}\int_0^\infty \{1 - B(x)\}\,dx = \frac{\lambda\mu}{\sigma} < 1.$$

Thus h is the density of a distribution whose total mass is less than 1 with a defect of $1-\lambda\mu/\sigma$. Equation (8.4.6) is called a *defective renewal equation*.

Asymptotic expansion for the ruin probability

A very useful asymptotic expansion of $Q(x)$ can be given when it is assumed that the probability density of the claim size (service time) is not heavy-tailed. To be more precise, the following assumption is made.

Assumption 8.4.1 *There are positive numbers a and b such that the complementary distribution function* $1 - B(y) \leq ae^{-by}$ *for all y sufficiently large.*

This assumption excludes probability distributions with long tails like the log-normal distribution. The assumption implies that the number s_0 defined by

$$s_0 = \sup\left\{s \mid \int_0^\infty e^{sy}\{1 - B(y)\}\,dy < \infty\right\}$$

exists and is positive (possibly $s_0 = \infty$). In addition to Assumption 8.4.1 we make the technical assumption

$$\lim_{s \to s_0} \frac{\lambda}{\sigma} \int_0^\infty e^{sy}\{1 - B(y)\}\,dy > 1.$$

Then it is readily verified that the equation

$$\frac{\lambda}{\sigma} \int_0^\infty e^{\delta y}\{1 - B(y)\}\,dy = 1 \tag{8.4.7}$$

has a unique solution δ on the interval $(0, s_0)$. Next we convert the defective renewal equation (8.4.6) into a standard renewal equation. This enables us to apply the key renewal theorem to obtain the asymptotic behaviour of $Q(x)$. Let

$$h^*(x) = \frac{\lambda}{\sigma} e^{\delta x}\{1 - B(x)\}, \quad x \geq 0.$$

Then $h^*(x)$, $x \geq 0$ is a probability density with finite mean. Multiplying both sides of equation (8.4.6) by $e^{\delta x}$ and defining the functions

$$Q^*(x) = e^{\delta x} Q(x) \quad \text{and} \quad a^*(x) = e^{\delta x} a(x), \quad x \geq 0,$$

we find that the defective renewal function (8.4.6) is equivalent to

$$Q^*(x) = a^*(x) + \int_0^x Q^*(x - y) h^*(y)\,dy, \quad x \geq 0. \tag{8.4.8}$$

This is a standard renewal equation to which we can apply the key renewal theorem. The function $a^*(x)$ is directly Riemann integrable as can be shown by verifying that $|a^*(x)| \leq ce^{-(a-\delta)x}$ as $x \to \infty$ for finite constants $c > 0$ and $a > \delta$. Using definition (8.4.7) for δ and the relation $\int_0^\infty \{1 - B(y)\}\,dy = \mu$, we find

$$\begin{aligned}
\int_0^\infty a^*(x)\,dx &= \int_0^\infty e^{\delta x}\left[\frac{\lambda}{\sigma}\int_x^\infty \{1 - B(y)\}\,dy\right]dx \\
&= \frac{\lambda}{\sigma}\int_0^\infty \{1 - B(y)\}\left[\int_0^y e^{\delta x}\,dx\right]dy \\
&= \frac{\lambda}{\delta\sigma}\int_0^\infty \left(e^{\delta y} - 1\right)\{1 - B(y)\}\,dy = \frac{1 - \rho}{\delta},
\end{aligned}$$

where the load factor ρ is defined by $\rho = \lambda\mu/\sigma$. Applying the key renewal theorem from Section 8.2 to the renewal equation (8.4.8), we find

$$\lim_{x\to\infty} Q^*(x) = \gamma,$$

where the constant γ is given by

$$\gamma = \frac{(1-\rho)}{\delta}\left[\frac{\lambda}{\sigma}\int_0^\infty ye^{\delta y}\{1-B(y)\}\,dy\right]^{-1}.$$

This yields the asymptotic expansion

$$Q(x) \sim \gamma e^{-\delta x} \quad \text{as } x \to \infty, \tag{8.4.9}$$

where $f(x) \sim g(x)$ as $x \to \infty$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. This is an extremely important result. The asymptotic expansion is very useful for practical purposes in view of the remarkable finding that already for relatively small values of x the asymptotic estimate predicts quite well the exact value of $Q(x)$ when the load factor ρ is not very small. To illustrate this, Table 8.4.1 gives the numerical values of $Q(x)$ and the asymptotic estimate $Q_{asy}(x) = \gamma e^{-\delta x}$ for several examples. We take $\mu = 1$ and $\sigma = 1$. The squared coefficient of variation c_X^2 of the claim size X is $c_X^2 = 0$ (deterministic distribution), $c_X^2 = 0.5$ (E_2 distribution) and $c_X^2 = 1.5$ (H_2 distribution with balanced means). The load factor ρ is 0.2, 0.5 and 0.8. It turns out that the closer ρ is to 1, the earlier the asymptotic expansion applies.

Table 8.4.1 Exact and asymptotic values for $Q(x)$

		$c_X^2 = 0$		$c_X^2 = 0.5$		$c_X^2 = 1.5$	
	x	$Q(x)$	$Q_{asy}(x)$	$Q(x)$	$Q_{asy}(x)$	$Q(x)$	$Q_{asy}(x)$
$\rho = 0.2$	0.5	0.11586	0.07755	0.12462	0.14478	0.13667	0.09737
	1	0.02288	0.03007	0.07146	0.07712	0.09669	0.07630
	2	0.00196	0.00210	0.02144	0.02188	0.05234	0.04685
	3	0.00015	0.00015	0.00617	0.00621	0.03025	0.02877
	5	7.20E-7	7.20E-7	0.00050	0.00050	0.01095	0.01085
$\rho = 0.5$	0.5	0.35799	0.30673	0.37285	0.38608	0.39390	0.34055
	1	0.17564	0.18817	0.26617	0.26947	0.31629	0.28632
	2	0.05304	0.05356	0.13106	0.13126	0.21186	0.20239
	5	0.00124	0.00124	0.01517	0.01517	0.07179	0.07149
	10	2.31E-6	2.31E-6	0.00042	0.00042	0.01262	0.01262
$\rho = 0.8$	0.5	0.70164	0.67119	0.71197	0.71709	0.72705	0.70204
	1	0.55489	0.56312	0.62430	0.62549	0.66522	0.65040
	2	0.36548	0.36601	0.47582	0.47589	0.56345	0.55825
	5	0.10050	0.10050	0.20959	0.20959	0.35322	0.35299
	10	0.01166	0.01166	0.05343	0.05343	0.16444	0.16444

Heavy-tailed distributions

The probability distribution function $B(x)$ of the claim sizes (service times) is said to be *heavy-tailed* when $B(x)$ does not satisfy Assumption 8.4.1. An important subclass of heavy-tailed distributions is the class of subexponential distributions. Let $X_1, X_2, \ldots$ be a sequence of non-negative independent random variables which are distributed according to the probability distribution function $B(x)$. The distribution function $B(x)$ is said to be *subexponential* if $B(x) < 1$ for all $x > 0$ and

$$P\{X_1 + \cdots + X_n > x\} \sim nP\{X_1 > x\} \quad \text{as } x \to \infty \tag{8.4.10}$$

for all $n \geq 2$. It can be shown that (8.4.10) holds for all $n \geq 2$ if it holds for $n = 2$. A physical interpretation of subexponentiality follows by noting that condition (8.4.10) is equivalent to

$$P\{X_1 + \cdots + X_n > x\} \sim P\{\max(X_1, \ldots, X_n) > x\} \quad \text{as } x \to \infty \tag{8.4.11}$$

for all $n \geq 2$. In other words, subexponentiality means that a very large value of a finite sum of independent subexponential random variables is most likely caused by a very large value of one of the random variables. This property makes subexponentiality a commonly used paradigm in insurance mathematics, especially in modelling catastrophes. The class of subexponential distributions is a natural subclass of heavy-tailed distributions. This subclass includes the lognormal distribution, the Pareto distribution and the Weibull distribution with a shape parameter less than 1. The equivalence of (8.4.10) and (8.4.11) is easily proved. Therefore note that

$$\begin{aligned} P\{\max(X_1, \ldots, X_n) > x\} &= 1 - [B(x)]^n \\ &= [1 - B(x)] \sum_{k=0}^{n-1} [B(x)]^k \sim n[1 - B(x)] \\ &\quad \text{as } x \to \infty \end{aligned}$$

and so $P\{\max(X_1, \ldots, X_n) > x\} \sim nP\{X_1 > x\}$ as $x \to \infty$. From this result the equivalence of (8.4.10) and (8.4.11) follows.

Denote by

$$B_e(x) = \frac{1}{\mu} \int_0^x \{1 - B(y)\}\, dy, \quad x \geq 0$$

the equilibrium excess distribution function associated with $B(x)$. Then the following result can be proved:

$$Q(x) \sim \frac{\rho}{1 - \rho}[1 - B_e(x)] \quad \text{as } x \to \infty \tag{8.4.12}$$

if and only if $B(x)$ is subexponential. Here $\rho = \lambda\mu/\sigma$. This result is mainly of theoretical importance. Unlike the asymptotic expansion (8.4.9) for the light-tailed

case, the asymptotic expansion (8.4.12) for the heavy-tailed case is typically bad for x-values of interest. It takes very large x before the asymptotic expansion (8.4.12) applies. In practice one has to use numerical Laplace inversion to calculate the tail probabilities $Q(x)$ in the heavy-tailed case; see Appendix F.

We give no rigorous proof for the result (8.4.12), but we do make it plausible. To do so, we first establish the relation

$$Q(x) = \sum_{n=0}^{\infty} (1-\rho)\rho^n [1 - B_{n,e}(x)], \quad x \geq 0, \tag{8.4.13}$$

where $B_{0,e}(x) = 1$ for all $x \geq 0$ and $B_{n,e}(x)$ is the n-fold convolution of $B_e(x)$ with itself for $n \geq 1$. The formula (8.4.13) does not require any condition on the distribution function $B(x)$. To prove (8.4.13), denote the Laplace transform of $Q(x)$ by $Q^*(s) = \int_0^\infty e^{-sx} Q(x)\, dx$. Taking Laplace transforms of both sides of (8.4.5), we find

$$Q^*(s) = \frac{\rho}{s} - \frac{\rho b_e^*(s)}{s} + \rho Q^*(s) b_e^*(s),$$

where $b_e^*(s)$ is the Laplace transform of the derivative $b_e(x) = (1/\mu)[1 - B(x)]$ of the equilibrium excess distribution function $B_e(x)$. This gives

$$\begin{aligned} Q^*(s) &= \frac{\rho - \rho b_e^*(s)}{s[1 - \rho b_e^*(s)]} = \frac{\rho - \rho b_e^*(s)}{s} \sum_{n=0}^{\infty} \rho^n [b_e^*(s)]^n \\ &= \frac{1}{s} - (1-\rho) \sum_{n=0}^{\infty} \rho^n \frac{[b_e^*(s)]^n}{s}. \end{aligned} \tag{8.4.14}$$

It is left to the reader to verify that $[b_e^*(s)]^n/s$ is the Laplace transform of $B_{n,e}(x)$; see also relation (E.12) in Appendix E. Inversion of (8.4.14) yields

$$Q(x) = 1 - \sum_{n=0}^{\infty} (1-\rho)\, \rho^n B_{n,e}(x) = \sum_{n=0}^{\infty} (1-\rho)\, \rho^n [1 - B_{n,e}(x)],$$

proving (8.4.13). Next the expansion (8.4.12) can be made plausible. Assume that $B(x)$ is subexponential. If in addition an integrability condition is imposed on $B(x)$ to exclude pathological cases, it can be shown that the equilibrium excess distribution function $B_e(x)$ is subexponential as well. Then $1 - B_{n,e}(x) \sim n[1 - B_e(x)]$ as $x \to \infty$ for all n and thus

$$Q(x) \sim \sum_{n=0}^{\infty} (1-\rho)\rho^n n [1 - B_e(x)] \quad \text{as } x \to \infty,$$

which yields (8.4.12) by noting that $\sum_{n=0}^{\infty} (1-\rho)\rho^n n = \rho/(1-\rho)$.

EXERCISES

8.1 Use Laplace transform theory to verify the following results:

(a) The renewal function associated with the interoccurrence-time density $f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$ is

$$M(x) = \frac{x}{E(X)} + \frac{1}{2}(c_X^2 - 1)[1 - e^{-(p\lambda_1 + (1-p)\lambda_2)x}], \qquad x \geq 0,$$

where the random variable X denotes the interoccurrence time.

(b) The renewal function associated with the interoccurrence-time density $f(x) = p\lambda e^{-\lambda x} + (1-p)\lambda^2 x e^{-\lambda x}$ is

$$M(x) = \frac{x}{E(X)} + \frac{1}{2}(c_X^2 - 1)[1 - e^{-\lambda(2-p)x}], \qquad x \geq 0.$$

8.2 For a renewal process let $M_2(t) = E[N^2(t)]$ be the second moment of the number of renewals up to time t. Verify that $M_2(t)$ satisfies the renewal equation

$$M_2(t) = 2M(t) - F(t) + \int_0^t M_2(t-x) f(x)\, dx, \qquad t \geq 0,$$

where $f(x)$ is the probability density of the interoccurrence times. Next verify that

$$\lim_{t\to\infty} E[N^2(t)] - \left\{ \frac{t^2}{\mu_1^2} + \left(\frac{2\mu_2}{\mu_1^3} - \frac{3}{\mu_1} \right) t \right\} = \frac{3\mu_2^2}{2\mu_1^4} - \frac{2\mu_3}{3\mu_1^3} - \frac{3\mu_2}{2\mu_1^2} + 1,$$

where μ_k denotes the kth moment of the density $f(x)$. Also, prove that

$$\lim_{t\to\infty} \int_0^t E[N^2(y)]\, dy - \left[\frac{t^3}{3\mu_1^3} + \left(\frac{\mu_2}{\mu_1^3} - \frac{3}{2\mu_1} \right) t^2 + \left(\frac{3\mu_2^2}{2\mu_1^4} - \frac{2\mu_3}{3\mu_1^3} - \frac{3\mu_2}{2\mu_1^2} + 1 \right) t \right]$$
$$= \frac{\mu_4}{6\mu_1^3} - \frac{\mu_2\mu_3}{\mu_1^4} + \frac{\mu_2^3}{\mu_1^5} + \frac{\mu_3}{2\mu_1^2} - \frac{3\mu_2^2}{4\mu_1^3}.$$

8.3 Consider a renewal process generated by the interoccurrence times $X_1, X_2, \ldots$ with mean μ_1 and second moment μ_2. Let L_1 be the length of the interoccurrence time covering epoch t. Derive a renewal equation for $E(L_t)$. Verify the following results:

(a) $E(L_t) = 2\mu_1 - \mu_1 e^{-t/\mu_1}$ for all t when the X_i are exponentially distributed.

(b) $\lim_{t\to\infty} E(L_t) = \mu_2/\mu_1$ when the X_i are continuously distributed.

Also derive a renewal equation for $P\{L_t > x\}$. Prove that the limiting distribution of L_t has the density $xf(x)/\mu_1$ when the X_i have a probability density $f(x)$. Can you give a heuristic explanation of why $E(L_t) \geq \mu_1$?

8.4 Consider an alternating renewal process in which the on-times and the off-times are generally distributed. The on-times are assumed to have a probability density. Let $P_{on}(t)$ be the probability that the process is in the on-state at time t given that an on-time starts at epoch 0.

(a) Prove that

$$P_{on}(t) = 1 - F_{on}(t) + \int_0^t [1 - F_{on}(t-x)] m(x)\, dx, \qquad t \geq 0,$$

where $F_{on}(t)$ is the probability distribution function of the on-time and $m(x)$ is the renewal density for the renewal process in which the interoccurrence time is distributed as the sum of an on-time and an off-time. Express the Laplace transform of $P_{on}(t)$ in terms of the Laplace transforms of the on-time density and the off-time density. Give an expression for the Laplace transform of $E(U(t)) = \int_0^t P_{on}(u)\,du$, where the random variable $U(t)$ denotes the cumulative on-time during $[0, t]$.

(b) Use the result of (a) to verify that

$$P_{on}(t) = \sum_{k=0}^{[t/D]} \frac{(t - kD)^k}{\mu^k k!} e^{-(t-kD)/\mu}, \quad t \geq 0,$$

when the off-time is a constant D and the on-time has an exponential distribution with mean $1/\mu$.

8.5 Consider the alternating renewal process in which both the on-times and the off-times have a general probability distribution. Assuming that an on-time starts at epoch 0, denote by the random variable $U(t)$ the cumulative amount of time the system is in the on-state during $[0, t]$.

(a) Use Theorem 2.2.5 to verify that $U(t)$ is asymptotically normally distributed with mean $\mu_{on}t/(\mu_{on} + \mu_{off})$ and variance $(\mu_{on}^2\sigma_{off}^2 + \mu_{off}^2\sigma_{on}^2)t/(\mu_{on} + \mu_{off})^3$, where $\mu_{on}(\mu_{off})$ and $\sigma_{on}^2(\sigma_{off}^2)$ denote the mean and the variance of the on-time (off-time).

(b) Derive a renewal equation for $E(U(t))$. Assuming that the on-time distribution and the off-time distribution are not both arithmetic, prove that

$$\lim_{t\to\infty} \left[E(U(t)) - \frac{\mu_{on}}{\mu_{on} + \mu_{off}} t \right] = \frac{\mu_{on}\sigma_{off}^2 - \mu_{off}\sigma_{on}^2}{2(\mu_{on} + \mu_{off})^2} + \frac{\mu_{on}\mu_{off}}{2(\mu_{on} + \mu_{off})}.$$

8.6 Consider the alternating renewal process in which both the on-times and the off-times have a general probability distribution. Let μ_{on} and μ_{off} denote the respective means of an on-time and an off-time. Denote by $G_{on}(x, t)$ the joint probability that the system is on at time t and that the residual on-time at time t is no more than x. Derive a renewal equation for $G_{on}(x, t)$. Assuming that the distribution functions of the on-time and off-time are not both arithmetic, prove that

$$\lim_{t\to\infty} G_{on}(x, t) = \frac{\mu_{on}}{\mu_{on} + \mu_{off}} \times \frac{1}{\mu_{on}} \int_0^x [1 - F_{on}(y)]\,dy, \quad x \geq 0,$$

where $F_{on}(x)$ denotes the probability distribution function of the on-time.

8.7 Consider the alternating renewal process. Let $F_{on}(t)$ and $F_{off}(t)$ denote the probability distribution functions of the on-time and the off-time. Assume that these distribution functions have respective densities $f_{on}(t)$ and $f_{off}(t)$. For any fixed $t > 0$, define $H_{on}(t, x)$ $(H_{off}(t, x))$ as the probability that the cumulative on-time during $[0, t]$ is no more than x given that an on-time (off-time) starts at epoch 0.

(a) Argue the integral equations

$$H_{on}(t, x) = \int_0^x H_{off}(t - u, x - u) f_{on}(u)\,du, \quad 0 \leq x < t$$

$$H_{off}(t, x) = 1 - F_{off}(t - x) + \int_0^{t-x} H_{on}(t - u, x) f_{off}(u)\,du, \quad 0 \leq x < t.$$

(b) By repeated substitution, verify that

$$H_{on}(t,x) = \sum_{n=0}^{\infty} \{F_{off}^{n*}(t-x) - F_{off}^{(n+1)*}(t-x)\} F_{on}^{(n+1)*}(x), \quad 0 \le x < t,$$

$$H_{off}(t,x) = \sum_{n=0}^{\infty} \{F_{off}^{n*}(t-x) - F_{off}^{(n+1)*}(t-x)\} F_{on}^{n*}(x), \quad 0 \le x < t,$$

where $F^{n*}(x)$ denotes the n-fold convolution of a probability distribution of $F(x)$ with itself for $n \ge 1$ and $F^{0*}(x) = 1$ for all $x \ge 0$.

8.8 Consider Exercise 8.7 again. Define for any fixed $t_0 > 0$,

$$\Omega(t_0, x) = \lim_{t \to \infty} P\{\text{the cumulative on-time during the time interval } [t, t+t_0] \text{ is no more than } x\}$$

for $0 \le x < t_0$. Use results from Exercise 8.7 to argue that $\Omega(t_0, x)$ is given by

$$\frac{\mu_{on}}{\mu_{on}+\mu_{off}} \sum_{n=0}^{\infty} \{F_{off}^{n*}(t_0-x) - F_{off}^{(n+1)*}(t_0-x)\} F_{on}^{e} * F_{on}^{n*}(x)$$

$$+ \frac{\mu_{off}}{\mu_{on}+\mu_{off}} \sum_{n=0}^{\infty} \{F_{off}^{e} * F_{off}^{n*}(t_0-x) - F_{off}^{e} * F_{off}^{(n+1)*}(t_0-x)\} F_{on}^{(n+1)*}(x)$$

$$+ \frac{\mu_{off}}{\mu_{on}+\mu_{off}} \{1 - F_{\text{off}}^{e}(t_0-x)\}, \quad 0 \le x < t_0,$$

where $F^e(x)$ denotes the equilibrium excess distribution function of a probability distribution function $F(x)$ and $A * B(x)$ denotes the convolution of two distribution functions $A(x)$ and $B(x)$.

8.9 Consider an age-replacement model in which preventive replacements are only possible at special times. Opportunities for preventive replacements occur according to a Poisson process with rate λ. The item is replaced by a new one upon failure or upon a preventive replacement opportunity occurring when the age of the item is T or more, whichever occurs first. The lifetime of the item has a probability density $f(x)$. The cost of replacing the item upon failure is c_0 and the cost of a preventive replacement is c_1 with $0 < c_1 < c_0$. Determine the long-run average cost per time unit. This problem is motivated by Dekker and Smeitink (1994).

8.10 A production machine gradually deteriorates in time. The machine has N possible working conditions $1, \ldots, N$ which describe increasing degrees of deterioration. Here working condition 1 represents a new system and working condition N represents a failed system. If the system reaches the working condition i, it stays in this condition during an exponentially distributed time with mean $1/\mu$ for each i with $1 \le i < N$. A change of the working condition cannot be observed except for a failure which is detected immediately. The machine is replaced by a new one upon failure or upon having worked during a time T, whichever occurs first. Each planned replacement involves a fixed cost of $J_1 > 0$, whereas a replacement because of a failure involves a fixed cost of $J_2 > 0$. The replacement time is negligible in both cases. Also, the system incurs an operating cost of $a_i > 0$ for each time unit the system is operating in working condition i. Use Lemma 1.1.4 to verify that

the long-run average cost per time unit is given by

$$\left\{ T \sum_{k=0}^{N-2} p_k + \frac{N-1}{\mu} \left(1 - \sum_{k=0}^{N-1} p_k \right) \right\}^{-1} \times \left\{ J_2 + (J_2 - J_1) \sum_{k=0}^{N-2} p_k + \sum_{k=0}^{N-2} p_k \sum_{i=1}^{k+1} a_i \frac{T}{k+1} + \sum_{i=1}^{N-1} a_i \sum_{k=N-1}^{\infty} p_k \frac{T}{k} \right\},$$

where $p_k = e^{-\mu T}(\mu T)^k / k!$. This problem is motivated by Luss (1976).

8.11 Consider a two-unit reliability model with one operating unit and one unit in warm standby. The operating unit has a constant failure rate of λ_0, while the unit in warm standby has a constant failure rate of λ_1. Upon failure of the operating unit, the unit in warm standby is put into operation if available. The repair time of a failed unit has a general probability distribution function $G(x)$ with density $g(x)$ and mean μ_R. The system is down when both units have failed. For the case of a single repair facility, prove that the long-run fraction of time the system is down is as follows. This problem is based on Gaver (1963).

$$\frac{\mu_R - \int_0^\infty \{1 - G(x)\} e^{-\lambda_0 x}\, dx}{\mu_R + (\lambda_0 + \lambda_1)^{-1} \int_0^\infty e^{-\lambda_0 x} g(x)\, dx}.$$

8.12 Consider an unreliable production unit whose output is temporarily stored in a finite buffer with capacity K. The buffer serves for the demand process as protection against random interruptions in the production process. For the output there is a constant demand at rate ν. When operating, the production unit produces at a constant rate $P > \nu$ if the buffer is not full and produces at the demand rate ν otherwise. If demand occurs while the unit is down and the buffer is empty then it is lost. The operating time of the unit is exponentially distributed with mean $1/\lambda$. If a failure occurs, the unit enters repair for an exponentially distributed time with mean $1/\mu$. Determine the long-run fraction of demand lost and determine the average inventory level in the buffer. (*Hint*: define the state of the system as $(1, x)$ and $(0, x)$ respectively when the inventory in the buffer is x and the unit is operating or down. The process regenerates itself each time the system enters state (0,0). Use differential equations to get the desired performance measures). This problem is motivated by Wijngaard (1979).

BIBLIOGRAPHIC NOTES

The key renewal theorem has a long history, and analytic proofs were given under rather restrictive conditions. The reader is referred to the book of Feller (1971) for a transparent proof under the weak condition of direct Riemann integrability; see also Asmussen (1987). The results for the alternating renewal process in Section 8.3 are proved in greater generality in Takács (1957). The material from Example 8.3.1 is based on the paper of Van der Heijden (1987). The renewal-theoretic method used in Section 8.4 to derive asymptotic estimates for ruin and waiting-time probabilities comes from Feller (1971). Another application of this powerful method to a storage problem for dams is given in De Kok *et al.* (1984). A good reference on heavy-tailed distributions is Embrechts *et al.* (1997).

REFERENCES

Asmussen, S. (1987) *Applied Probability and Queues*. John Wiley & Sons, Inc., New York.

Dekker, R., and Smeitink, E. (1994) Preventive maintenance at opportunities of restricted duration. *Naval Res. Logist*., **41**, 335–353.

De Kok, A.G., Tijms, H.C. and Van der Duyn Schouten, F.A. (1984) Approximations for the single-product production-inventory problem with compound Poisson demand and service level constraints. *Adv. Appl. Prob*., **16**, 378–401.

Den Iseger, P.W., Smith, M.A.J. and Dekker, R. (1997) Computing compound Poisson distributions faster. *Insurance Mathematics and Economics*, **20**, 23–34.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events*. Springer-Verlag, Berlin.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd edn. John Wiley & Sons, Inc., New York.

Gaver, D.P. (1963) Time to failure and availability of paralleled systems with repair. *IEEE Trans. Reliab*., **12**, 30–38.

Lorden, G. (1970) On excess over the boundary. *Ann. Math. Statist*., **41**, 520–527.

Luss, H. (1976) Maintenance policies when deterioration can be observed by inspections. *Operat. Res*., **24**, 359–366.

Takács, L.J. (1957) On certain sojourn time problems in the theory of stochastic processes. *Acta Mathematica Academiae Scientiarum Hungaricae*, **8**, 169–191.

Van der Heijden, M.C. (1987) Interval availability distribution for a 1-out-of-2 reliability system. *Prob. Engng. Inform. Sci.*, **2**, 211–224.

Wijngaard, J. (1979) The effect of interstage buffer storage on the output of two unreliable production units in series with different production rates. *AIEE Trans.*, **11**, 42–47.

Xie, M. (1989) On the solution of renewal-type integral equations. *Commun. Statist.*, **18**, 281–293.

CHAPTER 9

Algorithmic Analysis of Queueing Models

9.0 INTRODUCTION

Queueing models have their origin in the study of design problems of automatic telephone exchanges and were first analysed by the queueing pioneer A.K. Erlang in the early 1900s. In planning telephone systems to meet given performance criteria, questions were asked such as: How many lines are required in order to give a certain grade of service? What is the probability that a delayed customer has to wait more than a certain time before getting a connection? Similar questions arise in the design of many other systems: How many terminals are needed in a computer system so that 80% of the users get access to a terminal within 20 seconds? What will be the effect on the average waiting time of customers when changing the size of a maintenance staff to service leased equipment? How much storage space is needed in buffers at workstations in an assembly line in order to keep the probability of blocking below a specified acceptable level?

These design problems and many others concern, in fact, facilities serving a community of users, where both the times at which the users ask for service and the durations that the requests for service will occupy facilities are stochastic, so that inevitably congestion occurs and queues may build up. In the first stage of design the system engineer usually needs quick answers to a variety of questions like those posed above. Queueing theory constitutes a basic tool for making first-approximation estimates of queue sizes and probabilities of delays. Such a simple tool should in general be preferred to simulation, especially when it is possible to have a large number of different configurations in the design problem.

In this chapter we discuss a number of basic queueing models that have proved to be useful in analysing a wide variety of stochastic service systems. The emphasis will be on algorithms and approximations rather than on mathematical aspects. We feel that there is a need for such a treatment in view of the increased use of queueing models in modern technology. Actually, the application of queueing theory in the performance analysis of computer and communication systems has

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

stimulated much practically oriented research on computational aspects of queueing models. It is to these aspects that the present chapter is addressed. Here considerable attention is paid to robustness results. While it was seen in Section 5.2 that many loss systems (no access of arrivals finding all servers busy) are exactly or nearly insensitive to the distributional form of the service time except for its first moment, it will be demonstrated in this chapter that many delay systems (full access of arrivals) and many delay-loss systems (limited access of arrivals) allow for two-moment approximations. The approximate methods for complex queueing models are usually based on exact results for simpler related models and on asymptotic expansions. The usefulness of asymptotic expansions can hardly be overestimated.

Algorithmic analysis of queueing systems is more than getting numerical answers. The essence of algorithmic probability is to find probabilistic ideas which make the computations transparent and natural. However, once an algorithm has been developed according to these guidelines, one should always verify that it works in practice. The algorithms presented in this chapter have all been thoroughly tested. The cornerstones of the algorithms are:

- the embedded Markov chain method,
- the continuous-time Markov chain approach,
- renewal-theoretic methods,
- asymptotic expansions,
- discrete FFT method and numerical Laplace inversion.

This chapter is organized as follows. Section 9.1 reviews some basic concepts including phase-type distributions and Little's formula. In Section 9.2 we derive algorithms for computing the state probabilities and the waiting-time probabilities in the single-server queue with Poisson input and general service times ($M/G/1$ queue). These results are extended in Section 9.3 to the single-server queue with batch Poisson input. In Section 9.4 we consider the finite-buffer $M/G/1$ queue and the $M/G/1$ queue with impatient customers. The solution of these queueing systems can be expressed in terms of the solution for the infinite-capacity $M/G/1$ queue. The single-server queue with general interarrival times and service times is the subject of Section 9.5. Section 9.6 deals with multi-server queues with Poisson input, including both the case of single arrivals and the case of batch arrivals. Tractable exact results are only obtained for the special case of deterministic services and exponential services. For the case of general service times we derive several approximations. These approximations include two-moment approximations that are based on exact results for simpler models and use a linear interpolation with respect to the squared coefficient of variation of the service time. In Section 9.7 the multi-server queue with renewal input is discussed. In particular, attention is paid to the tractable models with exponential services and deterministic services. In Section 9.8 we consider finite-capacity queueing systems with

limited access of arrivals. In particular, attention is paid to approximations for the rejection probability. Throughout this chapter numerical results are given in order to provide insight into the performance of the solution methods. Indispensable tools for the solution of queues are the discrete Fast Fourier Transform (FFT) method and numerical Laplace inversion. This is a remarkable twist in the history of queueing analysis. The irony is that complaints about the 'Laplacian curtain' stimulated to a large extent the development of algorithmic analysis for queues. Most of the results for queues in the post-war period were in terms of generating functions or Laplace transforms. For a long time it was believed that such results were not very useful for computational purposes. However, the situation dramatically changed with the invention of the discrete FFT method in 1965, one of the greatest breakthroughs in numerical analysis. The power of this method was directly realized in the field of engineering, but it took some time before the immense usefulness of the discrete FFT method was recognized in the field of applied probability as well.

9.1 BASIC CONCEPTS

In this section we discuss a number of basic concepts for queueing systems. The discussion is restricted to queueing systems with only one service node. However, the fundamental results below are also useful for networks of queues.

Let us start by giving Kendall's notation for a number of standard queueing models in which the source of population of potential customers is assumed to be infinite. The customers arrive singly and are served singly. In front of the servers there is a common waiting line. A queueing system having waiting room for an unlimited number of customers can be described by a three-part code $a/b/c$. The first symbol a specifies the interarrival-time distribution, the second symbol b specifies the service-time distribution and the third symbol c specifies the number of servers. Some examples of Kendall's shorthand notation are:

1. $M/G/1$: Poisson (Markovian) input, general service-time distribution, 1 server.
2. $M/D/c$: Poisson input, deterministic service times, c servers.
3. $GI/M/c$: general, independently distributed interarrival times, exponential (Markovian) service times, c servers.
4. $GI/G/c$: general, independently distributed interarrival times, general service-time distribution, c servers.

The above notation can be extended to cover other queueing systems. For example, queueing systems have waiting room only for K customers (excluding those in service) are often abbreviated by a four-part code $a/b/c+K$. The notation $GI^X/G/c$ is used for infinite-capacity queueing systems in which customers arrive in batches and the batch size is distributed according to the random variable X.

Phase-type distributions

In queueing applications it is often convenient to approximate the interarrival time and/or the service time by distributions that are built out of a finite sum or a finite mixture of *exponentially* distributed components, or a combination of both. These distributions are called *phase-type distributions*. For practical purposes it usually suffices to use finite mixtures of Erlangian distributions with the *same* scale parameters or Coxian-2 distributions. These distributions are discussed in detail in Appendix B. The class of Coxian-2 distributions contains the hyperexponential distribution of order 2 as special case. The hyperexponential distribution always has a coefficient of variation greater than or equal to 1. This distribution is particulary suited to model irregular interarrival (or service) times which have the feature that most outcomes tend to be small and large outcomes occur only occasionally. The class of mixtures of Erlangian distributions with the same scale parameters is much more versatile than the class of Coxian-2 distributions and allows us to cover any positive value of the coefficient of variation. In particular, a mixture of E_{k-1} and E_k distributions with the same scale parameters is convenient to represent regular interarrival (or service) times which have a coefficient of variation smaller than or equal to 1. The theoretical basis for the use of mixtures of Erlangian distributions with the same scale parameters is provided by Theorem 5.5.1. This theorem states that each non-negative random variable can be approximated arbitrarily closely by a random sum of exponentially distributed phases with the same means. This explains why finite mixtures of Erlangian distributions with the same scale parameters are widely used for queueing calculations.

Performance measures

It is convenient to use the $GI/G/c/c+N$ queue as a vehicle to introduce some basic notation. Thus, we assume a multi-server queue with c identical servers and a waiting room of capacity N ($\leq \infty$) for customers awaiting to be served. A customer who finds $c+N$ other customers present upon arrival is rejected and has no further influence on the system. Otherwise, the arriving customer is admitted to the system and waits in queue until a server becomes available. The customers arrive according to a renewal process. In other words, the interarrival times are positive, independent random variables having a common probability distribution function $A(t)$. The service times of the customers are independent random variables with a common probability distribution function $B(x)$ and are also independent of the arrival process. The queue discipline specifying which customer is to be served next is first-come first-served (FCFS) unless stated otherwise. A server cannot be idle when customers are waiting in queue and a busy server works at unity rate. A customer leaves the system upon service completion. Let

$$\lambda = \text{the long-run average arrival rate of customers,}$$

$$E(S) = \text{the mean service time of a customer.}$$

The random variable S denotes the service time of a customer. Note that $\lambda = 1/E(A)$, where the random variable A denotes the interarrival time. An important quantity is the *offered load*, which is defined as $\lambda E(S)$. This dimensionless quantity indicates the average amount of work that is offered to the system per time unit. In the $GI/G/c$ queue ($N = \infty$) the offered load should be less than the maximum load the system can handle, otherwise infinitely long queues ultimately build up. Letting

$$\rho = \frac{\lambda E(S)}{c},$$

the following assumption is made.

Assumption 9.1.1 *For the $GI/G/c$ queue the load factor ρ is below* 1.

It will be seen below that in the $GI/G/c$ queue the quantity ρ can be interpreted as the long-run fraction of time that a given server is busy. This explains why ρ is called the *server utilization* in the $GI/G/c$ queue. In addition to Assumption 9.1.1 we make the following technical assumption.

Assumption 9.1.2 *(a) The interarrival-time distribution $A(t)$ or the service-time distribution $B(t)$ has a positive density on some interval.*

(b) The probability that the interarrival time A is larger than the service time S is positive.

Define a cycle as the time elapsed between two consecutive arrivals that find the system empty. Then, under Assumptions 9.1.1 and 9.1.2, it can be shown that the expected value of the cycle length is always finite. The proof of this result is quite deep and is not given here; see Wolff (1989). Let us now define the following random variables:

$L(t)$ = the number of customers in the system at time t (including those in service),
$L_q(t)$ = the number of customers in the queue at time t (excluding those in service),
D_n = the amount of time spent by the nth accepted customer in the queue (excluding service time),
U_n = the amount of time spent by the nth accepted customer in the system (including service time).

The continuous-time stochastic process $\{L(t)\}$ and $\{L_q(t)\}$ and the discrete-time stochastic processes $\{D_n\}$ and $\{U_n\}$ are all regenerative. The regeneration epochs are the epochs at which an arriving customer finds the system empty. The regeneration cycles have finite means. Thus the following long-run averages exist:

$$L = \lim_{t\to\infty} \frac{1}{t} \int_0^t L(u)\, du \quad \text{(the long-run average number in system)}$$

$$L_q = \lim_{t\to\infty} \frac{1}{t} \int_0^t L_q(u)\, du \quad \text{(the long-run average number in queue)}$$

$$W_q = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} D_k \quad \text{(the long-run average delay in queue)}$$

$$W = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} U_k \quad \text{(the long-run average wait in system).}$$

These long-run averages are constants with probability 1. The steady-state probabilities p_j and the steady-state waiting-time distribution function $W_q(x)$ are defined by

$$p_j = \lim_{t\to\infty} P\{L(t) = j\}, \quad j = 0, 1, \ldots$$

and

$$W_q(x) = \lim_{n\to\infty} P\{D_n \le x\}, \quad x \ge 0.$$

These limits exist and represent proper probability distributions; see Theorem 2.2.4. As pointed out in Section 2.2, it is often preferable to interpret p_j and $W_q(x)$ as the long-run fraction of time that j customers are in the system and as the long-run fraction of accepted customers whose delay in queue is at most x. In batch-arrival queues the above limits need not exist, while p_j and $W_q(x)$ can still be defined as long-run averages. The long-run averages L_q and W_q can be expressed in terms of the state probabilities p_j and the waiting-time probabilities $W_q(x)$:

$$L_q = \sum_{j=c}^{c+N} (j-c)p_j \quad \text{and} \quad W_q = \int_0^\infty \{1 - W_q(x)\}\, dx.$$

It is important to note that the distribution of the number of customers in the system is invariant to the order of service, provided that the queue discipline is *service-time independent* and *work-conserving*. Here 'service-time independent' means that the rule for selecting the next customer to be served does not depend on the service time of a customer, while 'work-conserving' means that the work or service requirement of a customer is not affected by the queue discipline. Queue disciplines having these properties include first-come first-served, last-come first-served and service in random order. The waiting-time distribution will obviously depend on the order of service.

Let the random variable $I_n = 1$ if the nth arrival is rejected and let $I_n = 0$ otherwise. Then the long-run fraction of customers who are rejected is given by the constant

$$P_{rej} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k.$$

Little's formula

The most basic result for queueing systems is Little's formula. This formula relates certain averages like the average number of customers in queue and the average delay in queue per customer. Little's formula is valid for almost any queueing system. In particular, for the $GI/G/c/c+N$ queue, we have the fundamental relations

$$L_q = \lambda(1 - P_{rej})W_q, \quad L = \lambda(1 - P_{rej})W, \tag{9.1.1}$$

$$\text{the long-run average number of busy servers} = (1 - P_{rej})E(S). \tag{9.1.2}$$

Note that $P_{rej} = 0$ if $N = \infty$. A heuristic but insightful motivation of these formulas was given in Section 2.3. The result (9.1.2) has two interesting implications. First, since each of the c servers carries on average the same load,

$$\text{the long-run fraction of time a given server is busy} = \frac{1}{c}\lambda(1 - P_{rej})E(S).$$

In particular, the long-run fraction of time a given server is busy equals ρ in the $GI/G/c$ queue. Second, since p_j represents the long-run fraction of time that j customers are present, the long-run average number of busy servers is also given by the expression $\sum_{j=0}^{c-1} jp_j + c\sum_{j\geq c} p_j$. Thus we obtain the useful identity

$$\sum_{j=1}^{c-1} jp_j + c\left(1 - \sum_{j=0}^{c-1} p_j\right) = \lambda(1 - P_{rej})E(S). \tag{9.1.3}$$

In particular, we find the relation $p_0 = 1 - \lambda E(S)$ for the $GI/G/1$ queue. The above relations can be directly extended to queueing systems with batch arrivals.

9.2 THE $M/G/1$ QUEUE

In the $M/G/1$ queue, customers arrive according to a Poisson process with rate λ and the service times of the customers are independent random variables with a common general probability distribution function $B(x)$ with $B(0) = 0$. There is a single server and an infinite waiting room. Denoting by the random variable S the service time of a customer, it is assumed that the server utilization $\rho = \lambda E(S)$ is smaller than 1.

In Section 9.2.1 we derive a recursive algorithm for the computation of the state probabilities. Several derivations are possible for the recursion relation. Our derivation uses the so-called *regenerative approach*, which involves simple renewal-theoretic arguments. The regenerative approach directly leads to a numerically stable recursion scheme for the state probabilities and also allows in a natural way for generalizations to more complex queueing models. Using the technique of generating functions, we also derive an asymptotic expansion for the state probabilities. Since an explicit expression is available for the generating function of

the state probabilities, the discrete FFT method provides an alternative method to compute the state probabilities. In Section 9.2.2 we discuss the computation of the waiting-time probabilities when service is in order of arrival. Also attention is paid to an approximation for the waiting-time distribution. This approximation is based on the asymptotic expansion of the tail of the waiting-time distribution. Further, we discuss a simple but generally useful two-moment approximation for the waiting-time percentiles. Section 9.2.3 discusses the probability distribution of the length of a busy period and the computation of the waiting-time probabilities when the last-come first-served discipline is used. The distribution of work in system is the subject of Section 9.2.4.

9.2.1 The State Probabilities

The time-average probability p_j can be interpreted as the long-run fraction of time that j customers are in the system. Using a basic result from the theory of regenerative processes and a simple up- and downcrossing argument, we derive a numerically stable recursion scheme for the state probabilities p_j.

Theorem 9.2.1 *The state probabilities p_j satisfy the recursion*

$$p_j = \lambda a_{j-1} p_0 + \lambda \sum_{k=1}^{j} a_{j-k} p_k, \quad j = 1, 2, \ldots, \tag{9.2.1}$$

where the constants a_n are given by

$$a_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} \{1 - B(t)\}\, dt, \quad n = 0, 1, \ldots .$$

Proof The stochastic process $\{L(t), t \geq 0\}$ describing the number of customers in the system is regenerative. The process regenerates itself each time an arriving customer finds the system empty. Denoting by a cycle the time elapsed between two consecutive arrivals who find the system empty, we define the random variables

T = the length of one cycle,

T_j = the amount of time that j customers are present during one cycle

for $j = 0, 1, \ldots$. The expected length of one cycle is finite (this is in fact a by-product of the analysis in Section 2.6). By Theorem 2.2.3,

$$p_j = \frac{E(T_j)}{E(T)}, \quad j = 0, 1, \ldots . \tag{9.2.2}$$

By the lack of memory of the Poisson process, $E(T_0) = 1/\lambda$ and so

$$p_0 = \frac{1}{\lambda E(T)}. \tag{9.2.3}$$

The following simple idea is crucial for the derivation of a recurrence relation for the probabilities p_j. Divide a cycle into a random number of disjoint intervals separated by the service completion epochs and calculate $E(T_j)$ as the sum of the contributions from the disjoint intervals to the expected sojourn time in state j during one cycle. Thus, for $k = 0, 1, \ldots$, we define the random variable N_k by

$$N_k = \text{the number of service completion epochs in one cycle at which } k \text{ customers are left behind.}$$

Using the lack of memory of the Poisson arrival process, define

$$A_{kj} = \text{the expected amount of time that } j \text{ customers are present during a given service time that starts with } k \text{ customers present.}$$

Then, noting that the first service in a cycle starts with one customer present,

$$E(T_j) = A_{1j} + \sum_{k=1}^{j} E(N_k)A_{kj}, \quad j = 1, 2, \ldots . \tag{9.2.4}$$

It should be pointed out that Wald's equation is used to justify that $E(N_k)A_{kj}$ is the contribution to $E(T_j)$ of those service intervals starting with k customers present. To find another relation between $E(T_j)$ and $E(N_k)$, observe that for each $k = 0, 1, \ldots$, the number of downcrossings from state $k+1$ to state k in one cycle equals the number of upcrossings from state k to state $k+1$ in one cycle. The expected number of downcrossings of the $\{L(t)\}$ process from state $k+1$ to state k in one cycle equals $E(N_k)$ by definition. On the other hand, since the arrival process is a Poisson process, we have by Corollary 2.4.2 that the expected number of upcrossings from state k to state $k+1$ in one cycle equals $\lambda E(T_k)$. Thus

$$E(N_k) = \lambda E(T_k), \quad k = 0, 1, \ldots . \tag{9.2.5}$$

Together the relations (9.2.2) to (9.2.5) imply that

$$p_j = \lambda p_0 A_{1j} + \sum_{k=1}^{j} \lambda p_k A_{kj}, \quad j = 1, 2, \ldots . \tag{9.2.6}$$

To specify the constants A_{kj}, suppose that at epoch 0 a service starts when k customers are present. Define the random variable $I_j(t) = 1$ if at time t the service is still in progress and j customers are present and let $I_j(t) = 0$ otherwise. Then, for $j \geq k$,

$$\begin{aligned} A_{kj} &= E\left[\int_0^\infty I_j(t)\,dt\right] = \int_0^\infty E[I_j(t)]\,dt \\ &= \int_0^\infty P\{I_j(t) = 1\}\,dt = \int_0^\infty \{1 - B(t)\}e^{-\lambda t}\frac{(\lambda t)^{j-k}}{(j-k)!}\,dt. \end{aligned} \tag{9.2.7}$$

Together (9.2.6) and (9.2.7) yield the desired result.

The recursion (9.2.1) enables us to compute recursively $p_1, p_2, \ldots$ starting with $p_0 = 1 - \rho$. In Section 2.5 we proved that $p_0 = 1 - \rho$; see also relation (9.1.3). The recursion scheme is numerically stable, since the calculations involve only additions with positive numbers and thus cannot cause a loss of significant digits. For many service-time distributions of practical interest, numerical integration can be avoided for the computation of the constants a_n. Explicit expressions for the a_n can be given for the cases of deterministic and phase-type services.

Define the generating function $P(z)$ by

$$P(z) = \sum_{j=0}^{\infty} p_j z^j, \quad |z| \leq 1.$$

Multiplying both sides of (9.2.1) by z^j and summing over j, it is a matter of simple algebra to derive that

$$P(z) - p_0 = \lambda p_0 z \sum_{n=0}^{\infty} a_n z^n + \lambda \{P(z) - p_0\} \sum_{n=0}^{\infty} a_n z^n.$$

Since $p_0 = 1 - \rho$, we obtain

$$P(z) = (1 - \rho) \frac{1 - \lambda(1 - z)\alpha(z)}{1 - \lambda\alpha(z)}, \tag{9.2.8}$$

where $\alpha(z) = \sum_{n=0}^{\infty} a_n z^n$ is given by

$$\alpha(z) = \int_0^{\infty} \{1 - B(t)\} e^{-\lambda(1-z)t}\, dt.$$

Expression (9.2.8) for $P(z)$ coincides with expression (2.5.8), since $\int_0^{\infty} e^{-\lambda(1-z)t} b(t)\, dt = 1 - \lambda(1 - z)\alpha(z)$ when $b(t)$ is the probability density of the service time. The discrete FFT method provides an alternative method for the computation of the state probabilities using the explicit expression (9.2.8) for the generating function $P(z)$. A by-product of (9.2.8) is the famous Pollaczek–Khintchine formula

$$L_q = \frac{\lambda^2 E(S^2)}{2(1 - \rho)} \tag{9.2.9}$$

for the long-run average queue size. Using Little's formula $L_q = \lambda W_q$, it follows that the long-run average delay in queue per customer is given by

$$W_q = \frac{\lambda E(S^2)}{2(1 - \rho)}. \tag{9.2.10}$$

Asymptotic expansion for the state probabilities

The representation (9.2.8) shows that the generating function $P(z)$ is the ratio of two functions, $N(z)$ and $D(z)$. These functions allow for an analytic continuation outside the unit circle when the following assumption is made.

Assumption 9.2.1 *(a)* $\int_0^\infty e^{st}\{1 - B(t)\}\,dt < \infty$ *for some* $s > 0$.
(b) $\lim_{s \to B} \int_0^\infty e^{st}\{1 - B(t)\}\,dt = \infty$, *where* B *is the supremum over all* s *with*

$$\int_0^\infty e^{st}\{1 - B(t)\}\,dt < \infty.$$

The assumption requires that the service-time distribution is not heavy-tailed. This is the case in most situations of practical interest. Under Assumption 9.2.1, it can be obtained from Theorem C.1 in Appendix C that

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty, \tag{9.2.11}$$

where τ is the unique solution of the equation

$$\int_0^\infty e^{-\lambda(1-\tau)t}\{1 - B(t)\}\,dt = \frac{1}{\lambda} \tag{9.2.12}$$

on the interval $(1, 1 + B/\lambda)$ and the constant σ is given by

$$\sigma = \frac{(1-\rho)}{\lambda^2}\left[\int_0^\infty te^{-\lambda(1-\tau)t}\{1 - B(t)\}\,dt\right]^{-1}. \tag{9.2.13}$$

It is empirically found that the asymptotic expansion (9.2.11) already applies for relatively small values of j. The asymptotic expansion can be used to reduce the computational effort of the recursion scheme (9.2.1). Since $p_{j-1}/p_j \approx \tau$ for j large enough, the recursive calculations can be halted as soon as the ratio p_{j-1}/p_j has sufficiently converged to the constant τ.

9.2.2 The Waiting-Time Probabilities

In this subsection we discuss the computation of the waiting-time probabilities under the assumption that customers are served in order of arrival. Both exact methods and approximate methods are discussed.

Exact methods

The following exact methods can be used for the computation of $W_q(x)$:

(a) discretization,

(b) Laplace-inversion,

(c) phase method.

(a) By relation (8.4.5),

$$W_q(x) = W_q(0) + \lambda \int_0^x W_q(x-y)\{1 - B(y)\}\, dy, \quad x \geq 0 \tag{9.2.14}$$

with $W_q(0) = 1 - \rho$. This integral equation can be solved by using the discretization method discussed in Section 8.1.2. However, when a high accuracy is required, this method is computationally rather demanding even when it is combined with the asymptotic expansion for $W_q(x)$ to be given below.

(b) By (2.5.13), the Laplace transform of $1 - W_q(x)$ is given by

$$\int_0^\infty e^{-sx}\{1 - W_q(x)\}\, dx = \frac{\rho s - \lambda + \lambda b^*(s)}{s(s - \lambda + \lambda b^*(s))}, \tag{9.2.15}$$

where $b^*(s) = \int_0^\infty e^{-sx} b(x)\, dx$ is the Laplace transform of the service-time density $b(x)$. In Appendix F the computation of $W_q(x)$ by numerical Laplace inversion is discussed.

(c) In Section 5.5 it was shown that any service-time distribution function $B(x)$ can be arbitrarily closely approximated by a distribution function of the form

$$\sum_{j=1}^{\infty} q_j \left[1 - \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!} \right], \quad x \geq 0,$$

where $q_j \geq 0$ and $\sum_{j=1}^{\infty} q_j = 1$. This distribution function is a mixture of Erlangian distribution functions with the *same* scale parameters. It allows us to interprete the service time as a random sum of independent phases each having the *same* exponential distribution. Example 5.5.1 explains how to use continuous-time Markov chain analysis for the computation of $W_q(x)$ when the service-time distribution has the above form. This approach leads to a simple and fast algorithm.

A simple approximation to the waiting-time probabilities

Assume that Assumption 9.2.1 holds. Then, as was shown in Section 8.4,

$$1 - W_q(x) \sim \gamma e^{-\delta x} \quad \text{as } x \to \infty, \tag{9.2.16}$$

with

$$\delta = \lambda(\tau - 1) \quad \text{and} \quad \gamma = \frac{\sigma}{\tau - 1}, \tag{9.2.17}$$

where the constants τ and σ are given by (9.2.12) and (9.2.13).

We found empirically that the asymptotic expansion for $1 - W_q(x)$ is accurate enough for practical purposes for relatively small values of x. However, why

not improve this first-order estimate by adding a second exponential term? This suggests the following approximation to $1 - W_q(x)$:

$$1 - W_{app}(x) = \alpha e^{-\beta x} + \gamma e^{-\delta x}, \quad x \geq 0. \tag{9.2.18}$$

The constants α and β are found by matching the behaviour of $W_q(x)$ at $x = 0$ and the first moment of $W_q(x)$. Since $1 - W_q(0) = P_{delay}$ and $W_q = \int_0^\infty \{1 - W_q(x)\}\, dx$, it follows that

$$\alpha = P_{delay} - \gamma \quad \text{and} \quad \beta = \alpha(W_q - \gamma/\delta)^{-1}, \tag{9.2.19}$$

where $P_{delay} = \rho$ and an explicit expression for W_q is given by (9.2.10). It should be pointed out that the approximation (9.2.18) can be applied only if $\beta > \delta$, otherwise $1 - W_{app}(x)$ for x large would not agree with the asymptotic expansion (9.2.16). Numerical experiments indicate that $\beta > \delta$ holds for a wide class of service-time distributions of practical interest. Further support to (9.2.18) is provided by the fact that the approximation is exact for Coxian-2 services.

Numerical investigations show that the approximation (9.2.18) performs quite satisfactorily for all values of x. Table 9.2.1 gives the exact values of $1 - W_q(x)$, the approximate values (9.2.18) and the asymptotic values (9.2.16) for E_{10} and E_3 service-time distributions. The server utilization ρ is 0.2, 0.5, 0.8. In all examples the normalization $E(S) = 1$ is used.

A two-moment approximation for the waiting-time percentiles

In applications it often happens that only the first two moments of the service time are available. In these situations, two-moment approximations may be very helpful.

Table 9.2.1 The waiting-time probabilities

		Erlang-10			Erlang-3		
	x	exact	approx	asymp	exact	approx	asymp
$\rho = 0.2$	0.10	0.1838	0.1960	0.3090	0.1839	0.1859	0.2654
	0.25	0.1590	0.1682	0.2222	0.1594	0.1615	0.2106
	0.50	0.1162	0.1125	0.1282	0.1209	0.1212	0.1432
	0.75	0.0755	0.0694	0.0739	0.0882	0.0875	0.0974
	1.00	0.0443	0.0413	0.0427	0.0626	0.0618	0.0663
$\rho = 0.5$	0.10	0.4744	0.4862	0.5659	0.4744	0.4764	0.5332
	0.25	0.4334	0.4425	0.4801	0.4342	0.4361	0.4700
	0.50	0.3586	0.3543	0.3651	0.3664	0.3665	0.3810
	0.75	0.2808	0.2745	0.2887	0.3033	0.3026	0.3088
	1.00	0.2127	0.2102	0.2111	0.2484	0.2476	0.2502
$\rho = 0.8$	0.10	0.7833	0.7890	0.8219	0.7834	0.7844	0.8076
	0.25	0.7557	0.7601	0.7756	0.7562	0.7571	0.7708
	0.50	0.7020	0.6998	0.7042	0.7074	0.7074	0.7131
	0.75	0.6413	0.6381	0.6394	0.6577	0.6573	0.6597
	1.00	0.5812	0.5801	0.5805	0.6097	0.6093	0.6103

However, such approximations should not be used blindly. Numerical experiments indicate that the waiting-time probabilities are rather insensitive to more than the first two moments of the service time S provided that the squared coefficient of variation c_S^2 is not too large (say, $0 \leq c_S^2 \leq 2$) and the service-time density satisfies a reasonable shape constraint. The sensitivity becomes less and less manifest when the traffic intensity ρ gets closer to 1.

The motivation for the two-moment approximation is provided by the Pollaczek–Khintchine formula for the average delay in queue. The expression (9.2.10) for W_q can be written as

$$W_q = \frac{1}{2}(1 + c_S^2)\frac{E(S)}{1 - \rho}, \tag{9.2.20}$$

where $c_S^2 = \sigma^2(S)/E^2(S)$. Denote by $W_q(\exp)$ and $W_q(\det)$ the average delay in queue for the special cases of exponential services ($c_S^2 = 1$) and deterministic services ($c_S^2 = 0$). The formula (9.2.20) is equivalent to the representations

$$W_q = \frac{1}{2}(1 + c_S^2)W_q(\exp), \tag{9.2.21}$$

and

$$W_q = (1 - c_S^2)W_q(\det) + c_S^2 W_q(\exp). \tag{9.2.22}$$

A natural question is whether the representations (9.2.21) and (9.2.22) can be used as a basis for approximations to the waiting-time probabilities. Numerical investigations reveal that the waiting-time probabilities themselves do not allow for two-moment approximations of the forms (9.2.21) and (9.2.22), but the waiting-time percentiles do allow for such two-moment approximations. The pth percentile $\xi(p)$ of the waiting-time distribution function $W_q(x)$ is defined as the solution to $W_q(x) = p$. In statistical equilibrium the percentage of customers having a delay in queue no more than $\xi(p)$ is $100p\%$. Since $W_q(0) = 1 - \rho$, the percentile $\xi(p)$ is only defined for $1 - \rho \leq p < 1$. Denote by $\xi_{exp}(p)$ and $\xi_{det}(p)$ the percentile $\xi(p)$ for the cases of exponential services and deterministic services with the same means $E(S)$. The representation (9.2.21) suggests the first-order approximation

$$\xi_{app1}(p) = \frac{1}{2}(1 + c_S^2)\xi_{exp}(p), \tag{9.2.23}$$

while the representation (9.2.22) suggests the second-order approximation

$$\xi_{app2}(p) = (1 - c_S^2)\xi_{det}(p) + c_S^2 \xi_{exp}(p). \tag{9.2.24}$$

In Section 5.1 it was shown that $1 - W_q(x) = \rho \exp[-\mu(1 - \rho)x]$ for all $x \geq 0$ when the service time has an exponential distribution with mean $1/\mu = E(S)$. Hence $\xi_{exp}(p)$ is simply computed as $\xi_{exp}(p) = E(S)\ln[\rho/(1 - p)]/(1 - \rho)$.

A relatively simple algorithm for the computation of $\xi_{det}(p)$ is given in Section 9.6.2 in the more general context of the $M/D/c$ queue. For higher values

Table 9.2.2 The waiting-time percentiles $\eta(p)$

		$c_S^2 = 0.5$					$c_S^2 = 2$				
ρ	p	0.2	0.5	0.9	0.99	0.999	0.2	0.5	0.9	0.99	0.999
0.2	exa	0.25	0.70	2.06	3.90	5.73	0.32	1.20	4.53	9.30	14.1
	app1	0.21	0.65	2.16	4.32	6.48	0.42	1.30	4.32	8.63	13.0
	app2	0.26	0.73	1.98	3.87	5.76	0.31	1.14	4.67	9.52	14.4
0.5	exa	0.39	1.09	3.34	6.54	9.75	0.54	2.00	7.12	14.5	21.8
	app1	0.33	1.04	3.45	6.91	10.4	0.67	2.08	6.91	13.8	20.7
	app2	0.41	1.10	3.33	6.55	9.77	0.53	1.96	7.16	14.5	21.9
0.8	exa	0.91	2.64	8.52	16.9	25.4	1.53	5.14	17.49	35.2	52.8
	app1	0.84	2.60	8.63	17.3	25.9	1.67	5.20	17.27	34.5	51.8
	app2	0.93	2.63	8.52	16.9	25.4	1.50	5.14	17.49	35.2	52.9

of p (say $p \geq 1 - \frac{1}{2}\rho$) the percentile $\xi_{det}(p)$ can be simply computed from the asymptotic expansion of $W_q(x)$ for deterministic services.

Table 9.2.2 gives some numerical results. In the table we work with the percentiles of the waiting-time distribution of the delayed customers. The probability that a delayed customer has to wait longer than x is $[1 - W_q(x)]/P_{delay}$, where $P_{delay} = 1 - W_q(0)$. The percentile $\eta(p)$ is defined as the solution to

$$1 - \frac{1 - W_q(x)}{P_{delay}} = p.$$

The conditional percentiles $\eta(p)$ are defined for all $0 \leq p < 1$. Note that $\eta(p_1) = \xi(p_0)$ when $p_0 = 1-(1-p_1)\rho$. Table 9.2.2 gives the exact and approximate values of $\eta(p)$ for E_2 services ($c_S^2 = 0.5$) and H_2 services with the gamma normalization ($c_S^2 = 2$). The numerical results show an excellent performance of the second-order approximation for all values of ρ and p. The first-order approximation $(1/2)(1 + c_S^2)\eta_{exp}(p)$ is only useful for quick engineering calculations when ρ is not too small (say, $\rho > 0.5$) and p is sufficiently close to 1 (say, $p > 1 - \rho$).

9.2.3 Busy Period Analysis

The busy period is an important concept in queueing. A busy period begins when an arriving customer finds the system empty and ends when a departing customer leaves the system empty behind. In this subsection we derive the Laplace transform of the probability distribution of the length of a busy period in the $M/G/1$ queue. Also it will be seen that both the transient emptiness probability and the steady-state waiting-time distribution under the last-come first-served discipline are closely related to the distribution of a busy period.

Denote by the random variable B the length of a busy period and let $\beta(x)$ be the probability density of B. Then the Laplace transform

$$\beta^*(s) = \int_0^\infty e^{-sx}\beta(x)\,dx \ (= E(e^{-sB}))$$

of the busy period density is determined by the functional equation

$$\beta^*(s) = b^*(s + \lambda - \lambda\beta^*(s)), \tag{9.2.25}$$

where $b^*(s) = \int_0^\infty e^{-sx} b(x)\,dx$ is the Laplace transform of the probability density $b(x)$ of the service time of a customer. By relation (E.8) in Appendix E, the Laplace transform of $P\{B > x\}$ is given by

$$\int_0^\infty e^{-sx} P\{B > x\}\,dx = \frac{1 - \beta^*(s)}{s}. \tag{9.2.26}$$

The key to the proof of (9.2.25) is the assertion that the amount of time needed to empty the system when the system starts with n customers present is distributed as the sum of the lengths of n independent busy periods $B_1, \ldots, B_n$. To see this, note first that the order of service has no effect on the amount of time needed to empty the system. Following Takács (1962), imagine now the following service discipline. The initial n customers $C_1, \ldots, C_n$ are separated. Customer C_1 is served first, after which all customers (if any) are served who have arrived during the service time of customer C_1, and this way of service is continued until the system is free of all customers but $C_2, \ldots, C_n$. Next this procedure is repeated with customer C_2, etc. This verifies the above assertion. The remainder of the proof is now simple. Let the random variables S_1 and ν_1 denote the length of the service initiating the busy period and the number of customers arriving during that first service time. Then, by conditioning on S_1 and ν_1, we find

$$\begin{aligned} E(e^{-sB}) &= \int_0^\infty \left[\sum_{n=0}^\infty e^{-\lambda t}\frac{(\lambda t)^n}{n!} E(e^{-sB} \mid S_1 = t, \nu_1 = n)\right] b(t)\,dt \\ &= \int_0^\infty \left[\sum_{n=0}^\infty e^{-\lambda t}\frac{(\lambda t)^n}{n!} E(e^{-s(t+B_0+\cdots+B_n)})\right] b(t)\,dt, \end{aligned}$$

where $B_0 = 0$ and $B_1, \ldots, B_n$ are independent random variables each having the same distribution as the busy period B. Thus we find

$$\begin{aligned} \beta^*(s) &= \int_0^\infty \left[\sum_{n=0}^\infty e^{-\lambda t}\frac{(\lambda t)^n}{n!} e^{-st}[\beta^*(s)]^n\right] b(t)\,dt \\ &= \int_0^\infty e^{-st} e^{-\lambda[1-\beta^*(s)]t} b(t)\,dt = b^*(s + \lambda - \lambda\beta^*(s)), \end{aligned}$$

as was to be proved. In the same way as (9.2.25) was derived, we can derive the generating function of the random variable N which is defined as the number of customers served in one busy period. Letting $F(z) = \sum_{k=0}^\infty P\{N = k\} z^k$, it is left to the reader to verify that

$$F(z) = z b^*(\lambda - \lambda F(z)), \quad |z| \le 1. \tag{9.2.27}$$

Using relation (E.2) in Appendix E, it easily follows from (9.2.25) that the first two moments of the length of a busy period are given by

$$E(B) = \frac{E(S)}{1-\rho} \quad \text{and} \quad E(B^2) = \frac{E(S^2)}{(1-\rho)^3}, \tag{9.2.28}$$

where the random variable S denotes the service time of a customer. The result (9.2.28) shows that the squared coefficient of variation of the length of a busy period equals $c_B^2 = (1+c_S^2)/(1-\rho)$, where c_S^2 is the squared coefficient of variation of the service time S. The value of c_B^2 explodes when ρ approaches 1. Consequently, the density of the busy period has a very long tail for ρ close to 1. As an illustration, consider the case of gamma services with $E(S) = 1$ and $c_S^2 = 2$. Then the tail probability $P\{B > 1000\}$ has the respective values 4.70×10^{-4}, 3.63×10^{-3} and 1.15×10^{-2} for $\rho = 0.90$, 0.95 and 0.99. These values have been computed by using the general formula

$$P\{B \le x\} = \sum_{n=1}^{\infty} \int_0^x e^{-\lambda y} \frac{(\lambda y)^{n-1}}{n!} b_n(y)\, dy, \quad x \ge 0, \tag{9.2.29}$$

where $b_n(x)$ denotes the probability density of the sum $S_1 + \cdots + S_n$ of n service times $S_1, \ldots, S_n$. The reader is referred to Takács (1962) for a proof of this formula. The numerical evaluation of this infinite series offers no difficulties when the service time has a gamma distribution. Then $b_n(x)$ is a gamma density as well, so that each term of the series can be written as an incomplete gamma integral; see Appendix B. Fast codes for the numerical evaluation of an incomplete gamma integral are widely available.

If the service times are not gamma distributed, one has to resort to numerical inversion of the Laplace transform (9.2.26) for the computation of $P\{B > x\}$. In inverting this Laplace transform, the problem is that $\beta^*(s)$ is not explicitly given but is given in the form of a functional equation. However, the value of $\beta^*(s)$ for a given point s can be simply computed by an iterative procedure.

Iterative procedure for $\beta^*(s)$

For a given point s, the function value $\beta^*(s)$ can be seen as a 'fixed point' of the equation

$$z = b^*(s + \lambda - \lambda z).$$

It was shown in Abate and Whitt (1992) that this equation can be solved by repeated substitution. Starting with $z_0 = 1$, compute the (complex) number z_n from

$$z_n = b^*(s + \lambda - \lambda z_{n-1}), \quad n = 1, 2, \ldots .$$

The sequence $\{z_n\}$ converges to the desired value $\beta^*(s)$.

Transient emptiness probability

The distribution of the length of the busy period is closely related to the transient emptiness probability $p_{00}(t)$ defined by

$$p_{00}(t) = P\{\text{no customers will be present at time } t \text{ when at the current epoch 0 the system is empty}\}$$

for $t \geq 0$. Defining the Laplace transform $p_{00}^*(s)$ by

$$p_{00}^*(s) = \int_0^\infty e^{-st} p_{00}(t)\, dt,$$

it holds that

$$p_{00}^*(s) = \frac{1}{\lambda + s - \lambda\beta^*(s)}. \tag{9.2.30}$$

The derivation is simple. By conditioning on the epoch of the first arrival and on the length of the subsequent busy period, it is readily seen that

$$p_{00}(t) = e^{-\lambda t} + \int_0^t h(t-x)\lambda e^{-\lambda x}\, dx, \quad t \geq 0,$$

where

$$h(u) = \int_0^u p_{00}(u-v)\beta(v)\, dv.$$

Taking the Laplace transform of both sides of the integral equation for $p_{00}(t)$ and using the convolution formula (E.6) in Appendix E, we obtain

$$p_{00}^*(s) = \frac{1}{s+\lambda} + \frac{\lambda}{s+\lambda} p_{00}^*(s)\beta^*(s).$$

Solving this equation gives the desired result (9.2.30).

Waiting-time probabilities for LCFS service

Under the last-come first-served discipline (LCFS) the latest arrived customer enters service when the server is free to start a new service. The LCFS discipline was in fact used in the derivation of the Laplace transform of the busy period. It will therefore be no surprise that under this service discipline the limiting distribution of the waiting time of a customer can be related to the distribution of the length of a busy period. Assuming the LCFS discipline, let D_n be the delay in queue of the nth arriving customer and let $W_q(x) = \lim_{n\to\infty} P\{D_n \leq x\}$. Then

$$\int_0^\infty e^{-sx}\{1 - W_q(x)\}\, dx = \frac{1}{s}\left\{\rho - \frac{\lambda(1-\beta^*(s))}{s + \lambda - \lambda\beta^*(s)}\right\}. \tag{9.2.31}$$

We give only a sketch of the proof. Let the random variable $D^{(\infty)}$ have $W_q(x)$ as probability distribution function. By relation (E.8) in Appendix E,

$$\int_0^\infty e^{-sx}\{1 - W_q(x)\}\,dx = \frac{1 - E(e^{-sD^{(\infty)}})}{s}.$$

To find $E(e^{-sD^{(\infty)}})$, let the random variable U_n be 0 if the server is idle upon the nth arrival and let U_n be the remaining service time of the service in progress upon the epoch of the nth arrival otherwise. Under the LCFS discipline, the delay D_n of the nth arrival depends only on U_n. The random variable D_n has a positive mass at $x = 0$. Thus

$$E(e^{-sD_n}) = P\{U_n = 0\} + E(e^{-sD_n} \mid U_n > 0)P\{U_n > 0\}.$$

Next the following observation is made. Under the condition that $U_n = u$ and that k new customers arrive during the remaining service time u, the delay in queue of the nth arrival is distributed as $u + \sum_{i=1}^{k} B_i$, where $B_1, \ldots, B_k$ are independent random variables each distributed as the length of a busy period. Hence

$$E(e^{-sD_n} \mid U_n = u) = \sum_{k=0}^{\infty} e^{-\lambda u}\frac{(\lambda u)^k}{k!}e^{-su}[\beta^*(s)]^k = e^{-[s+\lambda(1-\beta^*(s))]u}.$$

Define now the random variable R_t as the remaining service time of the service in progress at time t given that the server is busy at time t. Using the PASTA property, it follows that

$$\lim_{n\to\infty} P\{U_n = 0\} = 1 - \rho \quad \text{and} \quad \lim_{n\to\infty} P\{U_n \le u \mid U_n > 0\} = \lim_{t\to\infty} P\{R_t \le u\}.$$

Using the result

$$\lim_{t\to\infty} P\{R_t \le u\} = \frac{1}{E(S)}\int_0^u \{1 - B(y)\}\,dy, \quad u \ge 0, \tag{9.2.32}$$

it is a matter of some algebra to verify that

$$E(e^{-sD^{(\infty)}}) = \lim_{n\to\infty} E(e^{-sD_n}) = 1 - \rho + \frac{\lambda(1 - \beta^*(s))}{s + \lambda - \lambda\beta^*(s)}.$$

This result gives (9.2.31). A remark is made about the important result (9.2.32). It is tempting to conclude this result by considering only those times when the server is busy and next using the equilibrium excess distribution from renewal theory; see Theorem 8.2.5. However, more subtle renewal-theoretic arguments are needed to prove (9.2.32). A probabilistic proof is as follows. Fix $u \ge 0$. Let the random variable $I(t) = 1$ if the server is busy at time t and the remaining service time of the service in progress is larger than u and let $I(t) = 0$ otherwise. The stochastic process $\{I(t)\}$ is regenerative. The regeneration epochs are the service completion epochs at which the server becomes idle. The length of a regeneration

cycle is continuously distributed with a finite expectation. Thus, by Theorem 2.2.4, $\lim_{t\to\infty} P\{I(t) = 1\}$ exists and equals $E(D_1)/E(L_1)$, where L_1 is the length of one cycle and D_1 is the total amount of time in one cycle that a service is in progress with a remaining service time larger than u. Denoting by N the number of customers served in one cycle and using Wald's equation, we find

$$E(D_1) = E(N)\int_u^\infty (y-u)b(y)\,dy = E(N)\int_u^\infty \{1 - B(y)\}\,dy.$$

By (9.2.27) and (9.2.28), $E(N) = 1/(1-\rho)$ and $E(L_1) = 1/\lambda + E(S)/(1-\rho)$. This gives

$$\lim_{t\to\infty} P\{\text{the server is busy at time } t \text{ and the remaining service time of the service in progress is larger than } u\} = \lambda\int_u^\infty \{1 - B(y)\}\,dy.$$

Noting that $\lim_{t\to\infty} P\{\text{the server is busy at time } t\}$ exists and equals $\rho = \lambda E(S)$, the result (9.2.32) follows.

9.2.4 Work in System

Let the random variable V_t be defined by

$$V_t = \text{the total amount of work that remains to be done on all customers in the system at time } t.$$

In other words, V_t is the sum of the remaining service times of the customers in the system at time t. The stochastic process $\{V_t, t \geq 0\}$ is called the work-in-system process or the virtual-delay process. Let

$$V_\infty(x) = \lim_{t\to\infty} P\{V_t \leq x\}, \quad x \geq 0.$$

Also, $V_\infty(x)$ is the long-run fraction of time that the work in system is no more than x. By the PASTA property, it holds that $V_\infty(x)$ is identical to the limiting distribution function $W_q(x)$ of the waiting time of a customer when service is in order of arrival. In particular, by (2.5.13),

$$\int_0^\infty e^{-sx}\{1 - V_\infty(x)\}\,dx = \frac{\rho s - \lambda + \lambda b^*(s)}{s(s - \lambda + \lambda b^*(s))}, \tag{9.2.33}$$

where $b^*(s)$ is the Laplace transform of the service-time density $b(x)$. For later purposes, we mention here the following additional relations for $V_\infty(x)$:

$$V'_\infty(x) = \lambda V_\infty(x) - \lambda\int_0^x V_\infty(x-y)b(y)\,dy, \quad x > 0, \tag{9.2.34}$$

$$V'_\infty(x) = \lambda \frac{d}{dx} \int_0^x V_\infty(x-y)\{1 - B(y)\}\, dy, \quad x > 0. \tag{9.2.35}$$

Since $V_\infty(x) = W_q(x)$, these formulas follow from relations (8.4.2), (8.4.3) and (8.4.4); take $\sigma = 1$ in these relations. Also, by (8.4.9), it holds under Assumption 9.2.1 that

$$1 - V_\infty(x) \sim \gamma e^{-\delta x} \quad \text{as } x \to \infty, \tag{9.2.36}$$

where γ and δ are given by (9.2.17).

Unlike the waiting-time distribution, the distribution of the work in system is invariant among the so-called work-conserving queue disciplines. A queue discipline is called *work-conserving* when the amount of time a customer is in service is not affected by the queue discipline.

The maximum work in system during a busy period

Define the random variable V_{max} as

$$V_{max} = \text{the maximum amount of work in system during a busy period.}$$

A busy period is the time elapsed between the arrival epoch of a customer finding the system empty and the next epoch at which the system becomes empty. The following result holds:

$$P\{V_{max} > K\} = \frac{1}{\lambda} \frac{V'_\infty(K)}{V_\infty(K)}, \quad K > 0, \tag{9.2.37}$$

where $V'_\infty(x)$ is the derivative of $V_\infty(x)$ for $x > 0$. To prove this result, we fix $K > 0$ and define the probability $p_K(x)$ for $0 < x < K$ by

$$p_K(x) = \text{the probability that the work process } \{V_t\} \text{ reaches the level 0 before it exceeds the level } K \text{ when the current amount of work in system equals } x.$$

It will be shown that

$$p_K(x) = \frac{V_\infty(K-x)}{V_\infty(K)}, \quad 0 < x < K. \tag{9.2.38}$$

The proof of this result is as follows. If the amount of work in the system is $x < K$ upon arrival of a new customer, the workload remains below the level K only if the amount of work brought along by the customer is less than $K - x$. Thus, by conditioning on what may happen in a very small time interval of length $\Delta t = \Delta x$, we find

$$p_K(x + \Delta x) = (1 - \lambda \Delta x) p_K(x) + \lambda \Delta x \int_0^{K-x} p_K(x+y) b(y)\, dy + o(\Delta x).$$

This gives the following expression for the derivative of $p_K(x)$:

$$p'_K(x) = -\lambda p_K(x) + \lambda \int_0^{K-x} p_K(x+y)b(y)\,dy, \quad 0 < x < K.$$

Mimicking the derivation of (8.4.4) gives

$$p'_K(x) = \lambda \frac{d}{dx} \int_0^{K-x} p_K(x+y)\{1 - B(y)\}\,dy, \quad 0 < x < K.$$

Letting $q_K(x) = p_K(K-x)$ for $0 < x < K$, we thus have

$$q'_K(x) = \lambda \frac{d}{dx} \int_0^{x} q_K(x-y)\{1 - B(y)\}\,dy, \quad 0 < x < K.$$

This equation has a unique solution since it can be reduced to a renewal-type equation. Comparing this equation with equation (9.2.34) reveals that, for some constant c,

$$q_K(x) = cV_\infty(x), \quad 0 < x < K.$$

Since $\lim_{x\to 0} p_K(x) = 1$, the result (9.2.38) now follows. It remains to verify (9.2.37). To do so, note that

$$\begin{aligned} P\{V_{max} > K\} &= 1 - \int_0^K p_K(x)b(x)\,dx \\ &= \frac{V_\infty(K) - \int_0^K V_\infty(K-x)b(x)\,dx}{V_\infty(K)}. \end{aligned} \tag{9.2.39}$$

The numerator of the last expression equals $\lambda^{-1}V'_\infty(K)$ by relation (9.2.34). This completes the verification of (9.2.37).

The probability distribution (9.2.37) of V_{max} can be calculated by numerical inversion of the Laplace transforms of $V_\infty(x)$ and $V'_\infty(x)$. The Laplace transform of $1 - V_\infty(x)$ is given by (9.2.33). Letting $v_\infty(x)$ denote the derivative of $V_\infty(x)$ for $x > 0$ and noting that $V_\infty(x) = V_\infty(0) + \int_0^x v_\infty(y)\,dy$, we find

$$\int_0^\infty e^{-sx} v_\infty(x)\,dx = \frac{(1-\rho)\left[\lambda - \lambda b^*(s)\right]}{s - \lambda + \lambda b^*(s)}.$$

9.3 THE $M^X/G/1$ QUEUE

Queueing systems with customers arriving in batches rather than singly have many applications in practice, for example in telecommunication. A useful model is the single-server $M^X/G/1$ queue where batches of customers arrive according to a Poisson process with rate λ and the batch size X has a discrete probability distribution $\{\beta_j,\ j = 1, 2, \dots\}$ with finite mean β. The customers are served

individually by a single server. The service times of the customers are independent random variables with a common probability distribution function $B(t)$. Denoting by the random variable S the service time of a customer, it is assumed that the server utilization ρ defined by

$$\rho = \lambda\beta E(S)$$

is smaller than 1. The analysis for the $M/G/1$ queue can be extended to the $M^X/G/1$ queue. In Section 9.3.1 we give an algorithm for the state probabilities. The computation of the waiting-time probabilities is discussed in Section 9.3.2.

9.3.1 The State Probabilities

The stochastic process $\{L(t), t \geq 0\}$ describing the number of customers in the system is regenerative. The process regenerates itself each time an arriving batch finds the system empty. The cycle length has a continuous distribution with finite mean. Thus the process $\{L(t)\}$ has a limiting distribution $\{p_j\}$. The probability p_j can be interpreted as the long-run fraction of time that j customers are in the system. The probability p_0 allows for the explicit expression

$$p_0 = 1 - \rho. \tag{9.3.1}$$

To see this, we apply the 'reward principle' that was used in Section 2.3 to obtain Little's formula. Assume that the system earns a reward at rate 1 whenever a customer is in service. Then the average reward per time unit represents the fraction of time that the server is busy. The long-run average reward earned per customer is equal to $E(S)$, while the long-run average arrival rate of customers is $\lambda\beta$. Hence the long-run average reward earned per time unit equals $\lambda\beta E(S)$. The long-run fraction of time that the server is busy equals $1-p_0$. This shows that $1-p_0 = \lambda\beta E(S) = \rho$. A recursion scheme for the p_j is given in the following theorem.

Theorem 9.3.1 *The state probabilities p_j satisfy the recursion*

$$p_j = \lambda p_0 \sum_{s=1}^{j} \beta_s a_{j-s} + \lambda \sum_{k=1}^{j} \left(\sum_{i=0}^{k} p_i \sum_{s>k-i} \beta_s \right) a_{j-k}, \quad j = 1, 2, \ldots, \tag{9.3.2}$$

where

$$a_n = \int_0^\infty r_n(t)\{1 - B(t)\}\, dt, \quad n = 0, 1, \ldots$$

with $r_n(t) = P\{$a total of n customers will arrive in (0,t)$\}$.

Proof The proof is along the same lines as the proof of Theorem 9.2.1. The only modification is with respect to the up- and downcrossing relation (9.2.5). We now use the following up- and downcrossing argument: the number of downcrossings

from a state in the set $\{k+1, k+2, \dots\}$ to a state outside this set during one cycle equals the number of upcrossings from a state outside the set $\{k+1, k+2, \dots\}$ to a state in this set during one cycle. Thus relation (9.2.5) generalizes to

$$E(N_k) = \sum_{i=0}^{k} E(T_i)\lambda \sum_{s>k-i} \beta_s, \quad k = 0, 1, \dots .$$

The remainder of the proof is analogous to the proof of Theorem 9.2.1.

The recursion scheme (9.3.2) is not as easy to apply as the recursion scheme (9.2.1). The reason is that the computation of the constants a_n is quite burdensome. In general, numerical integration must be used, where each function evaluation in the integration procedure requires an application of Adelson's recursion scheme for the computation of the compound Poisson probabilities $r_n(t)$, $n \geq 0$; see Section 1.2.

The best general-purpose approach for the computation of the state probabilities is the discrete FFT method. An explicit expression for the generating function

$$P(z) = \sum_{j=0}^{\infty} p_j z^j, \quad |z| \leq 1$$

can be given. It is a matter of tedious algebra to derive from (9.3.2) that

$$P(z) = (1-\rho)\frac{1 - \lambda\alpha(z)\{1 - G(z)\}}{1 - \lambda\alpha(z)\{1 - G(z)\}/(1-z)}, \tag{9.3.3}$$

where

$$G(z) = \sum_{j=1}^{\infty} \beta_j z^j \quad \text{and} \quad \alpha(z) = \int_0^{\infty} e^{-\lambda\{1-G(z)\}t}(1 - B(t))\, dt.$$

The derivation uses that $e^{-\lambda\{1-G(z)\}t}$ is the generating function of the compound Poisson probabilities $r_n(t)$; see Theorem 1.2.1. Moreover, the derivation uses that the generating function of the convolution of two discrete probability distributions is the product of the generating functions of the two probability distributions. The other details of the derivation of (9.3.3) are left to the reader. For constant and phase-type services, no numerical integration is required to evaluate the function $\alpha(z)$ in the discrete FFT method.

Asymptotic expansion

The state probabilities allow for an asymptotic expansion when it is assumed that the batch-size distribution and the service-time distribution are not heavy-tailed. Let us make the following assumption.

Assumption 9.3.1 *(a) The convergence radius R of $G(z) = \sum_{j=1}^{\infty} \beta_j z^j$ is larger than 1. Moreover, $\int_0^\infty e^{st}\{1 - B(t)\}\,dt < \infty$ for some $s > 0$.*

(b) $\lim_{s \to B} \int_0^\infty e^{st}\{1 - B(t)\}\,dt = \infty$, where B is the supremum over all s with

$$\int_0^\infty e^{st}\{1 - B(t)\}\,dt] < \infty.$$

(c) $\lim_{x \to R_0} G(x) = 1 + B/\lambda$ for some number R_0 with $1 < R_0 \leq R$.

Under this assumption we obtain from Theorem C.1 in Appendix C that

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty, \tag{9.3.4}$$

where τ is the unique solution to the equation

$$\lambda\alpha(\tau)\{1 - G(\tau)\} = 1 - \tau \tag{9.3.5}$$

on $(1, R_0)$ and the constant σ is given by

$$\sigma = (1 - \rho)(1 - \tau)\left[\lambda\alpha'(\tau)\{1 - G(\tau)\} - \frac{(1 - \tau)G'(\tau)}{1 - G(\tau)} + 1\right]^{-1}. \tag{9.3.6}$$

A formula for the average queue size

The long-run average number of customers in queue is $L_q = \sum_{j=1}^{\infty}(j - 1)p_j$. Using the relation $P'(1) = \sum_{j=1}^{\infty} jp_j$, we obtain after some algebra from (9.3.3) that

$$L_q = \frac{1}{2}(1 + c_S^2)\frac{\rho^2}{1 - \rho} + \frac{\rho}{2(1 - \rho)}\left[\frac{E(X^2)}{E(X)} - 1\right],$$

where X denotes the batch size. Note that the first part of the expression for L_q gives the average queue size in the standard $M/G/1$ queue, while the second part reflects the additional effect of the batch size. The formula for L_q implies directly a formula for the long-run average delay in queue per customer. By Little's formula $L_q = \lambda\beta W_q$.

9.3.2 The Waiting-Time Probabilities

The concept of waiting-time distribution is more subtle for the case of batch arrivals than for the case of single arrivals. Let us assume that customers from each arrival group are numbered as $1, 2, \ldots$. Service to customers from the same arrival group is given in the order in which those customers are numbered. For customers from different batches the service is in order of arrival. Define the random variable D_n as the delay in queue of the customer who receives the nth service. In the batch-arrival queue, $\lim_{n \to \infty} P\{D_n \leq x\}$ need not exist. To see this, consider the particular case

of a constant batch size of 2. Then $P\{D_n > 0\} = 1$ for n even and $P\{D_n > 0\} < 1$ for n odd. The limit

$$W_q(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P\{D_k \leq x\}, \quad x \geq 0$$

always exists. To see this, fix x and imagine that a reward of 1 is earned for each customer whose delay in queue is no more than x. Using renewal-reward theory, it can be shown that the limit $W_q(x)$ exists and represents the long-run fraction of customers whose delay in queue is no more than x. If the batch size distribution is non-arithmetic, then $\lim_{n\to\infty} P\{D_n \leq x\}$ exists and equals $W_q(x)$.

Denote by

$$b^*(s) = \int_0^\infty e^{-sx} b(x)\, dx$$

the Laplace transform of the probability density $b(x)$ of the service time of a customer. Let $\beta_{SC}^*(s)$ be the Laplace transform of the probability density of the total time needed to serve all customers from one batch. It is left to the reader to verify that

$$\beta_{SC}^*(s) = \sum_{k=1}^{\infty} \beta_k [b^*(s)]^k = G(b^*(s)).$$

The following result now holds:

$$\int_0^\infty e^{-sx} \{1 - W_q(x)\}\, dx = \frac{1 - W_{SC}^*(s) W_r^*(s)}{s}, \tag{9.3.7}$$

where

$$W_{SC}(s^*) = \frac{(1-\rho)s}{s - \lambda + \lambda \beta_{SC}^*(s)} \quad \text{and} \quad W_r^*(s) = \frac{1 - G(b^*(s))}{\beta[1 - b^*(s)]}$$

with $\beta = \sum_{k=1}^{\infty} k\beta_k$ denoting the average batch size. The waiting-time probabilities $W_q(x)$ can be numerically obtained from (9.3.7) by using numerical Laplace inversion.

We give only a heuristic sketch of the proof of (9.3.7). A rigorous treatment is given in Van Ommeren (1988). An essential part of the proof is the following result. For $k = 1, 2, \ldots$, let

η_k = the long-run fraction of customers taking the kth position in their batch.

Then it holds that

$$\eta_k = \frac{1}{\beta} \sum_{j=k}^{\infty} \beta_j, \quad k = 1, 2, \ldots\ . \tag{9.3.8}$$

To prove this result, fix k and imagine that a reward of 1 is earned for each customer taking the kth position in its batch. Then the long-run average reward per customer is η_k by definition. By the renewal-reward theorem, the long-run average reward per customer equals the expected reward $\sum_{j=k}^{\infty} \beta_j$ earned for a single batch divided by the expected batch size β. This gives (9.3.8). Consider now a test customer belonging to a batch that arrives when the system has reached steady state. Denote by $D^{(\infty)}$ the delay in queue of this test customer. The delay $D^{(\infty)}$ can be written as $D^{(\infty)} = X_0 + X_1$, where X_0 is the delay caused by the customers present just before the batch of the test customer arrives and X_1 is the delay caused by customers belonging to the batch of the test customer. The random variables X_0 and X_1 are independent of each other and so $E(e^{-sD^{(\infty)}}) = E(e^{-sX_0})E(e^{-sX_1})$. Assuming that the position of the test customer in the batch is distributed according to $\{\eta_k\}$, we have by (9.3.8) that

$$E(e^{-sX_1}) = \sum_{k=1}^{\infty} \eta_k [b^*(s)]^{k-1} = \frac{1}{\beta} \sum_{k=1}^{\infty} [b^*(s)]^{k-1} \sum_{j=k}^{\infty} \beta_j$$

$$= \frac{1}{\beta} \sum_{j=1}^{\infty} \beta_j \sum_{k=1}^{j} [b^*(s)]^{k-1} = \frac{1 - G(b^*(s))}{\beta[1 - b^*(s)]}.$$

To find $E(e^{-sX_0})$, note that an arriving group of customers can be considered as a singly arriving supercustomer. The probability density of the total time to serve a supercustomer has the Laplace transform $\beta_{SC}^*(s)$. In other words, the delay in queue of the first customer of each batch can be described by a standard $M/G/1$ queue for which the service-time density has the Laplace transform $\beta_{SC}^*(s)$. Thus, using the result (2.5.12) for the $M/G/1$ queue,

$$E(e^{-sX_0}) = \frac{(1 - \rho)s}{s - \lambda + \lambda \beta_{SC}^*(s)}.$$

Since $\int_0^{\infty} e^{-sx}\{1 - W_q(x)\}\, dx = s^{-1}[1 - E(e^{-sD^{(\infty)}})]$ by relation (E.8) in Appendix E, we have now derived (9.3.7) heuristically.

Alternative algorithm

A simpler algorithm than numerical Laplace inversion can be given for the $M^X/D/1$ queue with deterministic services. This alternative algorithm is discussed in Section 9.5.3 in the more general context of the $M^X/D/c$ queue. A simple algorithm is also possible when the service time of a customer is a mixture of Erlangian distributions with the same scale parameters. In this case the service time of a customer can be interpreted as a random sum of independent phases each having an exponentially distributed length with the same mean. The $M^X/G/1$ queue with generalized Erlangian services is in fact an $M^Y/M/1$ queue in which the batch size Y is distributed as the total number of service phases generated by all customers in

one batch. For this particular $M^X/G/1$ queue the waiting-time probabilities $W_q(x)$ can be computed by a modification of the algorithm given in Example 5.5.1.

Approximations for the waiting-time probabilities

Suppose that Assumption 9.3.1 is satisfied and let $b(t)$ denote the density of the service-time distribution function $B(t)$. Then the following asymptotic expansion applies:

$$1 - W_q(x) \sim \gamma e^{-\delta x} \quad \text{as } x \to \infty,$$

where δ is the smallest positive solution to

$$\sum_{j=1}^{\infty} \beta_j \left\{ \int_0^{\infty} e^{\delta t} b(t)\, dt \right\}^j = 1 + \frac{\delta}{\lambda}$$

and the constant γ is given by

$$\gamma = \frac{(1-\rho)\delta}{\lambda\beta} \left[1 - \lambda \int_0^{\infty} te^{\delta t} b(t)\, dt \sum_{j=1}^{\infty} j\beta_j \left\{ \int_0^{\infty} e^{\delta t} b(t)\, dt \right\}^{j-1} \right]^{-1}$$
$$\times \left[1 - \int_0^{\infty} e^{\delta t} b(t)\, dt \right]^{-1}.$$

9.4 $M/G/1$ QUEUES WITH BOUNDED WAITING TIMES

In Section 9.2.4 we studied the limiting distribution function $V_\infty(x)$ of the work in system in the $M/G/1$ queue. This distribution function will play a key role in the analysis of both the finite-buffer $M/G/1$ queue with partial overflow and the $M/G/1$ queue with impatient customers.

9.4.1 The Finite-Buffer $M/G/1$ Queue

Consider the $M/G/1$ queue with a finite buffer, i.e. the finite dam model. Instead of a service time of a customer, we speak of the amount of work brought in by a customer. Customers arrive according to a Poisson process with rate λ. The amounts of work brought in by the customers are independent random variables having a common probability distribution function $B(x)$ with probability density $b(x)$. Denoting by μ the first moment of the amount of work brought in by a customer, it is assumed that $\rho = \lambda\mu$ is less than 1. Each customer puts their amount of work into a buffer. The buffer has a finite capacity of K. A customer who brings more work than can be stored in the buffer causes an overflow, where only the excess of work is lost (partial overflow). The buffer is emptied at a unity rate whenever there is work in the buffer. The finite-buffer $M/G/1$ queue has a variety of applications such as dam and production/inventory systems with a

finite storage space and telecommunication systems with a finite buffer for storing incoming data.

An important characteristic of the finite-buffer $M/G/1$ queue is

$$\pi(K) = \text{the long-run fraction of arrivals that cause a partial overflow.}$$

The following result can be proved:

$$\pi(K) = \frac{1}{\lambda} \frac{V'_\infty(K)}{V_\infty(K)}, \tag{9.4.1}$$

where $V_\infty(x)$ is defined in Section 9.2.4. It is remarkable that $\pi(K)$ is identical to the probability $P\{V_{max} > K\}$, where V_{max} is the maximal buffer content during a busy period in the infinite-buffer model; see relation (9.2.37). The proof of the result (9.4.1) is based on the proportionality relation

$$V_K(x) = \frac{V_\infty(x)}{V_\infty(K)} \quad \text{for } 0 \le x \le K, \tag{9.4.2}$$

where $V_K(x)$ is defined by

$$V_K(x) = \lim_{t \to \infty} P\{V_t^{(K)} \le x\}$$

with the random variable $V_t^{(K)}$ denoting the amount of work in the buffer at time t. We defer the proof of (9.4.2) to later. First we sketch how the result (9.4.1) can be obtained from the proportionality relation (9.4.2). A customer who finds an amount of work x in the buffer upon arrival causes an overflow only if the customer brings an amount of work larger than $K - x$. In statistical equilibrium the amount of work in the buffer seen by an arrival has $V_K(x)$ as probability distribution function by the PASTA property. Hence, by conditioning,

$$\pi(K) = \{1 - B(K)\} V_K(0) + \int_0^K \{1 - B(K - x)\} v_K(x)\, dx,$$

where $v_K(x)$ denotes the derivative of $V_K(x)$ for $x > 0$. Using (9.4.2), it is not difficult to verify by partial integration that

$$\pi(K) = \frac{1}{V_\infty(K)} \left[V_\infty(K) - \int_0^K V_\infty(K - x) b(x)\, dx \right].$$

By (9.2.34) the term between brackets equals $\lambda^{-1} V'_\infty(K)$, proving (9.4.1).

Assuming that the probability distribution function $B(x)$ satisfies Assumption 9.2.1, it follows from (9.2.36) that

$$\pi(K) \sim \frac{\gamma\delta}{\lambda} e^{-\delta K} \quad \text{as } K \to \infty,$$

where γ and δ are given by (9.2.17).

Derivation of the proportionality relation

Several proofs can be given for the proportionality relation (9.4.2). An insightful proof can be based on renewal-theoretic arguments. The workload process $\{V^{(K)}(t), t \geq 0\}$ regenerates itself each time the buffer becomes empty. Let a cycle be the time interval between two consecutive regeneration epochs. Denote by the random variable $T^{(K)}$ the length of one cycle and by the random variable $T^{(K)}(x)$ the amount of time in one cycle that the work in system is no more than x. The corresponding random variables for the workload process $\{V_t\}$ in the infinite-buffer $M/G/1$ queue are denoted by $T^{(\infty)}$ and $T^{(\infty)}(x)$. Then, by the theory of regenerative processes,

$$V_K(x) = \frac{E\left[T^{(K)}(x)\right]}{E(T^{(K)})} \quad \text{and} \quad V_\infty(x) = \frac{E\left[T^{(\infty)}(x)\right]}{E(T^{(\infty)})} \tag{9.4.3}$$

for $0 \leq x \leq K$. The crucial observation is that $T^{(K)}(x)$ has the same distribution as $T^{(\infty)}(x)$ for any $0 \leq x \leq K$. The assumption of Poisson arrivals and the assumption of *partial* overflow of work in excess of the buffer capacity are essential in order to establish this result. A rigorous proof requires a lot of technical machinery. The result can be made plausible as follows. In the infinite-buffer model the distribution of $T^{(\infty)}(x)$ for $0 \leq x \leq K$ does not depend on the duration of excursions of the workload process above the level K. The workload process in the infinite-buffer system returns to the level K after each upcrossing of the level K. However, by the lack of memory of the Poisson process, the situation in the infinite-buffer system at the epochs at which a return to level K occurs is probabilistically the same as in the finite-buffer system at the epochs at which an overflow of level K occurs. This explains why $T^{(K)}(x)$ and $T^{(\infty)}(x)$ have the same distribution for any $0 \leq x \leq K$. Thus we can conclude from (9.4.3) that, for the constant $\gamma = E[T^{(\infty)}]/E(T^{(K)})$,

$$V_K(x) = \gamma V_\infty(x), \quad 0 \leq x \leq K. \tag{9.4.4}$$

Since $V_K(K) = 1$, we next get the desired result (9.4.2). A rigorous proof of (9.4.4) can be found in Hooghiemstra (1987).

Other performance measures

Other performance measures of interest are:

$$f(K) = \text{the long-run fraction of input that overflows,}$$

$$I(K) = \text{the long-run average amount of work in the buffer.}$$

The following results hold:

$$f(K) = \frac{(1-\rho)[1 - V_\infty(K)]}{\rho V_\infty(K)} \tag{9.4.5}$$

$$I(K) = K - \frac{1}{V_\infty(K)} \int_0^K V_\infty(x)\, dx. \tag{9.4.6}$$

The proof of (9.4.5) is based on Little's formula for the average number of busy servers. The long-run fraction of time the server is busy equals $1 - V_K(0) = 1 - V_\infty(0)/V_\infty(K)$. Hence, by Little's formula,

$$\lambda(1 - f(K))\mu = 1 - \frac{V_\infty(0)}{V_\infty(K)}.$$

Since $V_\infty(0) = 1-\rho$, the formula (9.4.5) next follows. Using partial integration, the result (9.4.6) directly follows by using (9.4.2). The performance measures $\pi(K)$, $f(K)$ and $I(K)$ can be calculated by using numerical Laplace inversion for the computation of $V_\infty(x)$, $V'_\infty(x)$ and $\int_0^x V_\infty(y)\, dy$ from the corresponding Laplace transforms. The formula (9.4.5) for the overflow probability $f(K)$ has an interesting form. It is our conjecture that this structural form provides a useful approximation to the overflow probability in more complex finite-buffer models such as the finite-buffer fluid model with a Markov modulated Poisson input process determined by a number of independent on-off sources. The solution of the infinite-buffer version of this model is given in the classic paper of Anick *et al.* (1982); see also Schwartz (1996). In this paper the linear differential equations for the work in system are solved through eigenvalues and eigenvectors.

9.4.2 An $M/G/1$ Queue with Impatient Customers

A queueing system often encountered in practice is one in which customers wait for service for a limited time and leave the system if service has not begun within that time. Practical examples of queueing systems with customer impatience include real-time telecommunication systems in which data received after a hard deadline are useless, telecommunication systems in which subscribers give up due to impatience before the requested connection is established and inventory systems with perishable goods.

In this subsection we consider an $M/G/1$ queue in which customers arrive according to a Poisson process with rate λ. The service or work requirements of the customers are independent random variables having a general probability distribution function $B(x)$ with finite mean μ. It is assumed that $\rho = \lambda\mu$ is less than 1. The service discipline is first-come first-served. Each arriving customer enters the system, but is only willing to wait in queue for a fixed time $\tau > 0$. A customer who waits for a time τ without his service having begun leaves the system after that time τ and becomes a lost customer. A basic measure for the quality of service in such a system is the fraction of customers who are lost. Define the performance measure P_{loss} by

$$P_{loss} = \text{the long-run fraction of customers who are lost.}$$

The following result holds:

$$P_{loss} = \frac{(1-\rho)P\{W_q^{(\infty)} > \tau\}}{1-\rho P\{W_q^{(\infty)} > \tau\}}, \tag{9.4.7}$$

where the random variable $W_q^{(\infty)}$ is distributed as the steady-state delay in queue of a customer in the standard $M/G/1$ queue with service in order of arrival. That is, $P\{W_q^{(\infty)} \leq x\} = W_q(x)$. The computation of $W_q(x)$ is discussed in Section 9.2.2. The proof of (9.4.7) is very similar to that of (9.4.1). To obtain the formula for P_{loss}, it is no restriction on the mathematical analysis to assume that customers finding an amount of work in system larger than τ upon arrival do not enter the system but are immediately lost. Using this convention, denote by the random variable $V_t^{(\tau)}$ the amount of work in system at time t and let $V^{(\tau)}(x) = \lim_{t\to\infty} P\{V_t^{(\tau)} \leq x\}$ for $x \geq 0$. Then, using the PASTA property,

$$P_{loss} = 1 - V^{(\tau)}(\tau). \tag{9.4.8}$$

By the same arguments as used to obtain (9.4.4), there is a constant γ so that

$$V^{(\tau)}(x) = \gamma V_\infty(x), \quad 0 \leq x \leq \tau. \tag{9.4.9}$$

To find the constant γ, we use Little's formula for the average number of busy servers. Since $1 - V^{(\tau)}(0)$ gives the fraction of time the server is busy,

$$\lambda(1 - P_{loss})\mu = 1 - V^{(\tau)}(0). \tag{9.4.10}$$

Since $V^{(\tau)}(0) = \gamma V_\infty(0)$ and $V_\infty(0) = 1 - \rho$, we obtain from (9.4.10) that

$$P_{loss} = \frac{(1-\rho)(\gamma - 1)}{\lambda\mu}. \tag{9.4.11}$$

Also, by (9.4.8), $P_{loss} = 1 - \gamma V_\infty(\tau)$ and so

$$\gamma = \frac{1}{1 - \rho\,[1 - V_\infty(\tau)]}. \tag{9.4.12}$$

Finally, the desired result (9.4.7) follows by substituting (9.4.12) in (9.4.11) and noting that $V_\infty(x)$ equals the waiting-time distribution function $W_q(x)$. Assuming that the service-time distribution function satisfies Assumption 9.2.1, it follows from (9.4.7) and the asymptotic expansion (9.2.16) that

$$P_{loss} \sim \frac{(1-\rho)\,\gamma e^{-\delta\tau}}{1 - \rho\gamma e^{-\delta\tau}} \sim (1-\rho)\,\gamma e^{-\delta\tau} \quad \text{as } \tau \to \infty,$$

where γ and δ are given by (9.2.17). In other words, P_{loss} decreases *exponentially fast* as τ gets larger. The structural form of (9.4.7) is remarkable. The loss probability is expressed in terms of the waiting-time probability $P\{W_q^{(\infty)} > \tau\}$. The latter probability represents for the $M/G/1$ queue without impatience the

probability that a customer arriving in steady state has to wait longer than a time τ when service is in order of arrival. The result (9.4.7) can be shown to hold for the $M/M/c$ queue with impatient customers as well; see Boots and Tijms (1999). In fact the result (9.4.7) applies to both the $M^X/G/1$ queue and the $M^X/M/c$ queue with impatient customers. In Section 9.8 the structural form (9.4.7) will again be encountered in queueing systems with finite buffers. It will be seen that the loss probabilities in a finite-buffer queue can often be expressed in terms of the solution for the corresponding infinite-buffer queue. This finding is extremely useful from a computational point of view: to analyse the finite-buffer model for different buffer sizes it suffices to compute only once the solution for the infinite-buffer model.

9.5 THE $GI/G/1$ QUEUE

This section deals with the *GI/G/1* queue in which the interarrival times and the service times both have a general probability distribution. The server utilization ρ is assumed to be smaller than 1. Computationally tractable results can be obtained only for special cases. However, the exact results for simpler models may be used as a basis for approximations to the complex *GI/G/1* model; see also the discussion in Section 9.7. The discussion will concentrate on the computation of the waiting-time probabilities for the cases of phase-type services and phase-type arrivals. For these cases the computational method is based on numerical Laplace inversion. The embedded Markov chain method is an alternative approach when the service times are distributed as a mixture of Erlangian distributions with the same scale parameters. The probabilistic approach for this particular case will be discussed first. The discussion below assumes that service is in order of arrival.

9.5.1 Generalized Erlangian Services

Suppose that the service-time density $b(t)$ is given by

$$b(t) = \sum_{i=1}^{m} q_i \frac{\mu^i t^{i-1} e^{-\mu t}}{(i-1)!}, \quad t \geq 0,$$

where $q_m > 0$. In other words, with probability q_i the service time of a customer is the sum of i independent phases each having an exponential distribution with mean $1/\mu$. Thus we can define the embedded Markov chain $\{X_n\}$ by

X_n = the number of uncompleted service phases just before the arrival of the nth customer.

Denoting by $\{\pi_j,\ j = 0, 1, \ldots\}$ the equilibrium distribution of this Markov chain, we find by the same arguments as used to derive (5.1.7) that

$$W_q(x) = 1 - \sum_{k=0}^{\infty} e^{-\mu x} \frac{(\mu x)^k}{k!} \left(1 - \sum_{j=0}^{k} \pi_j\right), \quad x \geq 0. \tag{9.5.1}$$

Thus we have a computationally useful algorithm for the waiting-time distribution when the probabilities π_j can be efficiently computed. These probabilities are the unique solution to the equilibrium equations

$$\pi_j = \sum_{k=0}^{\infty} \pi_k p_{kj}, \quad j = 0, 1, \ldots \tag{9.5.2}$$

together with the normalizing equation $\sum_{j=0}^{\infty} \pi_j = 1$, where the p_{ij} are the one-step transition probabilities of the Markov chain $\{X_n\}$. The p_{ij} are easily found. Since service completions of phases occur according to a Poisson process with rate μ as long as the server is busy, it is readily seen that for any $i \geq 0$

$$p_{ij} = \sum_{k=\max(j-i,1)}^{m} q_k \int_0^{\infty} e^{-\mu t} \frac{(\mu t)^{i+k-j}}{(i+k-j)!} a(t)\, dt, \quad 1 \leq j \leq i+m,$$

where $a(t)$ denotes the probability density of the interarrival time. The geometric tail approach from Section 3.4.2 can be used to reduce the infinite system of linear equations (9.5.2) to a finite system of linear equations. To see that

$$\frac{\pi_{j+1}}{\pi_j} \sim \eta \quad \text{as } j \to \infty \tag{9.5.3}$$

for some constant $0 < \eta < 1$, note that for any $i \geq 0$ the one-step transition probability p_{ij} equals 0 for $j > i + m$ and depends on i and j only through the difference $j - i$ for $j \geq 1$. Next we can apply a general result from Section 3.4.2 to obtain (9.5.3). Using the expression for p_{ij}, the equation (3.4.9) reduces after some algebra to

$$w^m - \left\{ \int_0^{\infty} e^{-\mu(1-w)t} a(t)\, dt \right\} \sum_{i=1}^{m} q_i w^{m-i} = 0. \tag{9.5.4}$$

The decay factor η is the largest root on (0,1) of this equation. By replacing π_j for $j \geq M$ by $\pi_M \eta^{j-M}$ for an appropriately chosen integer M, we obtain a finite system of linear equations.

9.5.2 Coxian-2 Services

Suppose that the service time S of a customer has a Coxian-2 distribution with parameters (b, μ_1, μ_2). That is, S is distributed as U_1 with probability $1 - b$ and S is distributed as $U_1 + U_2$ with probability b, where U_1 and U_2 are independent exponentials with respective means $1/\mu_1$ and $1/\mu_2$. Then the waiting-time distribution function $W_q(x)$ allows for the explicit expression

$$1 - W_q(x) = a_1 e^{-\eta_1 x} + a_2 e^{-\eta_2 x}, \quad x \geq 0, \tag{9.5.5}$$

where η_1 and η_2 with $0 < \eta_1 < \min(\mu_1, \mu_2) \leq \eta_2$ are the roots of

$$x^2 - (\mu_1 + \mu_2)x + \mu_1\mu_2 - \{\mu_1\mu_2 - (1-b)\mu_1 x\} \int_0^\infty e^{-xt} a(t)\, dt = 0. \quad (9.5.6)$$

The function $a(t)$ denotes the interarrival-time density and

$$a_1 = [-\eta_1^2\eta_2 + \eta_1\eta_2(\mu_1 + \mu_2) - \eta_2\mu_1\mu_2] / [\mu_1\mu_2(\eta_1 - \eta_2)]$$
$$a_2 = [\eta_1\eta_2^2 - \eta_1\eta_2(\mu_1 + \mu_2) + \eta_1\mu_1\mu_2] / [\mu_1\mu_2(\eta_1 - \eta_2)]\,.$$

A derivation of this explicit result can be found in Cohen (1982). In particular, P_{delay} and W_q are given by

$$P_{delay} = 1 - \frac{\eta_1\eta_2}{\mu_1\mu_2} \quad \text{and} \quad W_q = -\frac{(\mu_1 + \mu_2)}{\mu_1\mu_2} + \frac{1}{\eta_1} + \frac{1}{\eta_2}. \quad (9.5.7)$$

Since the computation of the roots of a function of a single variable is standard fare in numerical analysis, the above results are very easy to use for practical purposes. Bisection is a safe and fast method to compute the roots.

9.5.3 The $GI/Ph/1$ Queue

The results in Section 9.5.2 can be extended to the $GI/Ph/1$ queue with phase-type services. Let $b^*(s) = \int_0^\infty e^{-st} b(t)\, dt$ denote the Laplace transform of the service-time density $b(t)$. For phase-type service $b^*(s)$ can be written as

$$b^*(s) = \frac{b_1(s)}{b_2(s)}$$

for *polynomials* $b_1(s)$ and $b_2(s)$, where the degree of $b_1(s)$ is smaller than the degree of $b_2(s)$. Let m be the degree of $b_2(s)$. It is no restriction to assume that $b_1(s)$ and $b_2(s)$ have no common zeros and that the coefficient of s^m in $b_2(s)$ is equal to 1. Also, let $a^*(s) = \int_0^\infty e^{-st} a(t)\, dt$ denote the Laplace transform of the interarrival-time density $a(t)$. It is assumed that $a^*(s)$ and $b_2(s)$ have no common zero. In Cohen (1982) it has been proved that

$$\int_0^\infty e^{-sx}\{1 - W_q(x)\}\, dx = \frac{1}{s}\left\{1 - \frac{b_2(s)}{b_2(0)} \prod_{i=1}^{m} \frac{\eta_i}{\eta_i + s}\right\}, \quad (9.5.8)$$

where $\eta_1, \ldots, \eta_m$ are the roots of

$$b_2(-s) - a^*(s)b_1(-s) = 0 \quad (9.5.9)$$

in the right half-plane $\{s | \mathrm{Re}(s) > 0\}$. Moreover,

$$P_{delay} = 1 - \frac{1}{b_2(0)} \prod_{i=1}^{m} \eta_i \quad (9.5.10)$$

and

$$W_q = -\frac{b_2'(0)}{b_2(0)} + \sum_{i=1}^{m} \frac{1}{\eta_i}, \tag{9.5.11}$$

where $b_2'(0)$ is the derivative of $b_2(s)$ at $s = 0$. Once the roots $\eta_1, \ldots, \eta_m$ have been computed, the waiting-time probabilities can be obtained by numerical Laplace inversion of (9.5.8). A few words are in order on the computation of the (complex) roots $\eta_1, \ldots, \eta_m$. If the interarrival-time density is a phase-type density as well, then equation (9.5.9) reduces to a polynomial equation. Standard methods are available to compute the roots of a polynomial equation; see Appendix G. Another important case is the case of constant interarrival times. For the $D/Ph/1$ queue, equation (9.5.9) becomes

$$b_2(-s) - e^{-sD} b_1(-s) = 0. \tag{9.5.12}$$

For Coxian-2 services this equation is a special case of (9.5.6) and has two real roots that are easily found by bisection. In general the equation (9.5.12) can be numerically solved by tools discussed in Appendix G. In Appendix G we give special attention to the numerical solution of (9.5.12) when the service-time distribution is a mixture of an Erlang $(m-1, \mu)$ distribution and an Erlang (m, μ) distribution.

9.5.4 The $Ph/G/1$ Queue

For phase-type arrivals the Laplace transform $a^*(s) = \int_0^\infty e^{-st} a(t)\, dt$ of the probability density $a(t)$ of the interarrival time can be written as

$$a^*(s) = \frac{a_1(s)}{a_2(s)},$$

for *polynomials* $a_1(s)$ and $a_2(s)$, where the degree of $a_1(s)$ is lower than the degree of $a_2(s)$. Let m be the degree of $a_2(s)$. It is no restriction to assume that $a_1(s)$ and $a_2(s)$ have no common zeros and that the coefficient of s^m in $a_2(s)$ is equal to 1. Also, let $b^*(s) = \int_0^\infty e^{-st} b(t)\, dt$ denote the Laplace transform of the service-time density $b(t)$. It is assumed that $b^*(s)$ and $a_2(s)$ have no common zero. For the case of $m \geq 2$, it follows from results in Cohen (1982) that

$$\int_0^\infty e^{-sx} \left\{1 - W_q(x)\right\} dx = \frac{1}{s} \left\{1 - \frac{-\alpha a_2(0) s(1-\rho)}{a_2(-s) - b^*(s) a_1(-s)} \prod_{i=1}^{m-1} \frac{\delta_i - s}{\delta_i}\right\}, \tag{9.5.13}$$

where $\delta_1, \ldots, \delta_{m-1}$ are the roots of

$$a_2(-s) - b^*(s) a_1(-s) = 0 \tag{9.5.14}$$

in the right half-plane $\{s|\mathrm{Re}(s) > 0\}$ and

$$\alpha = \frac{a_2'(0) - a_1'(0)}{a_2(0)}.$$

As usual, $a_2'(0)$ and $a_1'(0)$ denote the derivatives of $a_2(s)$ and $a_1(s)$ at $s = 0$. Moreover,

$$P_{delay} = 1 - (1-\rho)\alpha a_2(0) \prod_{i=1}^{m-1} \delta_i \tag{9.5.15}$$

and

$$W_q = \frac{\rho}{2(1-\rho)E(S)} \left\{ E(S^2) + E(A^2) + 2E(S)\frac{a_1'(0)}{a_1(0)} - 2\alpha\frac{a_2'(0)}{a_2(0)} + \sum_{i=1}^{m-1} \frac{1}{\delta_i} \right\}, \tag{9.5.16}$$

where the random variables S and A represent the service time and the interarrival time. If $m = 1$ (i.e. Poisson input), formulas (9.5.13), (9.5.15) and (9.5.16) remain valid provided we put the empty product equal to 1 and the empty sum equal to 0. Note that there is a subtle difference between equations (9.5.9) and (9.5.14): equation (9.5.9) has m roots with $\mathrm{Re}(s) > 0$ and the other equation has $m-1$ roots. The explanation lies in the asymmetric role of the interarrival time A and the service time S in the ergodicity condition $E(S)/E(A) < 1$. For the numerical computation of the roots of equation (9.5.14) the same remarks apply as for equation (9.5.9). In particular, the $Ph/D/1$ queue is important. It will be seen in Section 9.7 that the waiting-time distribution in the multi-server $GI/D/c$ queue can be found through an appropriate $Ph/D/1$ queue.

9.5.5 Two-moment Approximations

The general $GI/G/1$ queue is very difficult to analyse. In general one has to resort to approximations. There are several approaches to obtain approximate numerical results for the waiting-time probabilities:

(a) Approximate the service-time distribution by a mixture of Erlangian distributions or a Coxian-2 distribution.

(b) Approximate the continuous-time model by a discrete-time model and use the discrete FFT method.

(c) Use two-moment approximations.

Approach (a) has been discussed in Sections 9.5.1 and 9.5.2. This approach should only be used when the squared coefficient of variation of the service time is not too large, say $0 \leq c_S^2 \leq 2$.

Let us now briefly discuss approach (b) for the $GI/G/1$ queue. This approach is based on Lindley's integral equation. Define the random variables

D_n = the delay in queue of the nth customer,

S_n = the service time of the nth customer,

A_n = the interarrival time between the nth and $(n+1)$th customers.

For ease, let us assume that the service times and interarrival times have probability densities $b(t)$ and $a(t)$. In the same way as in Section 8.4, we obtain

$$D_{n+1} = \max(0, D_n + U_n), \quad n = 1, 2, \ldots, \tag{9.5.17}$$

where $U_n = S_n - A_n$. Using this recurrence equation, it is not difficult to show that the waiting-time distribution function $W_q(x)$ satisfies the so-called Lindley integral equation

$$W_q(x) = \int_{-\infty}^{x} W_q(x-t)c(t)\,dt, \quad x \geq 0, \tag{9.5.18}$$

where $c(t)$ is the probability density of the U_n. Note that $c(t)$ is the convolution of $a(-t)$ and $b(t)$. A discretized version of Lindley's integral equation can be effectively solved by using the discrete FFT method. The details will not be given here, but can be found in Ackroyd (1980) and Tran-Gia (1986). In De Kok (1989) a moment-approximation method is suggested to solve Lindley's integral equation. This method is generally applicable and yields good approximations to the waiting-time probabilities. In particular, the moment-approximation method is well suited for both the $GI/D/1$ queue and the $D/G/1$ queue.

KLB approximation

Using a hybrid combination of basic queueing results and experimental analysis, the following two-moment approximations for the delay probability and the average delay in queue per customer were obtained by Krämer and Langenbach-Belz (1976):

$$P_{delay}^{KLB} = \rho + (c_A^2 - 1)\rho(1-\rho) \times \begin{cases} \dfrac{1 + c_A^2 + \rho c_S^2}{1 + \rho(c_S^2 - 1) + \rho^2(4c_A^2 + c_S^2)} & \text{if } c_A^2 \leq 1, \\ \dfrac{4\rho}{c_A^2 + \rho^2(4c_A^2 + c_S^2)} & \text{if } c_A^2 > 1, \end{cases}$$

$$W_q^{KLB} = \frac{\rho E(S)}{2(1-\rho)}(c_A^2 + c_S^2) \times \begin{cases} \exp\left\{\dfrac{-2(1-\rho)(1-c_A^2)^2}{3\rho(c_A^2 + c_S^2)}\right\} & \text{if } c_A^2 \leq 1, \\ \exp\left\{\dfrac{-(1-\rho)(c_A^2 - 1)}{c_A^2 + 4c_S^2}\right\} & \text{if } c_A^2 > 1. \end{cases}$$

Table 9.5.1 Some numerical results for the $GI/G/1$ queue

		$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.8$	
		P_{delay}	W_q	P_{delay}	W_q	P_{delay}	W_q
$D/E_4/1$	exact	0.000	0.000	0.047	0.017	0.446	0.319
	KLB	0.005	0.000	0.091	0.009	0.457	0.257
$D/E_2/1$	exact	0.001	0.000	0.116	0.078	0.548	0.757
	KLB	0.009	0.000	0.143	0.066	0.557	0.717
$E_4/D/1$	exact	0.009	0.002	0.163	0.050	0.578	0.386
	KLB	0.021	0.000	0.188	0.028	0.621	0.344
$E_2/D/1$	exact	0.064	0.024	0.323	0.177	0.702	0.903
	KLB	0.064	0.016	0.313	0.179	0.719	0.920
$E_2/H_2/1$	exact	0.110	0.203	0.405	1.095	0.752	4.825
	KLB	0.088	0.239	0.375	1.169	0.743	4.917
$H_2/E_2/1$	exact	0.336	0.387	0.650	1.445	0.870	5.281
	KLB	0.255	0.256	0.621	1.103	0.869	4.756

These approximations are only useful as rough estimates for practical engineering purposes provided that the traffic load on the system is not small and c_A^2 is not too large. In fact, one should be very careful in using the KLB approximation when c_A^2 is larger than 1. A reason for this is that performance measures in queueing systems are usually much more sensitive to the shape of the interarrival-time density than to the shape of the service-time density, particularly when the traffic load on the system is light. To illustrate the KLB approximation, Table 9.5.1 gives some numerical results. The H_2 distributions in the table refer to a hyperexponential distribution with gamma normalization and a squared coefficient of variation equal to 2.

9.6 MULTI-SERVER QUEUES WITH POISSON INPUT

Multi-server queues are notoriously difficult and a simple algorithmic analysis is possible only for special cases. In principle any practical queueing process could be modelled as a Markov process by incorporating sufficient information in the state description, but the dimensionality of the state space would grow quickly beyond any practical bound and would therefore obstruct an exact solution. In many situations, however, one resorts to approximation methods for calculating measures of system performance. Useful approximations for complex queueing systems are often obtained through exact results for simpler related queueing systems.

In this section we discuss both exact and approximate solution methods for the state probabilities and the waiting-time probabilities in multi-server queues with Poisson arrivals. The general $M/G/c$ queue does not allow for a tractable exact solution except for the special cases of the $M/M/c$ queue and the $M/D/c$ queue. The $M/M/c$ queue was analysed in detail in Section 5.1. An exact analysis for

the $M/D/c$ queue will be given in Section 9.6.1. In Section 9.6.2 we consider the $M/G/c$ queue with general service times and give several approximations including two-moment approximations based on exact results for the $M/M/c$ queue and the $M/D/c$ queue. In Section 9.6.3 we consider the $M^X/G/c$ queue with batch arrivals and general service times. In particular, the $M^X/M/c$ queue and the $M^X/D/c$ queue are dealt with.

9.6.1 The $M/D/c$ Queue

In this model the arrival process of customers is a Poisson process with rate λ, the service time of a customer is a constant D, and c identical servers are available. It is assumed that the server utilization $\rho = \lambda D/c$ is smaller than 1.

An exact algorithm analysis of the $M/D/c$ queue goes back to Crommelin (1932) and is based on the following observation. Since the service times are equal to the constant D, any customer in service at time t will have left the system at time $t + D$, while the customers present at time $t + D$ are exactly those customers either waiting in queue at time t or having arrived in $(t, t + D)$. Let $p_j(s)$ be the probability of having j customers in the system at time s. Then, by conditioning on the number of customers present at time t,

$$p_j(t + D) = \sum_{k=0}^{c} p_k(t) e^{-\lambda D} \frac{(\lambda D)^j}{j!} + \sum_{k=c+1}^{c+j} p_k(t) e^{-\lambda D} \frac{(\lambda D)^{j-k+c}}{(j-k+c)!}$$

for $j = 0, 1, \ldots$, since the number of arrivals in a time D is Poisson distributed with mean λD. Next, by letting $t \to \infty$ in these equations, we find that the time-average probabilities p_j satisfy the linear equations

$$p_j = e^{-\lambda D} \frac{(\lambda D)^j}{j!} \sum_{k=0}^{c} p_k + \sum_{k=c+1}^{c+j} p_k e^{-\lambda D} \frac{(\lambda D)^{j-k+c}}{(j-k+c)!}, \quad j \geq 0. \tag{9.6.1}$$

Also, we have the normalizing equation $\sum_{j=0}^{\infty} p_j = 1$. This infinite system of linear equations can be reduced to a finite system of linear equations by using the geometric tail approach discussed in Section 3.4.2. It will be shown below that the state probabilities p_j exhibit the geometric tail behaviour

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty, \tag{9.6.2}$$

where τ is the unique solution of the equation

$$e^{\lambda D(1-\tau)} \tau^c = 1 \tag{9.6.3}$$

on the interval $(1, \infty)$ and the constant σ is given by

$$\sigma = (c - \lambda D \tau)^{-1} \sum_{k=0}^{c-1} p_k (\tau^k - \tau^c). \tag{9.6.4}$$

Since $p_j/p_{j-1} \approx \tau^{-1}$ for j large enough, we replace p_j for $j \geq M$ by $p_M \tau^{-(j-M)}$ for an appropriately chosen integer M. Then the infinite system of linear equations (9.6.1) together with the normalizing equation $\sum_{j=0}^{\infty} p_j = 1$ is reduced to a finite system of linear equations of dimension $M+1$. A relatively small value of M is usually good enough for practical purposes. The value of M does not grow beyond any practical bound when the traffic load on the system gets close to 1. It is an empirical fact that the asymptotic expansion (9.6.2) already applies for relatively small values of j. For practical purposes the value $M = \frac{1}{2}(1+\rho)c + 10\rho\sqrt{c}$ seems large enough to obtain the state probabilities to at least nine decimal places (e.g. for $c = 25$ and $\rho = 0.99$ we have $M = 75$, which is in marked contrast with the brute-force value $N = 1056$ that is required when the infinite system of linear equations is truncated such that $\Sigma_{i=N}^{\infty} p_i \leq 10^{-9}$). In general the geometric tail approach leads to a relatively small system of linear equations that can usually be solved by a standard Gaussian elimination method. This approach requires that beforehand we compute the constant τ from (9.6.3). Using logarithms, the equation (9.6.3) is equivalent to $\lambda D(1-\tau) + c\ln(\tau) = 0$. Noting that $\lambda D = c\rho$ and using the transformation $\eta = 1/\tau$, it follows that τ can be obtained by computing the unique $\eta \in (0, 1)$ satisfying

$$\rho(1-\eta) + \eta\ln(\eta) = 0.$$

We can conclude that the state probabilities in the $M/D/c$ queue can be routinely computed by solving a finite system of linear equations. An accuracy check on the calculated values of the p_j is Little's relation

$$\sum_{j=1}^{c-1} j p_j + c\left(1 - \sum_{j=0}^{c-1} p_j\right) = \lambda D \tag{9.6.5}$$

for the average number of busy servers. An alternative and more advanced method for computing the state probabilities is based on the discrete FFT method. Before giving this method, we derive the generating function of the state probabilities. This generating function will also be used to verify the asymptotic expansion (9.6.2).

Generating function

Let $P(z) = \sum_{j=0}^{\infty} p_j z^j$ for $|z| \leq 1$. Multiplying both sides of (9.6.1) by z^j and summing over j gives

$$\begin{aligned} P(z) &= e^{\lambda D(z-1)} \sum_{k=0}^{c} p_k + \sum_{j=1}^{\infty} z^j \sum_{k=c+1}^{c+j} p_k e^{-\lambda D} \frac{(\lambda D)^{j-k+c}}{(j-k+c)!} \\ &= e^{\lambda D(z-1)} \sum_{k=0}^{c} p_k + \sum_{k=c+1}^{\infty} p_k z^{k-c} \sum_{j=k-c}^{\infty} e^{-\lambda D} \frac{(\lambda D)^{j-k+c}}{(j-k+c)!} z^{j-k+c} \end{aligned}$$

$$= e^{\lambda D(z-1)} \sum_{k=0}^{c} p_k + e^{\lambda D(z-1)} \sum_{k=c+1}^{\infty} p_k z^{k-c}$$

$$= e^{\lambda D(z-1)} z^{-c} \sum_{k=0}^{c} p_k z^c + e^{\lambda D(z-1)} z^{-c} \left[P(z) - \sum_{k=0}^{c} p_k z^k \right].$$

This gives the desired result

$$P(z) = \frac{\sum_{k=0}^{c-1} p_k (z^k - z^c)}{1 - z^c e^{\lambda D(1-z)}}. \tag{9.6.6}$$

The generating function $P(z)$ is the ratio of two functions that allow for an analytic continuation outside the unit circle. Next the asymptotic expansion (9.6.2) follows by applying Theorem C.1 in Appendix C. Also, we obtain after considerable algebra from (9.6.6) that the average queue size is given by

$$L_q = \frac{1}{2c(1-\rho)} \left[(c\rho)^2 - c(c-1) + \sum_{j=2}^{c-1} \{c(c-1) - j(j-1)\} p_j \right]. \tag{9.6.7}$$

An expression for the average delay in queue per customer next follows by using Little's formula $L_q = \lambda W_q$.

The discrete FFT method for the state probabilities

An alternative method for the computation of the probabilities p_j is to use the discrete FFT method. We cannot directly apply this method to (9.6.6) since the expression for $P(z)$ involves the unknowns $p_0, \ldots, p_{c-1}$. However, by a generally useful method, we can obtain from (9.6.6) an *explicit* expression for $P(z)$. The method is to compute first the zeros of the denominator on the right-hand side of (9.6.6) in the region $|z| \leq 1$ in the complex plane. The denominator $1 - z^c e^{\lambda D(1-z)}$ has c distinct zeros $z_0, z_1, \ldots, z_{c-1}$ inside or on the unit circle, where $z_0 = 1$. A simple algorithm for the computation of these roots is given in Appendix G. Each zero z_k must also be a zero of the numerator on the right-hand side of (9.6.6) for the simple reason that $P(z) = \sum_{j=0}^{\infty} p_j z^j$ is analytic for $|z| \leq 1$. Thus we can write (9.6.6) as

$$P(z) = \frac{\delta(z-1)}{1 - z^c e^{\lambda D(1-z)}} \prod_{k=1}^{c-1} (z - z_k)$$

for some constant δ. Since $P(1) = 1$, we find by using L'Hôpital's rule that

$$\delta = -c(1-\rho) / \prod_{k=1}^{c-1} (1 - z_k).$$

This gives for $P(z)$ the following *explicit* expression:

$$P(z) = \frac{c(1-\rho)(1-z)}{1 - z^c e^{\lambda D(1-z)}} \prod_{k=1}^{c-1} \left(\frac{z - z_k}{1 - z_k} \right), \quad |z| \leq 1. \tag{9.6.8}$$

This expression for $P(z)$ allows for a direct application of the FFT method. It is important to have an accuracy check on the calculated complex roots z_k and the subsequent calculations by the discrete FFT method. Such an accuracy check is provided by Little's relation (9.6.5). Another accuracy check is obtained by calculating the average queue size L_q both from formula (9.6.7) and from the direct expression $L_q = \sum_{j=c}^{\infty} (j-c) p_j$.

Waiting-time probabilities

In the paper of Crommelin (1932) an explicit expression has been derived for $W_q(x)$ in terms of an infinite alternating series. However, this explicit expression turns out to be of little computational use and is therefore not further discussed. It is possible to deduce a recursion scheme for $W_q(x)$ from Crommelin's original derivation, but this recursion scheme is also hampered by numerical difficulties. It took more than sixty years before a satisfying solution was found for the computation of $W_q(x)$. An elegant and numerically stable algorithm was found by Franx (2001) using an ingenious argument. The following expression holds for $W_q(x)$:

$$W_q(x) = \sum_{j=0}^{kc-1} Q_{kc-1-j} e^{-\lambda(kD-x)} \frac{[\lambda(kD-x)]^j}{j!}, \quad (k-1)D \leq x < kD \tag{9.6.9}$$

for $k = 1, 2, \ldots$, where

$$Q_j = \sum_{i=0}^{c+j} p_i, \qquad j = 0, 1, \ldots .$$

The first step in the proof is to assume that the arriving customers are assigned in cyclic order to the servers: the customers with labels $i, i+c, i+2c, \ldots$ are assigned to server i for $i = 1, \ldots, c$ (the nth arriving customer gets label n). This service discipline is not violating the assumption of service in order of arrival since the service times are deterministic. Denote by W_j the waiting time in queue of the customer with label j. It is assumed that there is a single queue in front of all c servers. Fix now $x > 0$. Also fix the positive integer k by $(k-1)D \leq x < kD$. Next choose any integer n such that $n > kc$. Consider now the customers with the labels $n - kc$ and n. Both customers are served by the same server. This server is called the marked server. To derive $P(W_n \leq x)$, we condition upon the number of waiting customers in the queue just after the epoch at which the customer with label $n - kc$ enters service with the marked server. Distinguish between two cases:

(a) There are at least kc customers waiting in queue just after the epoch at which the customer with label $n - kc$ enters service. Then the customer with label n must be among those waiting customers and its waiting time in queue is $D + (k-1)D$ which is larger than x.

(b) There are $i \leq kc - 1$ customers waiting in queue just after the epoch at which the customer with label $n - kc$ enters service. Denote this epoch by S^*. Since the customer with label n is the (kc)th customer to enter service after epoch S^*, the customer with label n is not yet present at epoch S^* and is the $(kc - i)$th customer to arrive after epoch S^*. Suppose that the customer with label n arrives at epoch $S^* + y$. Distinguish between the two cases $y < kD - x$ and $y \geq kD - x$.

(b1) $y < kD - x$. Since $y < kD - x \leq kD - (k-1)D = D$ the customer with label n arrives during the service time of the customer with label $n - kc$. Thus the waiting time in queue of the customer with label n equals $D - y + (k-1)D$, which is larger than x.

(b2) $y \geq kD - x$. The amount of time that the customer with label n spends in queue during the service time of the customer with label $n - kc$ equals $\max(D - y, 0)$. The customer with label n is the kth customer to be served by the marked server after the customer with label $n - kc$. Hence

$$\begin{aligned} W_n &\leq \max(D - y, 0) + (k-1)D \\ &\leq \max(D - (kD - x), 0) + (k-1)D \\ &= x - (k-1)D + (k-1)D = x. \end{aligned}$$

Denote now by S_j the epoch at which the customer with label j enters service and let L_j^+ be the number of customers waiting in queue just after epoch S_j, $j = 1, 2, \ldots$. Under the condition that $L_{n-kc}^+ = i$ with $i \leq kc - 1$ the customer with label n will be the $(kc - i)$th customer to arrive after epoch S_{n-kc}. Denote by A_n the number of arrivals during the interval $[S_{n-kc}, S_{n-kc} + kD - x]$. The random variable A_n is Poisson distributed with mean $\lambda(kD - x)$. The above arguments show that

$$W_n \leq x \quad \text{if and only if } L_{n-kc}^+ \leq kc - 1 \text{ and } A_n \leq kc - 1 - L_{n-kc}^+.$$

This leads to

$$P(W_n \leq x) = \sum_{i=0}^{kc-1} P(L_{n-kc}^+ = i) \sum_{\ell=0}^{kc-1-i} e^{-\lambda(kD-x)} \frac{[\lambda(kD-x)]^\ell}{\ell!}.$$

For fixed x and k, we now let $n \to \infty$. This gives

$$W_q(x) = \sum_{i=0}^{kc-1} q_i \sum_{\ell=0}^{kc-1-i} e^{-\lambda(kD-x)} \frac{[\lambda(kD-x)]^\ell}{\ell!}, \tag{9.6.10}$$

where $q_i = \lim_{j\to\infty} P(L_j^+ = i)$. It remains to find the limiting probabilities q_i. These limiting probabilities can be obtained by a simple up- and downcrossing argument: the long-run fraction of customers finding k other customers in queue upon arrival equals the long-run fraction of customers leaving k other customers behind in queue when entering service. This holds for any integer $k \geq 0$. For $k \neq 0$ we also have that the long-run fraction of arrivals finding k other customers in queue equals the long-run fraction of arrivals who find $k + c$ other customers in the system. This latter fraction equals the time-average probability p_{c+k} by the PASTA property. Hence we find

$$q_i = p_{c+i} \quad \text{for } i = 1, 2, \ldots \quad \text{and} \quad q_0 = \sum_{j=0}^{c} p_j.$$

Interchanging the order of summation in (9.6.10), the result (9.6.9) now follows.

Asymptotic expansion

It is also possible to give an asymptotic expansion for $1 - W_q(x)$:

$$1 - W_q(x) \sim \gamma e^{-\lambda(\tau-1)x} \quad \text{as } x \to \infty, \tag{9.6.11}$$

where

$$\gamma = \frac{\sigma}{(\tau - 1)\tau^{c-1}}$$

with τ and σ as in (9.6.3) and (9.6.4). To prove this result, we fix u with $0 \leq u < D$ and let x run through $(k-1)D + u$ for $k = 1, 2, \ldots$. Defining

$$b_r(u) = \sum_{j=0}^{r} Q_{r-j} e^{-\lambda(D-u)} \frac{[\lambda(D-u)]^j}{j!} \quad \text{for } r = 0, 1, \ldots,$$

we have by (9.6.9) that

$$1 - W_q(x) = 1 - b_{kc-1}(u) \quad \text{for} \quad x = (k-1)D + u.$$

Next consider the generating function $\overline{B}_u(z) = \sum_{r=0}^{\infty}(1 - b_r(u))z^r$. Since the generating function of the convolution of two discrete sequences is the product of the generating functions of the separate sequences, it follows that

$$\overline{B}_u(z) = \frac{1}{1-z} - Q(z)e^{\lambda(D-u)(z-1)},$$

where $Q(z) = \sum_{j=0}^{\infty} Q_j z^j$. Since $Q_j = \sum_{k=0}^{c+j} p_k$, we find after some algebra that

$$Q(z) = \frac{z^{-c}}{1-z}\left[P(z) - \sum_{k=0}^{c-1} p_k(z^k - z^c)\right] = \frac{e^{\lambda D(1-z)} \sum_{k=0}^{c-1} p_k(z^k - z^c)}{(1-z)(1 - z^c e^{\lambda D(1-z)})},$$

where the latter equality uses (9.6.6). This leads to

$$\overline{B}_u(z) = \frac{\left[1 - z^c e^{\lambda D(1-z)} - e^{\lambda u(1-z)} \sum_{k=0}^{c-1} p_k(z^k - z^c)\right]/(1-z)}{1 - z^c e^{\lambda D(1-z)}}.$$

Next, by Theorem C.1 in Appendix C and $\tau^c e^{\lambda D(1-\tau)} = 1$, we find

$$1 - b_j(u) \sim \frac{\sigma e^{-\lambda(\tau-1)u}}{\tau - 1} \tau^{-j} \quad \text{as } j \to \infty.$$

Take now $j = kc - 1$ and $x = (k-1)D + u$. Then $1 - b_j(u) = 1 - W_q(x)$. Since the equation $\tau^c e^{\lambda D(1-\tau)} = 1$ implies $\tau^{-(k-1)c} = e^{-\lambda(\tau-1)(k-1)D}$, we obtain

$$1 - b_{kc-1} \sim \frac{\sigma e^{-\lambda(\tau-1)x}}{(\tau-1)\tau^{c-1}} \quad \text{as } k \to \infty,$$

which proves the desired result (9.6.11).

9.6.2 The $M/G/c$ Queue

In this multi-server model with c servers the arrival process of customers is a Poisson process with rate λ and the service time S of a customer has a general probability distribution function $B(t)$. It is assumed that the server utilization $\rho = \lambda E(S)/c$ is smaller than 1.

The $M/G/c$ queue with general service times permits no simple analytical solution, not even for the average waiting time. Useful approximations can be obtained by the regenerative approach discussed in Section 9.2.1. In applying this approach to the multi-server queue, we encounter the difficulty that the number of customers left behind at a service completion epoch does not provide sufficient information to describe the future behaviour of the system. In fact we need the additional information of the elapsed service times of the other services (if any) still in progress. A full inclusion of this information in the state description would lead to an intractable analysis. However, as an approximation, we will aggregate the information of the elapsed service times in such a way that the resulting approximate model enables us to carry through the regenerative analysis. A closer look at the regenerative approach reveals that we need only a suitable approximation to the probability distribution of the time elapsed between service completions. We now make the following approximation assumption with regard to the behaviour of the process at the service completion epochs.

Assumption 9.6.1 (approximation assumption) *(a) If at a service completion epoch, k customers are left behind in the system with $1 \leq k < c$, then the time until the next service completion epoch is distributed as $\min(S_1^e, \ldots, S_k^e)$, where*

$S_1^e, \ldots, S_k^e$ are independent random variables that have the equilibrium excess distribution function

$$B_e(t) = \frac{1}{E(S)} \int_0^t \{1 - B(x)\} dx, \quad t \geq 0,$$

as probability distribution function.

(b) If at a service completion epoch, k customers are left behind in the system with $k \geq c$, then the time until the next service completion is distributed as S/c, where S denotes the original service time of a customer.

This approximation assumption can be motivated as follows. First, if not all c servers are busy, the $M/G/c$ queueing system may be treated as an $M/G/\infty$ queueing system in which a free server is immediately provided to each arriving customer. For the $M/G/\infty$ queue in statistical equilibrium it was shown by Takács (1962) that the remaining service time of any busy server is distributed as the residual life in a renewal process with the service times as the interoccurrence times. The same is true for the $M/G/1$ queue; see formula (9.2.32). The equilibrium excess distribution of the service time is given by $B_e(t)$; see Theorem 8.2.5. Second, if all of the c servers are busy, then the $M/G/c$ queue may be approximated by an $M/G/1$ queue in which the single server works c times as fast as each of the c servers in the original multi-server system. It is pointed out that the approximation assumption holds exactly for both the case of the $c = 1$ server and the case of exponentially distributed service times.

Approximations to the state probabilities

Under the approximation assumption the recursion scheme derived in Section 9.2.1 for the $M/G/1$ queue can be extended to the $M/G/c$ queue to yield approximations p_j^{app} to the state probabilities p_j. These approximations are given in the next theorem, whose lengthy proof may be skipped at first reading. The approximation to the state probabilities implies an approximation to the waiting-time probabilities. The latter approximation is discussed in Exercise 9.11.

Theorem 9.6.1 *Under the approximation assumption,*

$$p_j^{app} = \frac{(c\rho)^j}{j!} p_0^{app}, \quad j = 0, 1, \ldots, c-1, \tag{9.6.12}$$

$$p_j^{app} = \lambda a_{j-c} p_{c-1}^{app} + \lambda \sum_{k=c}^{j} b_{j-k} p_k^{app}, \quad j = c, c+1, \ldots, \tag{9.6.13}$$

where the constants a_n and b_n are given by

$$a_n = \int_0^\infty \{1 - B_e(t)\}^{c-1} \{1 - B(t)\} e^{-\lambda t} \frac{(\lambda t)^n}{n!} dt, \quad n = 0, 1, \ldots,$$

$$b_n = \int_0^\infty \{1 - B(ct)\} e^{-\lambda t} \frac{(\lambda t)^n}{n!}\, dt, \quad n = 0, 1, \ldots .$$

Proof In the same way as in the proof of Theorem 9.2.1, we find

$$p_j^{app} = \lambda p_0^{app} A_{0j} + \sum_{k=1}^{j} \lambda p_k^{app} A_{kj}, \quad j = 1, 2, \ldots , \tag{9.6.14}$$

where the constant A_{kj} is defined by

A_{kj} = the expected amount of time that j customers are present during the time until the next service completion epoch when a service has just been completed with k customers left behind in the system.

By the same argument as used to derive (9.2.7), we find under the approximation assumption that

$$A_{kj} = \int_0^\infty \{1 - B(ct)\} e^{-\lambda t} \frac{(\lambda t)^{j-k}}{(j-k)!}\, dt, \quad k \geq c \text{ and } j \geq k. \tag{9.6.15}$$

However, the problem is to find a tractable expression for A_{kj} when $0 \leq k \leq c - 1$. An explicit expression for A_{kj} involves a multidimensional integral when $0 \leq k \leq c - 1$. Fortunately, this difficulty can be circumvented by defining, for any $1 \leq k \leq c$ and $j \geq k$, the probability $M_{kj}(t)$ by

$M_{kj}(t) = P\{j - k$ customers arrive during the next t time units and the service of none of these customers is completed in the next t time units when only $c - k$ servers are available for the new arrivals$\}$.

Then, using the approximation assumption,

$$A_{kj} = \int_0^\infty \{1 - B_e(t)\}^k M_{kj}(t)\, dt, \quad 1 \leq k \leq c - 1, \ j \geq k. \tag{9.6.16}$$

Further, we have

$$A_{0j} = \int_0^\infty \{1 - B(t)\} M_{1j}(t)\, dt, \quad j \geq 1.$$

The definition of $M_{kj}(t)$ implies that

$$M_{kk}(t) = e^{-\lambda t}, \quad k \geq 1 \quad \text{and} \quad M_{cj}(t) = e^{-\lambda t} \frac{(\lambda t)^{j-c}}{(j-c)!}, \quad j \geq c.$$

Next we derive a differential equation for $M_{kj}(t)$ when $j > k$. By conditioning on what may happen in the first Δt time units, we find for any $1 \leq k \leq c - 1$ and $j > k$ that

$$M_{kj}(t + \Delta t) = (1 - \lambda \Delta t) M_{kj}(t) + \lambda \Delta t \{1 - B(t)\} M_{k+1,j}(t) + o(\Delta t), \quad t > 0.$$

Hence, for any $1 \le k \le c-1$ and $j > k$,

$$M_{kj}'(t) = -\lambda M_{kj}(t) + \lambda\{1 - B(t)\}M_{k+1,j}(t), \quad t > 0.$$

Multiplying both sides of this differential equation by $\{1-B_e(t)\}^k$, integrating over t and using (9.6.16), we find after partial integration that

$$A_{kj} = B_{k+1,j} - \frac{k}{\lambda E(S)} B_{kj}, \quad 1 \le k \le c-1, \ j > k, \tag{9.6.17}$$

where B_{kj} is a shorthand notation for

$$B_{kj} = \int_0^\infty \{1 - B_e(t)\}^{k-1}\{1 - B(t)\}M_{kj}(t)\,dt.$$

Next it is easy to establish the recursion scheme for p_j^{app}. To verify (9.6.12), we use induction. Obviously, (9.6.12) holds for $j = 0$. Suppose now that (9.6.12) holds for $j = 0, \ldots, n-1$ for some $1 \le n \le c-1$. Then, by (9.6.14) and (9.6.17)

$$\begin{aligned} p_n^{app}(1 - \lambda A_{nn}) &= \lambda p_0^{app} A_{0n} + \sum_{k=1}^{n-1} \lambda p_k^{app} \left\{ B_{k+1,n} - \frac{k}{\lambda E(S)} B_{kn} \right\} \\ &= \sum_{k=0}^{n-1} \lambda p_k^{app} B_{k+1,n} - \sum_{k=1}^{n-1} \lambda p_{k-1}^{app} B_{kn} = \lambda p_{n-1}^{app} B_{nn}, \end{aligned} \tag{9.6.18}$$

where the second equality uses $A_{0n} = B_{1n}$ and uses the induction assumption that $p_k^{app} = c\rho p_{k-1}^{app}/k$ for $1 \le k \le n-1$. Using partial integration it is readily verified that $B_{nn} = (1-\lambda A_{nn})E(S)/n$. Hence we obtain from (9.6.18) that $p_n^{app} = c\rho p_{n-1}^{app}/n$, which completes the induction step. To verify (9.6.13) we first note that

$$\lambda p_0^{app} A_{0j} + \sum_{k=1}^{c-1} \lambda p_k^{app} A_{kj} = \lambda p_{c-1}^{app} B_{cj}, \quad j \ge c. \tag{9.6.19}$$

The derivation of this relation is similar to that of (9.6.18). Inserting (9.6.19) into (9.6.14) and using (9.6.15), the desired result (9.6.17) follows.

Computational aspects

The recursion scheme for p_j^{app} is easy to apply in practice. In general the constants a_n and b_n have to be evaluated by numerical integration. An explicit expression for b_n can be given for deterministic and phase-type services. To compute the a_n, it is recommended to use Gauss–Legendre integration for deterministic services. To do so for phase-type services, the infinite integral for a_n must be first reduced to an integral over $(0, 1)$ by using that $E[g(X)] = E[g(F^{-1}(U))]$ when $F(x) = P\{X \le x\}$ and U is uniformly distributed on $(0, 1)$ (take $F(x) = B_e(x)$). The computational effort of the approximation algorithm depends only to a slight degree

on c, as opposed to exact methods for which the computing times quickly increase when c gets larger. For the first c state probabilities, we have

$$p_j^{app} = p_j^{exp}, \qquad j = 0, 1, \ldots, c-1, \tag{9.6.20}$$

where p_j^{exp} denotes the state probability p_j in the $M/M/c$ queue. To prove (9.6.20), sum both sides of (9.6.3) over $j \geq c$. This yields

$$\sum_{j=c}^{\infty} p_j^{app} = \frac{\rho}{1-\rho} p_{c-1}^{app}. \tag{9.6.21}$$

By (5.1.8) and (9.6.12),

$$p_j^{exp} = \frac{c\rho}{j} p_{j-1}^{exp} \quad \text{and} \quad p_j^{app} = \frac{c\rho}{j} p_{j-1}^{app} \quad \text{for } 1 \leq j \leq c-1,$$

Hence, for some constant γ, $p_j^{app} = \gamma p_j^{exp}$ for $0 \leq j \leq c-1$. To verify that $\gamma = 1$, we use (9.6.21) and (5.1.9) to obtain

$$\begin{aligned} \frac{\rho}{1-\rho} p_{c-1}^{app} &= 1 - \sum_{j=0}^{c-1} p_j^{app} = 1 - \sum_{j=0}^{c-1} \gamma p_j^{exp} = 1 - \gamma \left(1 - P_{delay}^{exp}\right) \\ &= 1 - \gamma + \frac{\gamma\rho}{1-\rho} p_{c-1}^{exp}. \end{aligned}$$

and so $\rho p_{c-1}^{app}/(1-\rho) = 1 - \gamma + \rho p_{c-1}^{app}/(1-\rho)$. This implies that $\gamma = 1$ and so (9.6.20) holds. The relation (9.6.20) says that the approximate queueing system behaves like an $M/M/c$ queue when not all of the c servers are busy. As a by-product of the above proof, we find for the delay probability $P_{delay} = \sum_{j=c}^{\infty} p_j$ that

$$P_{delay}^{app} = P_{delay}^{exp},$$

where P_{delay}^{exp} denote Erlang's delay probability in the $M/M/c$ queue. It has long been known that Erlang's delay probability gives a good approximation to the delay probability in the general $M/G/c$ queue. Further support for the quality of the approximation to the state probabilities p_j is provided by the result that p_j^{app}/p_{j-1}^{app} is asymptotically exact as $j \to \infty$. This result will be proved below.

The generating function

The algorithm in Section 5.1 gives a very simple scheme to compute $p_j^{app} = p_j^{exp}$ for $0 \leq j \leq c-1$. Define the generating function

$$P_q(z) = \sum_{j=0}^{\infty} p_{c+j}^{app} z^j, \qquad |z| \leq 1.$$

It is a matter of simple algebra to derive from (9.6.13) that

$$P_q(z) = \lambda p_{c-1}^{app} \frac{\alpha(z)}{1 - \lambda\beta(z)}, \tag{9.6.22}$$

where

$$\alpha(z) = \int_0^\infty \{1 - B_e(t)\}^{c-1}\{1 - B(t)\}e^{-\lambda(1-z)t}\,dt,$$
$$\beta(z) = \int_0^\infty \{1 - B(ct)\}e^{-\lambda(1-z)t}\,dt.$$

The discrete FFT method can be used to obtain the p_j^{app} for $j \geq c$.

Also, the generating function $P_q(z)$ enables us to obtain an approximation to the average queue size. Since $L_q = \sum_{j=c}^\infty (j-c)p_j$, the derivative $P_q'(1)$ yields an approximation to L_q. By differentiation of (9.6.22), we find after lengthy algebra that

$$L_q^{app} = \left[(1-\rho)\gamma_1 \frac{c}{E(S)} + \rho\frac{1}{2}(1 + c_S^2)\right] L_q(\text{exp}), \tag{9.6.23}$$

where $c_S^2 = \sigma^2(S)/E^2(S)$ and

$$\gamma_1 = \int_0^\infty \{1 - B_e(t)\}^c\,dt.$$

The quantity $L_q(\text{exp})$ denotes the average queue size in the $M/M/c$ queue. If $c_S^2 \leq 1$, the constant γ_1 is very well approximated by $(c+1)^{-1}c_S^2E(S) + c^{-1}(1-c_S^2)E(S)$. The approximation (9.6.23) has the term γ_1 in common with the approximation proposed in Boxma *et al.* (1979). This approximation improves the first-order approximation $\frac{1}{2}\left(1 + c_S^2\right) L_q(\text{exp})$ to L_q through

$$L_q^{Box} = \frac{1}{2}(1 + c_S^2)\frac{2L_q(\text{exp})L_q(\text{det})}{2\alpha L_q(\text{det}) + (1-\alpha)L_q(\text{exp})},$$

where $\alpha = \frac{1}{c-1}\left[\frac{E(S^2)}{\gamma_1 E(S)} - c - 1\right]$ and $L_q(\text{det})$ denotes the average queue size in the $M/D/c$ queue.

Table 9.6.1 gives for several examples the exact and approximate values of P_{delay} and L_q. We consider the cases of deterministic service ($c_S^2 = 0$), E_2 service ($c_S^2 = 0.5$) and H_2 service with the gamma normalization ($c_S^2 = 2$). In the table we also include the two-moment approximation

$$L_q^{app2} = (1 - c_S^2)L_q(\text{det}) + c_S^2 L_q(\text{exp}). \tag{9.6.24}$$

Table 9.6.1 Exact and approximate results

		$c_S^2 = 0$		$c_S^2 = 0.5$		$c_S^2 = 2$	
		P_{delay}	L_q	P_{delay}	L_q	P_{delay}	L_q
$c = 2$	exac	0.3233	0.177	0.3308	0.256	0.3363	0.487
$\rho = 0.5$	app	0.3333	0.194	0.3333	0.260	0.3333	0.479
	app2	—	0.176	—	0.255	—	0.491
$c = 5$	exa	0.1213	0.077	0.1279	0.104	0.1335	0.181
$\rho = 0.5$	app	0.1304	0.087	0.1304	0.107	0.1304	0.176
	app2	—	0.076	—	0.103	—	0.185
$c = 10$	exa	0.0331	0.024	0.0352	0.030	0.0373	0.048
$\rho = 0.5$	app	0.0361	0.025	0.0361	0.030	0.0361	0.047
	app2	—	0.023	—	0.030	—	0.049
$c = 2$	exa	0.7019	1.445	0.7087	2.148	0.7141	4.231
$\rho = 0.8$	app	0.7111	1.517	0.7111	2.169	0.7111	4.196
	app2	—	1.442	—	2.143	—	4.247
$c = 5$	exa	0.5336	1.156	0.5484	1.693	0.5611	3.250
$\rho = 0.8$	app	0.5541	1.256	0.5541	1.723	0.5541	3.191
	app2	—	1.155	—	1.686	—	3.277
$c = 25$	exact	0.1900	0.477	0.2033	0.661	0.2164	1.173
$\rho = 0.8$	approx	0.2091	0.495	0.2091	0.663	0.2091	1.178
	approx2	—	0.477	—	0.657	—	1.196
$c = 50$	exa	0.0776	0.214	0.0840	0.282	0.0908	0.471
$\rho = 0.8$	app	0.0870	0.207	0.0870	0.277	0.0870	0.488
	app2	—	0.211	—	0.279	—	0.485

This two-moment approximation can be found in Cosmetatos (1976) and Page (1972). The useful special-purpose approximation

$$L_q^{app}(\text{det}) = \frac{1}{2}\left[1 + (1-\rho)(c-1)\frac{\sqrt{4+5c}-2}{16c\rho}\right] L_q(\text{exp})$$

to $L_q(\text{det})$ was proposed in Cosmetatos (1976). The results in Table 9.6.1 for the approximation (9.6.24) use this approximation to $L_q(\text{det})$.

Asymptotic expansions

It is assumed that the probability distribution function $B_c(t) = B(ct)$ satisfies Assumption 9.2.1. In other words, the service-time distribution is not heavy-tailed. Let $B = \sup[s \mid \int_0^\infty e^{st}\{1 - B(ct)\}\,dt < \infty]$. Then, using (9.6.22) and Theorem C.1 in Appendix C, it is a routine matter to verify that

$$p_j^{app} \sim \sigma_{app}\tau^{-j} \quad \text{as } j \to \infty, \tag{9.6.25}$$

where τ is the unique solution to the equation

$$\int_0^\infty e^{-\lambda(1-\tau)t}\{1 - B(ct)\}\,dt = \frac{1}{\lambda} \tag{9.6.26}$$

on the interval $(1, 1 + B/\lambda)$. The constant σ_{app} is given by

$$\sigma_{app} = \frac{p_{c-1}^{app}\tau^{c-1}\int_0^\infty e^{-\lambda(1-\tau)t}\{1 - B_e(t)\}^{c-1}\{1 - B(t)\}dt}{\lambda\int_0^\infty te^{-\lambda(1-\tau)t}\{1 - B(ct)\}\,dt}. \tag{9.6.27}$$

In Section 9.7 we give asymptotic expansions for the state probabilities and the waiting-time probabilities in the general $GI/G/c$ queue. Using equation (9.6.26) and equation (9.7.4), it is not difficult to verify that p_j^{app}/p_{j-1}^{app} is asymptotically exact as $j \to \infty$. Also, an approximation to the asymptotic expansion of the waiting-time probabilities can be given. Using (9.6.25) and (9.7.1) to (9.7.4), we find

$$1 - W_q(x) \sim \gamma e^{-\lambda(\tau-1)x} \quad \text{as } x \to \infty, \tag{9.6.28}$$

where an approximation to γ is given by

$$\gamma_{app} = \frac{\sigma_{app}}{(\tau - 1)\tau^{c-1}}. \tag{9.6.29}$$

Two-moment approximations for the waiting-time percentiles

It is convenient to work with the percentiles $\eta(p)$ of the waiting-time distribution of the delayed customers. The percentiles $\eta(p)$ are defined for all $0 \le p < 1$; see Section 9.2.2. Just as in the $M/G/1$ case, we suggest the first-order approximation

$$\eta_{app1}(p) = \frac{1}{2}(1 + c_S^2)\eta_{exp}(p) \tag{9.6.30}$$

and the second-order approximation

$$\eta_{app2}(p) = (1 - c_S^2)\eta_{det}(p) + c_S^2\eta_{exp}(p), \tag{9.6.31}$$

where $\eta_{exp}(p)$ and $\eta_{det}(p)$ are the corresponding percentiles for the $M/M/c$ queue and the $M/D/c$ queue. Both approximations require that the squared coefficient of variation of the service time is not too large (say, $0 \le c_S^2 \le 2$) and the traffic load on the system is not very small. In the multi-server case the fraction of time that all servers are busy is an appropriate measure for the traffic load on the system. This fraction is given by P_{delay}. The second-order approximation (9.6.31) performs quite satisfactorily for all parameter values. The simple approximation (9.6.30) is only useful for quick engineering calculations when P_{delay} is not small and p is sufficiently close to 1 (say, $p > 1 - P_{delay}$). Table 9.6.2 gives for several examples the exact value and the approximate values (9.6.30) and (9.6.31) for the conditional waiting-time percentiles. It also includes the asymptotic value based on the approximation (9.6.28). We consider the cases of E_2 services ($c_S^2 = 0.5$) and H_2 services with gamma normalization ($c_S^2 = 2$).

Table 9.6.2 Conditional waiting-time percentiles

		$c_S^2 = 0.5$				$c_S^2 = 2$			
p		0.2	0.5	0.9	0.99	0.2	0.5	0.9	0.99
$c = 2$	exa	0.200	0.569	1.72	3.32	0.256	0.930	3.48	7.15
$\rho = 0.5$	app1	0.167	0.520	1.73	3.45	0.335	1.04	3.45	6.91
	app2	0.203	0.580	1.70	3.31	0.264	0.920	3.52	7.20
	asy	0.282	0.609	1.73	3.33	0.158	0.907	3.47	7.14
$c = 5$	exa	0.082	0.240	0.722	1.37	0.099	0.339	1.32	2.78
$\rho = 0.5$	app1	0.067	0.208	0.691	1.38	0.134	0.416	1.38	2.76
	app2	0.082	0.243	0.725	1.36	0.104	0.346	1.32	2.82
	asy	0.146	0.277	0.725	1.36	—	0.296	1.32	2.79
$c = 5$	exa	0.193	0.554	1.74	3.42	0.274	0.962	3.43	6.96
$\rho = 0.8$	app1	0.167	0.520	1.73	3.45	0.335	1.04	3.45	6.91
	app2	0.192	0.556	1.73	3.42	0.284	0.967	3.44	6.98
	asy	0.218	0.562	1.74	3.42	0.232	0.954	3.42	6.96
$c = 25$	exa	0.040	0.118	0.364	0.703	0.052	0.174	0.649	1.35
$\rho = 0.8$	app1	0.033	0.104	0.345	0.691	0.067	0.208	0.691	1.38
	app2	0.040	0.119	0.365	0.701	0.055	0.179	0.651	1.36
	asy	0.048	0.117	0.353	0.690	0.038	0.182	0.676	1.38

9.6.3 The $M^X/G/c$ Queue

In the $M^X/G/c$ queue the customers arrive in batches rather than singly. The arrival process of batches is a Poisson process with rate λ. The batch size has a probability distribution $\{\beta_j,\ j = 1, 2, \dots\}$ with finite mean β. The service times of the customers are independent of each other and have a general distribution with mean $E(S)$. There are c identical servers. It is assumed that the server utilization ρ, defined by

$$\rho = \frac{\lambda \beta E(S)}{c},$$

is smaller than 1. The customers from different batches are served in order of arrival and customers from the same batch are served in the same order as their positions in the batch. A computationally tractable analysis can only be given for the special cases of exponential services and deterministic services. We first analyse these two special cases. Next we discuss a two-moment approximation for the general $M^X/G/c$ queue.

The $M^X/M/c$ queue

The process $\{L(t)\}$ describing the number of customers present is a continuous-time Markov chain. Equating the rate at which the process leaves the set of states $\{i, i+1, \dots\}$ to the rate at which the process enters this set of states, we find for

the state probabilities p_j the recursion scheme

$$\min(i, c)\mu p_i = \sum_{k=0}^{i-1} p_k \lambda \sum_{s \geq i-k} \beta_s, \quad i = 1, 2, \ldots, \tag{9.6.32}$$

where $\mu = 1/E(S)$. Starting with $\overline{p}_0 := 1$, we successively compute $\overline{p}_1, \overline{p}_2, \ldots$ and next obtain the desired p_i by normalization. The normalization can be based on Little's relation

$$\sum_{j=0}^{c-1} j p_j + c(1 - \sum_{j=0}^{c-1} p_j) = c\rho \tag{9.6.33}$$

stating that the average number of busy servers equals $c\rho$. The computational effort of the recursion scheme can be reduced by using the asymptotic expansion

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty, \tag{9.6.34}$$

where τ is the unique solution of the equation

$$\lambda\tau[1 - \beta(\tau)] = c\mu(1 - \tau) \tag{9.6.35}$$

on the interval $(1, R)$ and the constant σ is given by

$$\sigma = \frac{(\tau - 1)\sum_{i=0}^{c-1}(c - i)p_i\tau^i/c}{1 - \lambda\tau^2\beta'(\tau)/(c\mu)}. \tag{9.6.36}$$

Here $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ and R is the convergence radius of the power series $\beta(z)$. To establish the asymptotic expansion, it is assumed that $R > 1$. In other words, the batch-size distribution is not heavy-tailed. The derivation of the asymptotic expansion (9.6.34) is routine. Define the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$, $|z| \leq 1$. It is a matter of simple algebra to derive from (9.6.32) that

$$P(z) = \frac{(1/c)\sum_{i=0}^{c-1}(c - i)p_i z^i}{1 - \lambda z\{1 - \beta(z)\}/\{c\mu(1 - z)\}}.$$

Next, by applying TheoremC.1 in Appendix C, we obtain (9.6.34).

From the generating function we also derive after considerable algebra that the long-run average queue size is given by

$$L_q = \frac{1}{c(1-\rho)} \sum_{j=1}^{c-1} j(c - j)p_j + \frac{\rho}{2(1-\rho)} \left\{ \frac{E(X^2)}{E(X)} - 1 \right\} + \frac{\rho}{1-\rho} - c\rho,$$

where the random variable X denotes the batch size.

Next we discuss the computation of the steady-state probability distribution function $W_q(x)$ of the waiting time of a customer. The function $W_q(x)$ is defined in the same way as in Section 9.3.2. To find $W_q(x)$, we need the probabilities

$$z_j = \text{the long-run fraction of customers who have } j \text{ other customers in front of them just after arrival}, \quad j = 0, 1, \ldots .$$

The delay in queue of a customer who has $j \geq c$ other customers in front of him just after arrival is the sum of $j - c + 1$ independent exponentials with common mean $1/(c\mu)$. Hence this conditional waiting time has an E_{j-c+1} distribution and so

$$1 - W_q(x) = \sum_{j=c}^{\infty} z_j \sum_{k=0}^{j-c} e^{-c\mu x} \frac{(c\mu x)^k}{k!}, \quad x \geq 0.$$

A computationally better representation for $W_q(x)$ is

$$1 - W_q(x) = \sum_{k=0}^{\infty} e^{-c\mu x} \frac{(c\mu x)^k}{k!} \left(1 - \sum_{j=0}^{k+c-1} z_j \right), \quad x \geq 0. \tag{9.6.37}$$

The probabilities z_j are easily expressed in terms of the p_j. To do so, let

$$\eta_k = \frac{1}{\beta} \sum_{j=k}^{\infty} \beta_j, \quad k = 1, 2, \ldots .$$

Then, as shown in Section 9.3.2, the probability η_k gives the long-run fraction of customers who take the kth position in their batch. Since the long-run fraction of batches finding m other customers present upon arrival equals p_m, we find

$$z_j = \sum_{m=0}^{j} p_m \eta_{j-m+1}, \quad j = 0, 1, \ldots .$$

For the case of exponential services this formula can be considerably simplified. Using the recursion relation (9.6.32), we have

$$z_j = \frac{\mu}{\lambda\beta} \min(j + 1, c) p_{j+1}, \quad j = 0, 1, \ldots . \tag{9.6.38}$$

This completes the specification of the exact algorithm (9.6.37) for the computation of the waiting-time probabilities $W_q(x)$. The computational effort can further be reduced by using an asymptotic expansion for $1 - W_q(x)$. Inserting (9.6.34) and (9.6.38) into (9.6.37), we find after some algebra that

$$1 - W_q(x) \sim \frac{\sigma \tau^{-c}}{\tau - 1} e^{-c\mu(1-1/\tau)x} \quad \text{as } x \to \infty, \tag{9.6.39}$$

where τ and σ are given by (9.6.35) and (9.6.36).

The $M^X/D/c$ queue

Suppose that the service time of each customer is a constant D. Denoting by $p_j(t)$ the probability that j customers are present at time t, we find by the same arguments as used in Section 9.6.2 that

$$p_j(t+D) = \sum_{k=0}^{c} p_k(t) r_j(D) + \sum_{k=c+1}^{c+j} p_k(t) r_{j-k+c}(D), \quad j = 0, 1, \ldots,$$

where the compound Poisson probability $r_j(D)$ is defined by

$$r_j(D) = \text{the probability that exactly } j \text{ customers arrive during a given time interval of length } D, \quad j = 0, 1, \ldots .$$

Letting $t \to \infty$, we find the system of linear equations

$$p_j = r_j(D) \sum_{k=0}^{c} p_k + \sum_{k=c+1}^{c+j} r_{j-k+c}(D) p_k, \quad j = 0, 1, \ldots \tag{9.6.40}$$

together with the normalizing equation $\sum_{j=0}^{\infty} p_j = 1$. Just as in the $M/D/c$ case, this infinite system of equations can be reduced to a finite system of linear equations by using the geometric tail behaviour of the p_j. It holds that

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty, \tag{9.6.41}$$

where τ is the unique root of the equation

$$\tau^c e^{\lambda D\{1-\beta(\tau)\}} = 1 \tag{9.6.42}$$

on the interval $(1, R)$ and the constant σ is given by

$$\sigma = [c - \lambda D \tau \beta'(\tau)]^{-1} \sum_{j=0}^{c-1} p_j(\tau^j - \tau^c). \tag{9.6.43}$$

As before, $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ and the number R denotes the convergence radius of the power series $\beta(z)$. It is assumed that $R > 1$.

In general, however, it is computationally simpler to compute the state probabilities p_j by applying the discrete FFT method to the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$. In the same way as (9.6.6) was derived, we obtain

$$P(z) = \frac{\sum_{j=0}^{c-1} p_j(z^j - z^c)}{1 - z^c e^{\lambda D\{1-\beta(z)\}}}, \tag{9.6.44}$$

since the generating function of the compound Poisson probabilities $r_j(D)$ is given by $e^{-\lambda D\{1-\beta(z)\}}$; see Theorem 1.2.1. Before the discrete FFT method can be applied, the unknown probabilities $p_0, \ldots, p_{c-1}$ must be removed from (9.6.44). To do so, we proceed in the same way as in Section 9.6.1 and rewrite $P(z)$ in the explicit form

$$P(z) = \frac{c(1-\rho)(1-z)}{1 - z^c e^{\lambda D\{1-\beta(z)\}}} \prod_{k=1}^{c-1} \left(\frac{z - z_k}{1 - z_k} \right), \tag{9.6.45}$$

where $z_0 = 1, z_1, \ldots, z_{c-1}$ are the c distinct roots of $z^c e^{\lambda D\{1-\beta(z)\}} = 1$ inside or on the unit circle. The computation of the (complex) roots $z_1, \ldots, z_{c-1}$ is discussed in Appendix G. The asymptotic expansion (9.6.41) follows from the generating function (9.6.44) and Theorem C.1 in Appendix C. Also, we obtain after considerable algebra from (9.6.44) that the long-run average queue size is given by

$$L_q = \frac{1}{2c(1-\rho)} \left[(c\rho)^2 - c(c-1) + \sum_{j=2}^{c-2} \{c(c-1) - j(j-1)\} p_j + c\rho \left(\frac{E(X^2)}{E(X)} - 1 \right) \right],$$

where the random variable X denotes the batch size. This relation can be used as an accuracy check on the calculated values of the probabilities p_j.

Waiting-time probabilities in the $M^X/D/c$ queue

In the batch-arrival $M^X/D/c$ queue, the waiting-time probability $W_q(x)$ is defined as the long-run fraction of customers whose waiting time in queue is no more than x, $x \geq 0$. The expression (9.6.9) for $W_q(x)$ in the $M/D/c$ queue can be extended to the $M^X/G/c$ queue. For any x with $(k-1)D \leq x < kD$ and $k = 1, 2, \ldots$, it holds that

$$W_q(x) = \sum_{m=0}^{kc-1} \eta_{m+1} \sum_{j=0}^{kc-1-m} Q_{kc-1-m-j}\ r_j(kD - x) \tag{9.6.46}$$

where $Q_j = \sum_{i=0}^{c+j} p_i$ for $j = 0, 1, \ldots$ and the probability η_r is defined by

$$\eta_r = \frac{1}{\beta} \sum_{j=r}^{\infty} \beta_j, \quad r = 1, 2, \ldots .$$

This result is due to Franx (2002). Its proof will be omitted. The asymptotic expansion

$$1 - W_q(x) \sim \gamma e^{-\lambda[\beta(\tau)-1]x} \quad \text{as } x \to \infty \tag{9.6.47}$$

holds with

$$\gamma = \frac{\sigma[\beta(\tau) - 1]}{(\tau - 1)^2 \tau^{c-1} \beta},$$

where τ and σ are given by (9.6.42) and (9.6.43). This result can be derived in a similar way as expansion (9.6.11) for the $M/D/c$ queue was obtained.

The $M^X/G/c$ queue

An exact and tractable solution for the $M^X/G/c$ queue is in general not possible except for the special cases of deterministic services and exponential services. Using the solutions for these special cases, we can give useful approximations for the general $M^X/G/c$ queue. A practically useful approximation to the average delay in queue per customer is

$$W_q^{app} = (1 - c_S^2) W_q(\text{det}) + c_S^2 W_q(\text{exp}),$$

provided that c_S^2 is not too large (say, $0 \leq c_S^2 \leq 2$) and the traffic load is not very small. It was pointed out in Section 9.3 that the first-order approximation $\frac{1}{2}(1 + c_S^2) W_q(\text{exp})$ is not applicable in the batch-arrival queue. A two-moment

Table 9.6.3 The percentiles $\eta(p)$ for the $M^X/E_2/c$ queue

			Constant batch size				Geometric batch size			
c	ρ	p	0.80	0.90	0.95	0.99	0.80	0.90	0.95	0.99
1	0.2	exa	2.927	3.945	4.995	7.458	5.756	8.122	10.49	15.98
		app	2.836	3.901	4.967	7.440	5.745	8.116	10.49	15.99
1	0.5	exa	5.107	7.170	9.231	14.02	9.044	12.84	16.64	25.45
		app	5.089	7.154	9.219	14.01	9.040	12.84	16.64	25.47
2	0.2	exa	1.369	1.897	2.431	3.661	2.989	4.172	5.355	8.101
		app	1.354	1.887	2.419	3.656	2.982	4.167	5.353	8.106
2	0.5	exa	2.531	3.561	4.592	6.985	4.600	6.498	8.395	12.80
		app	2.535	3.567	4.599	6.996	4.601	6.501	8.401	12.81
5	0.2	exa	0.621	0.845	1.063	1.560	1.298	1.773	2.246	3.345
		app	0.640	0.853	1.066	1.560	1.305	1.779	2.253	3.354
5	0.5	exa	1.063	1.476	1.889	2.846	1.898	2.657	3.417	5.179
		app	1.069	1.482	1.895	2.853	1.905	2.665	3.425	5.190
10	0.5	exa	0.553	0.764	0.971	1.451	0.980	1.360	1.740	2.622
		app	0.566	0.772	0.979	1.458	0.991	1.371	1.751	2.634
10	0.7	exa	0.923	1.295	1.667	2.530	1.547	2.181	2.815	4.287
		app	0.930	1.302	1.673	2.536	1.556	2.190	2.824	4.297

approximation to the percentiles $\eta(p)$ of the waiting-time distribution of the delayed customers is provided by

$$\eta_{app}(p) = (1 - c_S^2)\eta_{det}(p) + c_S^2\eta_{exp}(p), \quad 0 < p < 1.$$

However, it turns out that in the batch-arrival case the two-moment approximation to $\eta(p)$ works only for the higher percentiles. Fortunately, higher percentiles are usually the percentiles of interest in practice. Table 9.6.3 gives for the $M^X/E_2/c$ queue the exact and approximate values of the conditional waiting-time percentiles $\eta(p)$ both for the case of a constant batch size and the case of a geometrically distributed batch size. In both cases the mean batch size $E(X) = 3$. The normalization $E(S) = 1$ is used for the service time. The percentiles $\eta_{exp}(p)$ for exponential services and $\eta_{det}(p)$ for deterministic services have been computed from the asymptotic expansions (9.6.39) and (9.6.47). These asymptotic expansions already apply for moderate values of x provided the traffic load on the system is not very small. An appropriate measure for the traffic load is the probability that all servers are simultaneously busy. This probability is given by $P_B = 1 - \sum_{j=0}^{c-1} p_j$. As a rule of thumb, the asymptotic expansions can be used for practical purposes for $x \geq E(X)E(S)/\sqrt{c}$ when $P_B \geq 0.2$.

9.7 THE $GI/G/c$ QUEUE

It seems obvious that the general $GI/G/c$ queue offers enormous difficulties in getting practically useful results. Nevertheless, using specialized techniques for solving large-scale systems of linear equations for structured Markov chains, the continuous-time Markov chain approach has proved to be quite useful for an exact analysis of the $GI/G/c$ queue when the interarrival time and service time both have phase-type distributions; see also Van Hoorn and Seelen (1986) for an approximative analysis. By a detailed state description involving sufficient information about the number of customers present and the status of both the arrival in progress and the services in progress, it is possible to set up the equilibrium equations for the microstate probabilities. The resulting large-scale system of linear equations possesses a structure enabling the application of specialized algorithms to solve numerically the equations, provided the number of servers is not too large; see Seelen *et al.* (1985) and Takahashi and Takami (1976). However, this numerical approach is not suited to routine calculations. The specialized algorithms involve a clever use of asymptotic expansions for the $GI/G/c$ queue. It is assumed that the server utilization $\rho = \lambda E(S)/c$ is smaller than 1, where λ denotes the average arrival rate and $E(S)$ is the mean service time.

Asymptotic expansions

Under Assumption 9.2.1 with $B(t)$ replaced by $B(ct)$, asymptotic expansions can be given for the state probabilities p_j and the waiting-time probabilities $W_q(x)$. It

holds that

$$p_j \sim \sigma \tau^{-j} \quad \text{as } j \to \infty \tag{9.7.1}$$

and

$$1 - W_q(x) \sim \frac{\sigma \delta}{\lambda(\tau - 1)^2 \tau^{c-1}} e^{-\delta x} \quad \text{as } x \to \infty. \tag{9.7.2}$$

Assuming that the interarrival time and the service time have probability densities $a(x)$ and $b(x)$, the constant δ is the unique solution to the characteristic equation

$$\int_0^\infty e^{-\delta x} a(x)\, dx \int_0^\infty e^{\delta y/c} b(y)\, dy = 1 \tag{9.7.3}$$

on the interval $(0, B)$ with $B = \sup\{s \mid \int_0^\infty e^{st}\{1 - B(ct)\}\, dt < \infty\}$. The constant τ (> 1) is given by

$$\tau = \left[\int_0^\infty e^{-\delta x} a(x)\, dx \right]^{-1}. \tag{9.7.4}$$

An explicit expression for the constant σ cannot be given in general. A proof of the above asymptotic expansions is beyond the scope of this book. The asymptotic expansions were established by Takahashi (1981) for the case of a phase-type interarrival-time distribution and a phase-type service-time distribution. However, the class of phase-type distributions is dense in the class of all probability distributions on the non-negative axis. Thus, one might conjecture that the asymptotic expansions hold for a general interarrival-time distribution and a general service-time distribution provided that the service-time distribution is not heavy-tailed.

Two-moment approximations

In this section we restrict ourselves to the particular models of the $GI/M/c$ queue with exponential services and the $GI/D/c$ queue with deterministic services. These models allow for a relatively simple algorithmic analysis. The results for these models may serve as a basis for approximations to the complex $GI/G/c$ queue. Several performance measures P, such as the average queue length, the average waiting time per customer and the (conditional) waiting-time percentiles, can be approximated by using the familiar interpolation formula

$$P_{app} = (1 - c_S^2) P_{GI/D/c} + c_S^2 P_{GI/M/c} \tag{9.7.5}$$

provided c_S^2 is not too large and the traffic load on the system is not very light. In this formula $P_{GI/D/c}$ and $P_{GI/M/c}$ denote the exact values of the specific performance measure for the special cases of the $GI/D/c$ queue and the $GI/M/c$ queue with the same mean service time $E(S)$. Table 9.7.1 gives for several values

Table 9.7.1 Some numerical results for the $E_{10}/E_2/c$ queue

		$\rho = 0.5$			$\rho = 0.8$			$\rho = 0.9$		
c		L_q	$\eta(0.8)$	$\eta(0.95)$	L_q	$\eta(0.8)$	$\eta(0.95)$	L_q	$\eta(0.8)$	$\eta(0.95)$
1	exa	0.066	1.21	2.21	0.780	2.59	4.78	2.21	4.99	9.25
	app	0.082	1.19	2.17	0.813	2.57	4.76	2.25	5.14	9.25
5	exa	0.006	0.277	0.499	0.452	0.551	0.993	1.75	1.02	1.87
	app	0.009	0.243	0.452	0.466	0.530	0.968	1.76	1.02	1.86

of c and ρ the exact and approximate values of the average queue size L_q and the conditional waiting-time percentiles $\eta(0.8)$ and $\eta(0.95)$ for the $E_{10}/E_2/c$ queue. In all examples the normalization $E(S) = 1$ is used. The above linear interpolation formula is in general not to be recommended for the delay probability, particularly not when c_S^2 is close to zero. For example, the delay probability has the respective values 0.0776, 0.3285 and 0.3896 for the $E_{10}/D/5$ queue, the $E_{10}/E_2/5$ queue and the $E_{10}/M/5$ queue, each with $\rho = 0.8$. Interpolation formulas like the one above should always be accompanied by a caveat against their blind application. The above interpolation formula reflects the empirical finding that measures of system performance are in general much more sensitive to the interarrival-time distribution than to the service-time distribution, in particular when the traffic load is light.

9.7.1 The $GI/M/c$ Queue

In the $GI/M/c$ queue the service times of the customers are exponentially distributed with mean $1/\mu$. In addition to the time-average probabilities p_j, let

$$\pi_j = \text{the long-run fraction of customers who find } j \text{ other customers present upon arrival.}$$

There is a simple relation between the p_j and the π_j. We have

$$\min(j, c)\mu p_j = \lambda \pi_{j-1}, \quad j = 1, 2, \ldots . \tag{9.7.6}$$

This relation equates the average number of downcrossings from state j to state $j-1$ per time unit to the average number of upcrossings from state $j-1$ to state j per time unit; see also Section 2.7.

The probabilities π_j determine the waiting-time distribution function $W_q(x)$. Note that the conditional waiting-time of a customer finding $j \geq c$ other customers present upon arrival is the sum of $j - c + 1$ independent exponentials with mean $1/(c\mu)$ and thus has an Erlang distribution. Hence, by conditioning,

$$1 - W_q(x) = \sum_{j=c}^{\infty} \pi_j \sum_{k=0}^{j-c} e^{-c\mu x} \frac{(c\mu x)^k}{k!}, \quad x \geq 0. \tag{9.7.7}$$

This expression can be further simplified. To show this, we use that

$$\frac{\pi_{j+1}}{\pi_j} = \eta, \quad j \geq c-1 \tag{9.7.8}$$

for some constant $0 < \eta < 1$. The proof of this result is a replica of the proof of the corresponding result for the $GI/M/1$ queue; see (3.5.15). Hence

$$\pi_j = \eta^{j-c+1}\pi_{c-1}, \quad j \geq c-1. \tag{9.7.9}$$

As a by-product of (9.7.6) and (9.7.7) we have

$$p_j = \eta^{j-c} p_c, \quad j \geq c. \tag{9.7.10}$$

Substituting (9.7.9) into (9.7.8) yields

$$1 - W_q(x) = \frac{\eta}{1-\eta}\pi_{c-1}e^{-c\mu(1-\eta)x}, \quad x \geq 0. \tag{9.7.11}$$

The constant η is the unique solution of the equation

$$\eta = \int_0^\infty e^{-c\mu(1-\eta)t}a(t)\,dt \tag{9.7.12}$$

on the interval (0,1). To see this, note that $\{\pi_j\}$ is the equilibrium distribution of the embedded Markov chain describing the number of customers present just before an arrival epoch. Substituting (9.7.9) into the balance equations

$$\pi_j = \sum_{k=j-1}^{\infty} \pi_k \int_0^\infty e^{-c\mu t}\frac{(c\mu t)^{k+1-j}}{(k+1-j)!}a(t)\,dt, \quad j \geq c$$

easily yields the result (9.7.12).

By the relations (9.7.6), (9.7.9) and (9.7.10), the probability distributions $\{p_j\}$ and $\{\pi_j\}$ are completely determined once we have computed $\pi_0, \ldots, \pi_{c-1}$ or $p_0, \ldots, p_c$. These c unknowns can be rather easily computed for the special cases of deterministic, Coxian-2 and Erlangian interarrival times. If one is only interested in the waiting-time probabilities (9.7.11), these computations can be avoided. An explicit expression for the delay probability $\eta\pi_{c-1}/(1-\eta)$ is given in Takács (1962). For the case of $c = 1$ ($GI/M/1$ queue), $\eta\pi_{c-1}/(1-\eta) = \eta$.

Deterministic arrivals

Suppose there is a constant time D between two consecutive arrivals. Define the embedded Markov chain $\{X_n\}$ by

X_n = the number of customers present just before the nth arrival.

Denoting the one-step transition probabilities of this Markov chain by p_{ij}, the π_j are the unique solution to the equations

$$\pi_j = \sum_{k=j-1}^{\infty} \pi_k p_{kj}, \quad j = 1, 2, \ldots$$

together with the normalizing equation $\sum_{j=0}^{\infty} \pi_j = 1$. Substituting (9.7.9) into these equations yields that $\pi_0, \ldots, \pi_{c-1}$ are the unique solution to the finite system of linear equations

$$\pi_j = \sum_{k=j-1}^{c-2} \pi_k p_{kj} + \pi_{c-1} p^*_{c-1,\,j}, \quad 1 \le j \le c-1,$$

$$\sum_{j=0}^{c-2} \pi_j + \frac{\pi_{c-1}}{1-\eta} = 1, \tag{9.7.13}$$

where

$$p^*_{c-1,\,j} = \sum_{k=c-1}^{\infty} \eta^{k-c+1} p_{kj}, \quad 1 \le j \le c-1.$$

The constant η is the unique solution to the equation $\eta = \exp[-c\mu D(1-\eta)]$ on the interval (0,1). It remains to specify the p_{kj} for $1 \le j \le c-1$. Since the probability that an exponentially distributed service time is completed within a time D equals $1 - \exp(-\mu D)$, we have

$$p_{kj} = \binom{k+1}{j} e^{-\mu D j} (1 - e^{-\mu D})^{k+1-j}, \quad 0 \le k \le c-1 \text{ and } 0 \le j \le k+1.$$

The probabilities p_{kj} for $k > c-1$ require a little bit more explanation. We first note that the times between service completions are independent exponentials with common mean $1/(c\mu)$ as long as c or more customers are present. Thus, starting with $k+1 \ge c$ customers present, the time until the $(k+1-c)$th service completion has an E_{k+1-c} distribution. By conditioning on the epoch of this $(k+1-c)$th service completion, we find for any $k \ge c$ that

$$\begin{aligned} p_{kj} &= \int_0^D \binom{c}{j} e^{-\mu(D-x)j} \{1 - e^{-\mu(D-x)}\}^{c-j} (c\mu)^{k+1-c} \frac{x^{k-c}}{(k-c)!} e^{-c\mu x}\, dx \\ &= \binom{c}{j} e^{-j\mu D} c\mu \int_0^D \frac{(c\mu x)^{k-c}}{(k-c)!} (e^{-\mu x} - e^{-\mu D})^{c-j}\, dx, \quad 0 \le j \le c. \end{aligned}$$

This expression is needed to evaluate $p^*_{c-1,\,j}$. We find

$$p^*_{c-1,\,j} = p_{c-1,\,j} + c\mu\eta \binom{c}{j} e^{-j\mu D} \int_0^D e^{c\mu\eta x} (e^{-\mu x} - e^{-\mu D})^{c-j}\, dx$$

for $1 \leq j \leq c-1$. Numerical integration must be used to calculate $p^*_{c-1,j}$ for $1 \leq j \leq c-1$. A convenient method is Gauss–Legendre integration. The other coefficients p_{kj} of the linear equations (9.7.13) are simply computed as binomial coefficients. Once the linear equations (9.7.13) have been solved, we can compute the various performance measures.

The analysis for the $D/M/c$ queue can straightforwardly be generalized to the $GI/M/c$ queue. However, in general, the expression for p_{kj} with $k \geq c$ is quite complicated and leads to a cumbersome and time-consuming calculation of $p^*_{c-1,j}$. Fortunately, a much simpler alternative is available when the interarrival time has a phase-type distribution.

Coxian-2 arrivals

Suppose that the interarrival time has a Coxian-2 distribution with parameters $(b, \lambda_1, \lambda_2)$. In other words, the interarrival time first goes through phase 1 and next it is finished with probability $1-b$ or goes through a second phase 2 with probability b, where the phases are independent exponentials with respective means $1/\lambda_1$ and $1/\lambda_2$.

The state probabilities p_j for $0 \leq j \leq c$ can be calculated by using the continuous-time Markov chain approach. Define $X(t)$ as the number of customers present at time t and let $Y(t)$ be the phase of the interarrival time in progress at time t. The process $\{(X(t), Y(t))\}$ is a continuous-time Markov chain with state space $I = \{(n, i) \mid n = 0, 1, \ldots; i = 1, 2\}$. Denoting the equilibrium probabilities of this Markov chain by p_{ni}, we have $p_n = p_{n1} + p_{n2}$. By equating the rate at which the system leaves the set of states having at least n customers present to the rate at which the system enters this set, we obtain

$$\min(n, c)\mu(p_{n1} + p_{n2}) = \lambda_1(1-b)p_{n-1,1} + \lambda_2 p_{n-1,2}, \quad n \geq 1. \tag{9.7.14}$$

This system of equations is augmented by the equations

$$[\min(n, c)\mu + \lambda_2]p_{n2} = \min(n+1, c)\mu p_{n+1,2} + \lambda_1 b p_{n1}, \quad n \geq 0. \tag{9.7.15}$$

These equations follow by equating the rate out of state $(n, 2)$ to the rate into this state. A closer examination of equations (9.7.14) and (9.7.15) reveals that they cannot be solved recursively starting with $\overline{p}_0 := 1$. Nevertheless, a recursive computation of $p_0, \ldots, p_c$ is possible since

$$\frac{p_{n+1,i}}{p_{ni}} = \eta, \quad n \geq c \text{ and } i = 1, 2. \tag{9.7.16}$$

The relation (9.7.16) extends the relation $p_{n+1}/p_n = \eta$ for $n \geq c$. A proof of the relation (9.7.16) is not given here. It can be deduced from Lemma 3.5.10 and general results in Takahashi (1981). The constant η can be computed beforehand from equation (9.7.12). Using the expression for Coxian-2 density given in Appendix B,

this equation becomes

$$\frac{r_1\lambda_1}{c\mu(1-\eta)+\lambda_1}+\frac{r_2\lambda_2}{c\mu(1-\eta)+\lambda_2}=\eta, \tag{9.7.17}$$

where $r_1 = 1 - b\lambda_1/(\lambda_1 - \lambda_2)$ and $r_2 = 1 - r_1$. Here it is assumed that $\lambda_1 \neq \lambda_2$. Once η is known, we can express p_{c2} into p_{c1}. Substituting $p_{c+1,2} = \eta p_{c2}$ into (9.7.15) with $n = c$ yields

$$(c\mu+\lambda_2)p_{c2} = c\mu\eta p_{c2} + \lambda_1 b p_{c1}.$$

The following algorithm can now be given.

Algorithm

Step 0. Calculate first η as the unique root of equation (9.7.17) on (0,1). Let $\overline{p}_{c1} := 1$ and $\overline{p}_{c2} := \lambda_1 b\{c\mu(1-\eta)+\lambda_2\}^{-1}\overline{p}_{c1}$.
Step 1. For $k = c-1, \ldots, 0$, use equation (9.7.14) with $n = k+1$ and equation (9.7.15) with $n = k$ to solve for $\overline{p}_{k1}$ and $\overline{p}_{k2}$.
Step 2. Calculate $\overline{p}_n := \overline{p}_{n1} + \overline{p}_{n2}$ for $n = 0, 1, \ldots, c$ and next use relation (9.7.10) to normalize the $\overline{p}_n$ as

$$p_n := \left[\sum_{j=0}^{c-1}\overline{p}_j + \frac{\overline{p}_c}{1-\eta}\right]^{-1}\overline{p}_n, \quad n = 0, 1, \ldots, c.$$

Generalized Erlangian arrivals

Suppose that the interarrival time has density

$$a(t) = \sum_{i=1}^{m} q_i \alpha^i \frac{t^{i-1}}{(i-1)!}e^{-\alpha t}, \quad t \geq 0,$$

where $q_m > 0$. In other words, with probability q_i an interarrival time is the sum of i independent phases each having an exponential distribution with mean $1/\alpha$. We again use the continuous-time Markov chain approach to compute the probabilities p_j. Define $X(t)$ as the number of customers present at time t and let $Y(t)$ be the number of remaining phases of the interarrival time in progress at time t. The process $\{(X(t), Y(t))\}$ is a continuous-time Markov chain with state space $I = \{(n, i) \mid n \geq 0;\ 1 \leq i \leq m\}$. By equating the rate at which the system leaves the set of states having at least n customers present to the rate at which the system enters this set, we find

$$\min(n, c)\mu p_n = \alpha p_{n-1,1}, \quad n \geq 1. \tag{9.7.18}$$

Moreover, by rate out of state $(n, i) =$ rate into state (n, i),

$$[\min(n, c)\mu + \alpha]p_{ni} = \alpha p_{n,i+1} + \min(n + 1, c)\mu p_{n+1,i} + \alpha q_i p_{n-1,1}$$

for $n \geq 0$ and $1 \leq i \leq m$, where $p_{n,m+1} = p_{-1,1} = 0$ by convention. Again a rather simple solution procedure can be given in view of

$$\frac{p_{n+1,i}}{p_{n,i}} = \eta, \quad n \geq c \text{ and } 1 \leq i \leq m.$$

A proof of this result will not be given here. The decay factor η is the unique solution to the equation

$$\eta = \sum_{i=1}^{m} q_i \frac{\alpha^i}{[c\mu(1 - \eta) + \alpha]^i}$$

on the interval (0, 1). By substitution of (9.7.18) into the balance equation for p_{ni}, we obtain for each $n \geq 0$ that

$$[\min(n, c)\mu + \alpha]p_{ni} = \alpha p_{n,i+1} + \min(n + 1, c)\mu p_{n+1,i} + q_i \min(n, c)\mu \sum_{j=1}^{m} p_{nj}, \quad 1 \leq i \leq m. \tag{9.7.19}$$

In particular, since $p_{c+1,i} = \eta p_{ci}$ for $1 \leq i \leq m$,

$$(c\mu + \alpha)p_{ci} = \alpha p_{c,i+1} + c\mu\eta p_{ci} + q_i c\mu \sum_{j=1}^{m} p_{cj}, \quad 1 \leq i \leq m. \tag{9.7.20}$$

The probabilities $p_0, \ldots, p_c$ can now be computed as follows.

Algorithm

Step 0. Calculate the decay factor η. Let $\overline{p}_{c1} := 1$.
Step 1. Solve the linear equations (9.7.20) with $2 \leq i \leq m$ to obtain $\overline{p}_{ci}$ for $2 \leq i \leq m$.
Step 2. For $k = c - 1, \ldots, 0$, solve the linear equations (9.7.19) with $n = k$ to obtain $\overline{p}_{ki}$ for $1 \leq i \leq m$.
Step 3. Calculate $\overline{p}_n := \sum_{j=1}^{m} \overline{p}_{nj}$ for $n = 0, 1, \ldots, c$ and normalize the $\overline{p}_n$ as

$$p_n := \left[\sum_{j=0}^{c-1} \overline{p}_j + \frac{\overline{p}_c}{1 - \eta}\right]^{-1} \overline{p}_n, \quad n = 0, 1, \ldots, c.$$

The algorithm requires that a system of linear equations of order m is solved c times. This is computationally feasible provided m is not too large.

9.7.2 The $GI/D/c$ Queue

In the $GI/D/c$ queue the arrival process of customers is a renewal process and the service time of each customer is equal to the constant D. Let us consider the situation that the interarrival-time distribution has a probability density $a(t)$ with Laplace transform

$$a^*(s) = \int_0^\infty e^{-st} a(t)\, dt.$$

We first discuss the computation of the state probabilities

$$p_j = \lim_{t\to\infty} p_j(t), \quad j = 0, 1, \ldots,$$

where $p_j(t) = P\{j \text{ customers will be present at time } t\}$. In a similar way as in the $M/D/c$ queue, the probabilities p_j can be computed from a system of linear equations. Let

$$a_n(D) = \lim_{t\to\infty} a_n(t, D), \quad n = 0, 1, \ldots,$$

where $a_n(t, D) = P\{n \text{ customers will arrive in } (t, t + D]\}$, $t > 0$. Mimicking the derivation of (9.6.1), we obtain the equilibrium equations

$$p_j = a_j(D) \sum_{k=0}^{c} p_k + \sum_{k=c+1}^{c+j} p_k a_{j-k+c}(D), \quad j = 0, 1, \ldots. \tag{9.7.21}$$

These linear equations are obtained by letting $t \to \infty$ in

$$p_j(t + D) = \sum_{k=0}^{c} p_k(t) a_j(t, D) + \sum_{k=c+1}^{c+j} p_k(t) a_{j-k+c}(t, D).$$

To solve the linear equations (9.7.21) together with $\sum_{j=0}^{\infty} p_j = 1$, we need first to compute the probabilities $a_n(D)$. These probabilities can be numerically obtained by Laplace inversion. In Section 8.1 it was shown that

$$\int_0^\infty e^{-sx} a_0(x)\, dx = \frac{1}{s} - \frac{\lambda(1 - a^*(s))}{s^2} \tag{9.7.22}$$

and

$$\int_0^\infty e^{-sx} a_n(x)\, dx = \frac{\lambda[(1 - a^*(s)]^2 [a^*(s)]^{n-1}}{s^2}, \quad n \geq 1. \tag{9.7.23}$$

The infinite system of linear equations for the p_j can be reduced to a finite system by using the geometric tail approach discussed in Section 3.4.2. By (9.7.1),

$$\frac{p_j}{p_{j-1}} \sim \tau^{-1} \quad \text{as } j \to \infty,$$

where $\tau = e^{\delta D/c}$ and δ is the unique solution of the equation

$$e^{\delta D/c} \int_0^\infty e^{-\delta x} a(x)\, dx = 1 \tag{9.7.24}$$

on the interval $(0, \infty)$. Hence a finite system of linear equations is obtained for the p_j by replacing p_j by $p_M \tau^{-(j-M)}$ for $j \geq M$ with M a sufficiently large integer.

Waiting-time probabilities

In general it is not possible to give a tractable algorithm for the waiting-time probabilities in the $GI/D/c$ queue. An exception is the $E_k/D/c$ queue. The waiting-time probabilities in the $E_k/D/c$ queue are the same as the waiting-time probabilities in the $M/D/kc$ queue with the same server utilization as in the $E_k/D/c$ queue.

Theorem 9.7.1 *The waiting-time distribution function $W_q(x)$ in the multi-server $GI/D/c$ queue is the same as in the single-server $GI^{(c*)}/D/1$ queue in which the interarrival time is distributed as the sum of c interarrival times in the GI/D/c queue.*

Proof Since the service times are deterministic, it is no restriction to cyclically assign the customers to the c servers. Then server k gets the customers numbered as $k, k+c, k+2c, \ldots$ for $k = 1, \ldots, c$. This simple observation proves the theorem.

The theorem has the following important corollary.

Corollary 9.7.2 *The waiting-time distribution function $W_q(x)$ in the $E_k/D/c$ queue is identical to the waiting-time distribution in the $M/D/kc$ queue with the same server utilization.*

Proof An Erlang (k, α) distributed random variable has the same distribution as the sum of k independent random variables each having an exponential distribution with mean $1/\alpha$. Consider now the $E_k/D/c$ system with mean interarrival time k/α and the $M/D/kc$ system with mean interarival time $1/\alpha$. By Theorem 9.7.1, both the waiting-time distribution in the $E_k/D/c$ system and the waiting-time distribution in the $M/D/kc$ system are the same as the waiting-time distribution in the $E_{ck}/D/1$ queue with mean interarrival time ck/α. This gives the desired result.

What can be done for the case of a general interarrival-time distribution? Then an approximation to the waiting-time probabilities can be computed by using Theorem 9.7.1. The idea is to approximate the $GI^{(c*)}/D/1$ queue by an $Ph/D/1$ queue by replacing the interarrival-time distribution by a tractable phase-type distribution that matches the first two or three moments. Section 9.5.4 discusses algorithms to compute the waiting-time probabilities in the $Ph/D/1$ queue.

9.8 FINITE-CAPACITY QUEUES

This section considers queueing systems having room for only a finite number of customers. Each customer finding no waiting place available upon arrival is rejected. A rejected customer is assumed to have no further influence on the system. In finite-capacity systems the finite waiting room acts as a regulator on the queue size and so no *a priori* assumption on the offered load is needed. A practical problem of considerable interest is the calculation of the rejection probability. A basic problem in telecommunication and production is the design of finite buffers such that the rejection probability is below a prespecified value. In this section it will be shown that the rejection probability for the finite-buffer model can often be expressed in terms of the state probabilities for the corresponding *infinite-buffer* model. This result greatly simplifies the calculation of the smallest buffer size such that the rejection probability is below a prespecified value. Before discussing this result in Section 9.8.2, we first discuss in Section 9.8.1 an approximation to the state probabilities in the $M/G/c/c+N$ queue.

9.8.1 The $M/G/c/c+N$ Queue

The $M/G/c/c$ queueing model has a Poisson input with rate λ, a general service-time distribution, c identical servers and N waiting positions for customers to await service. An arriving customer who finds all c servers busy and all N waiting places occupied is rejected. A tractable exact solution of this model is only possible for the case of a single server ($M/G/1/N$ queue), the case of exponential services ($M/M/c/c+N$ queue) and the case of no waiting room ($M/G/c/c$ queue). The $M/G/c/c$ queue (Erlang loss model) was discussed in detail in Section 5.2 and the $M/M/c/c+N$ queue was dealt with in Exercise 5.1.

In the $M/G/c/c+N$ queue the service time S of a customer has a general probability distribution function $B(x)$ with $B(0)=0$. No restriction is imposed on the load factor ρ defined by $\rho=\lambda E(S)/c$. Let $\{p_j, 0 \le j \le N+c\}$ denote the limiting distribution of the number of customers present. The next theorem extends the approximation that was given in Theorem 9.6.1 for the state probabilities in the infinite-capacity $M/G/c$ queue. An approximation to the waiting-time probabilities (percentiles) in the $M/G/c/c+N$ is outlined in Exercise 9.14. This approximation is based on the approximation to the state probabilities.

Theorem 9.8.1 *Under Assumption 9.6.1, the state probabilities p_j are approximated by*

$$p_j^{app} = \frac{(c\rho)^j}{j!} p_0^{app}, \quad 0 \le j \le c-1,$$

$$p_j^{app} = \lambda p_{c-1}^{app} a_{j-c} + \lambda \sum_{k=c}^{j} p_k^{app} b_{j-k}, \quad c \le j \le N+c-1,$$

Table 9.8.1 Numerical results for P_{rej} in the $M/G/c/c+N$ queue ($c=5$).

		$\rho = 0.5$		$\rho = 0.8$		$\rho = 1.5$	
c_S^2		$N=1$	$N=5$	$N=1$	$N=5$	$N=1$	$N=5$
0	app	0.0286	0.00036	0.1221	0.0179	0.3858	0.3348
	exa	[0.0281–0.0293]	[0.00032–0.00038]	[0.1212–0.1236]	[0.0168–0.0182]	[0.3854–0.3886]	[0.3332–0.3372]
$\frac{1}{2}$	app	0.0311	0.0010	0.1306	0.0308	0.3975	0.3395
	exa	0.0314	0.0010	0.1318	0.0314	0.4000	0.3400
2	app	0.0370	0.0046	0.1450	0.0603	0.4114	0.3555
	exa	0.0366	0.0044	0.1435	0.0587	0.4092	0.3537

$$p_j^{app} = \rho p_{c-1}^{app} - (1-\rho) \sum_{k=c}^{N+c-1} p_k^{app},\ j = N+c,$$

where $\rho = \lambda E(S)/c$ and the constants a_n and b_n are the same as in Theorem 9.6.1.

Proof The proof of the theorem is a minor modification of the proof of Theorem 9.6.1. The details are left to the reader.

The result of Theorem 9.8.1 is exact for both the case of multiple servers with exponential service times and the case of a single server with general service times, since for these two special cases the approximation assumption holds exactly. Further support for the approximate result of the theorem is provided by the fact that the approximation is exact for the case of no waiting room ($N=0$).

Numerical investigations indicate that the approximation for the state probabilities is accurate enough for practical purposes. Table 9.8.1 gives the exact and approximate values of the rejection probability P_{rej} for several examples. The probability P_{rej} denotes the long-run fraction of customers who are rejected. By the PASTA property,

$$P_{rej} = p_{N+c}.$$

In all examples we take $c=5$ servers. Deterministic services ($c_S^2 = 0$), E_2 services ($c_S^2 = \frac{1}{2}$) and H_2 services with gamma normalization ($c_S^2 = 2$) are considered. For the latter two services, the exact values of P_{rej} are taken from the tabulations of Seelen *et al.* (1985). For deterministic services, computer simulation was used to find P_{rej}. In the table we give the 95% confidence intervals. It is interesting to point out that the results in Table 9.8.1 support the long-standing conjecture for the $GI/G/c/c+N$ queue that $P_{rej} \to 1 - 1/\rho$ as $N \to \infty$ when $\rho > 1$.

A proportionality relation

For the case of $\rho < 1$ the computational work can be considerably reduced when the approximation to P_{rej} must be computed for several values of N. Denote by

$p_j^{(\infty)}(\text{app})$ the approximation given in Theorem 9.6.1 to the state probability $p_j^{(\infty)}$ in the infinite-capacity $M/G/c$ queue. This approximation requires that $\rho < 1$. An inspection of the recursion schemes in Theorems 9.6.1 and 9.8.1 reveals that, for some constant γ,

$$p_j^{app} = \gamma p_j^{(\infty)}(\text{app}), \quad j = 0, 1, \ldots, N + c - 1. \tag{9.8.1}$$

The constant γ is given by $\gamma = [1 - \rho \sum_{j=N+c}^{\infty} p_j^{(\infty)}(\text{app})]^{-1}$. In the next section it will be seen that this proportionality relation implies

$$P_{rej}^{app} = \frac{(1-\rho) \sum_{j=N+c}^{\infty} p_j^{(\infty)}(\text{app})}{1 - \rho \sum_{j=N+c}^{\infty} p_j^{(\infty)}(\text{app})}, \tag{9.8.2}$$

where $P_{rej}^{app} = p_{N+c}^{app}$ denotes the approximation to P_{rej}. The computation of the probabilities $p_j^{(\infty)}(\text{app})$ was discussed in Section 9.6.2.

The approximations p_j^{app} and $p_j^{(\infty)}$ (app) are exact both for the case of multiple servers with exponential service times and for the case of a single server with general service times. Therefore relations (9.8.1) and (9.8.2) hold exactly for the $M/M/c/c+N$ queue and the $M/G/1/N+1$ queue. For these particular queueing models the proportionality relation (9.8.1) can be directly explained by a simple probabilistic argument. This will be done in the next subsection. It is noted that for the general $M/G/c/c+N$ queue the proportionality relation is not satisfied when the exact values of p_j and $p_j^{(\infty)}$ are taken instead of the approximate values.

9.8.2 A Basic Relation for the Rejection Probability

In this section a structural form will be revealed for the rejection probability. In many situations the rejection probability can be expressed in terms of the state probabilities in the infinite-capacity model. In the following, p_j and $p_j^{(\infty)}$ denote the time-average state probabilities for the finite-capacity model and the infinite-capacity model. To ensure the existence of the probabilities $p_j^{(\infty)}$, it is assumed that the server utilization ρ is smaller than 1.

Theorem 9.8.2 *Both for the $M/M/c/c+N$ queue and the $M/G/1/N+1$ queue it holds that*

$$p_j = \gamma p_j^{(\infty)}, \quad j = 0, 1, \ldots, N + c - 1 \tag{9.8.3}$$

for some constant $\gamma > 0$. The constant γ is given by $\gamma = [1 - \rho \sum_{j=N+c}^{\infty} p_j^{(\infty)}]^{-1}$ and the rejection probability is given by

$$P_{rej} = \frac{(1-\rho)\sum_{j=N+c}^{\infty} p_j^{(\infty)}}{1-\rho\sum_{j=N+c}^{\infty} p_j^{(\infty)}}. \tag{9.8.4}$$

Proof The proof of (9.8.3) is based on the theory of regenerative processes. The process describing the number of customers present is a regenerative stochastic process in both the finite-capacity model and the infinite-capacity model. For both models, let a cycle be defined as the time elapsed between two consecutive arrivals that find the system empty. For the finite-capacity model, we define the random variables

T = the length of one cycle,

T_j = the amount of time that j customers are present during one cycle.

The corresponding quantities for the infinite-capacity model are denoted by $T^{(\infty)}$ and $T_j^{(\infty)}$. By the theory of regenerative processes,

$$p_j = \frac{E(T_j)}{E(T)} \quad \text{and} \quad p_j^{(\infty)} = \frac{E(T_j^{(\infty)})}{E(T^{(\infty)})}, \quad j = 0, 1, \ldots, N+c. \tag{9.8.5}$$

The crucial observation is that the random variable T_j has the same distribution as $T_j^{(\infty)}$ for any $0 \le j \le N+c-1$ both in the $M/M/c/c+N$ queue and in the $M/G/1/N+1$ queue. This result can be roughly explained as follows. Suppose that at epoch 0 a cycle starts and let the processes $\{L(t)\}$ and $\{L^{(\infty)}(t)\}$ describe the number of customers present in the finite-capacity system and in the infinite-capacity system. During the first cycle the behaviour of the process $\{L(t)\}$ is identical to that of the process $\{L^{(\infty)}(t)\}$ as long as the processes have not reached the level $N+c$. Once the level $N+c$ has been reached, the process $\{L^{(\infty)}(t)\}$ may temporarily make an excursion above the level $N+c$. However, after having reached the level $N+c$, both the process $\{L(t)\}$ and the process $\{L^{(\infty)}(t)\}$ will return to the level $N+c-1$. This return to the level $N+c-1$ occurs at a service completion epoch. At a service completion epoch the elapsed service times of the other services in progress are not relevant. In the $M/G/1/N+1$ queue the reason is simply that no other services are in progress at a service completion epoch and in the $M/M/c/c+N$ queue the explanation lies in the memoryless property of the exponential service-time distribution. Also, it should be noted that at a service completion epoch the elapsed time since the last arrival is not relevant since the arrival process is a Poisson process. Thus we can conclude that after a downcrossing to the level $N+c-1$ the behaviour of the process $\{L^{(\infty)}(t)\}$ is again probabilistically the same as the behaviour of the process $\{L(t)\}$ as long as the number of customers present stays below the level $N+c$. These arguments make it plausible that the distribution of T_j is the same as that of $T_j^{(\infty)}$ for any

$0 \leq j \leq N + c - 1$. Next it follows from (9.8.5) that (9.8.3) holds with

$$\gamma = \frac{E(T^{(\infty)})}{E(T)}.$$

The proportionality relation (9.8.3) is the key to the proof of (9.8.4). We first note that in the finite-capacity model the average number of busy servers equals $\lambda(1 - P_{rej})E(S)$ by Little's formula. Writing $\lambda(1 - P_{rej})E(S)$ as $c\rho(1 - P_{rej})$, it follows that

$$c\rho(1 - P_{rej}) = \sum_{j=0}^{N+c} \min(j, c)p_j = \sum_{j=0}^{c-1} jp_j + c(1 - \sum_{j=0}^{c-1} p_j).$$

Substituting (9.8.3) in this equation gives

$$\begin{aligned} c\rho(1 - P_{rej}) &= \gamma \sum_{j=0}^{c-1} jp_j^{(\infty)} + c(1 - \gamma \sum_{j=0}^{c-1} p_j^{(\infty)}) \\ &= \gamma \sum_{j=0}^{c-1} jp_j^{(\infty)} + c[1 - \gamma(1 - \sum_{j=c}^{\infty} p_j^{(\infty)})] \\ &= \gamma \sum_{j=0}^{\infty} \min(j, c)p_j^{(\infty)} + c - c\gamma. \end{aligned}$$

By Little's formula, the average number of busy servers equals $c\rho$ in the infinite-buffer model and so $\sum_{j=0}^{\infty} \min(j, c)p_j^{(\infty)} = c\rho$. This leads to

$$c\rho(1 - P_{rej}) = \gamma c\rho + c - c\gamma.$$

Solving for γ gives

$$P_{rej} = \frac{(1 - \rho)(\gamma - 1)}{\rho}. \tag{9.8.6}$$

Also, using the PASTA property,

$$\begin{aligned} P_{rej} = p_{N+c} &= 1 - \sum_{j=0}^{N+c-1} p_j = 1 - \gamma \sum_{j=0}^{N+c-1} p_j^{(\infty)} \\ &= 1 - \gamma[1 - \sum_{j=N+c}^{\infty} p_j^{(\infty)}]. \end{aligned} \tag{9.8.7}$$

By (9.8.6) and (9.8.7), $\gamma = [1-\rho \sum_{j=N+c}^{\infty} p_j^{(\infty)}]^{-1}$. Next the result (9.8.4) follows.

It is important to point out that the assumption of a single server with general service times or multiple servers with exponential service times was only used for the proof of (9.8.3). The proof of (9.8.4) does not use this assumption, but is solely based on the proportionality relation (9.8.3).

Table 9.8.2 Numerical results for the $D/M/c/c+N$ queue

		$\rho = 0.8$			$\rho = 0.95$		
		$N=0$	$N=10$	$N=25$	$N=0$	$N=50$	$N=75$
$c=5$	app	1.02E-2	6.99E-4	6.59E-7	1.48E-1	1.39E-6	$6.83E$-9
	exa	1.11E-2	7.49E-4	7.06E-7	1.59E-1	1.44E-6	$6.74E$-9
$c=25$	app	1.59E-3	1.44E-4	1.37E-7	4.98E-2	7.54E-7	$3.53E$-9
	exa	1.71E-3	1.55E-4	1.46E-7	5.23E-2	7.80E-7	$3.65E$-9
$c=100$	app	2.16E-4	2.08E-6	1.97E-9	9.60E-3	1.94E-7	$9.07E$-10
	exa	2.32E-4	2.23E-6	2.11E-9	9.96E-3	2.00E-7	$9.39E$-10

Interpretation of formula (9.8.4)

Define for the infinite-capacity $M/G/c$ queue the tail probability

$$\Pi_{N+c}^{(\infty)} = \text{the long-run fraction of customers who find } N+c \text{ or more other customers present upon arrival.}$$

By the PASTA property $\Pi_{N+c}^{(\infty)} = \sum_{j=N+c}^{\infty} p_j^{(\infty)}$, and so formula (9.8.4) can be written in the more insightful form

$$P_{rej} = \frac{(1-\rho)\Pi_{N+c}^{(\infty)}}{1-\rho\Pi_{N+c}^{(\infty)}}. \tag{9.8.8}$$

Practitioners often use the tail probability $\Pi_{N+c}^{(\infty)}$ from the infinite-capacity model as an approximation to the rejection probability in the finite-capacity model. The formula (9.8.8) shows that this is a poor approximation when ρ is not very small. The approximation $\Pi_{N+c}^{(\infty)}$ differs by a factor $(1-\rho)^{-1}$ from the right-hand side of (9.8.8) when N gets large. The improved approximation (9.8.8) is just as easy to use as the approximation $\Pi_{N+c}^{(\infty)}$. In queueing systems in which the proportionality relation (9.8.3) does not necessarily holds, the structural form $(1-\rho)\Pi_{N+c}^{(\infty)}/(1-\rho\Pi_{N+c}^{(\infty)})$ can nevertheless be used as an approximation to P_{rej}. In Exercise 9.14 this will be illustrated for the single-server queue with a Markov modulated arrival process. Here we illustrate the performance of the approximation $(1-\rho)\Pi_{N+c}^{(\infty)}/(1-\rho\Pi_{N+c}^{(\infty)})$ to the rejection probability in the $D/M/c/c+N$ queue with deterministic arrivals. Table 9.8.2 gives the approximate and exact values of P_{rej} for several examples. The numerical result shows an excellent performance of the approximation. In all examples the approximate value of P_{rej} is of the same order of magnitude as the exact value. This is what is typically needed when a heuristic is used for dimensioning purposes.

9.8.3 The $M^X/G/c/c+N$ Queue with Batch Arrivals

Theorem 9.8.2 can be extended to the batch-arrival $M^X/G/c/c+N$ queue. In this model batches of customers arrive according to a Poisson process with rate λ and

the batch size X has a discrete probability distribution $\{\beta_j, j \geq 1\}$ with mean β. Denoting by μ the mean service time of a customer, it is assumed that the load factor $\rho = \lambda\beta\mu/c$ is smaller than 1. As before $\{p_j, 0 \leq j \leq N + c\}$ denotes the limiting distribution of the number of customers present. For finite-buffer queues with batch arrivals we must distinguish between these two cases:

(a) *Partial rejection*: an arriving batch whose size exceeds the remaining capacity of the buffer is partially rejected by turning away only those customers in excess of the remaining capacity.

(b) *Complete rejection*: an arriving batch whose size exceeds the remaining capacity of the buffer is rejected in its entirety.

The emphasis of the discussion will be on the case of partial rejection. We first derive an expression for the tail probability $\Pi_{N+c}^{(\infty)}$ in the infinite-capacity $M^X/G/c$ queue. Let $\{p_j^{(\infty)}\}$ denote the time-average probabilities in the infinite-capacity $M^X/G/c$ queue. Then, by the PASTA property,

the long-run fraction of batches finding k other customers present upon arrival

$$= p_k^{(\infty)}, \quad k = 0, 1, \ldots . \tag{9.8.9}$$

Suppose that the customers are numbered as $1, 2, \ldots$ in accordance with the order in which the batches arrive and in accordance with the relative positions the customers take within the same batch. Define for $j = 0, 1, \ldots$,

$\pi_j^{(\infty)} =$ the long-run fraction of customers who have j other customers in front of them just after arrival (including customers from the same batch).

In Section 9.3.2 we have already shown that

the long-run fraction of customers taking the rth position in their batch

$$= \frac{1}{\beta} \sum_{j=r}^{\infty} \beta_j, \quad r = 1, 2, \ldots .$$

This result in conjunction with (9.8.9) gives

$$\pi_j^{(\infty)} = \frac{1}{\beta} \sum_{k=0}^{j} p_k^{(\infty)} \sum_{s=j-k+1}^{\infty} \beta_s, \; j = 0, 1, \ldots . \tag{9.8.10}$$

Hence, in the infinite-capacity model, the long-run fraction of customers having $N + c$ or more customers in front of them just after arrival is given by

$$\Pi_{N+c}^{(\infty)} = \sum_{j=N+c}^{\infty} \frac{1}{\beta} \sum_{k=0}^{j} p_k^{(\infty)} \sum_{s=j-k+1}^{\infty} \beta_s. \tag{9.8.11}$$

As before, let P_{rej} denote the long-run fraction of customers who are rejected in the finite-capacity model. For the $M^X/G/c/c+N$ queue with *partial rejection*, we approximate P_{rej} by

$$P_{rej} \approx \frac{(1-\rho)\Pi_{N+c}^{(\infty)}}{1-\rho\Pi_{N+c}^{(\infty)}}. \tag{9.8.12}$$

The approximation (9.8.12) to P_{rej} holds *exactly* for the $M^X/G/1/N$ queue with partial rejection and the $M^X/M/c/c+N$ queue with partial rejection. It is left to the reader to verify that the proportionality relation (9.8.3) remains valid for these special cases. In the proof of Theorem 9.8.2 one needs only to modify formula (9.8.7). In the $M^X/G/c/c+N$ model with partial rejection,

$$P_{rej} = \frac{1}{\beta}\sum_{k=0}^{N+c} p_k \sum_{s=N+c-k+1}^{\infty} (k+s-N-c)\beta_s.$$

This result follows by noting that the fraction of customers rejected is the ratio of the average number of customers rejected per batch and the average batch size.

Complete rejection

In the $M^X/G/c/c+N$ queue with *complete* rejection it is no longer true that the proportionality relation (9.8.3) holds for the case of a single server with general service times and for the case of multiple servers with exponential service times. However, one might make the heuristic assumption that $p_j \approx \gamma p_j^{(\infty)}$ for $0 \le j \le N+c-1$. Exercise 9.19 is to verify that this heuristic assumption leads to the approximation

$$P_{rej} \approx \frac{(1-\rho)\left[1-\sum_{j=0}^{N+c-1} u_j^{(\infty)}\right]}{1-\rho\left[1-\sum_{j=0}^{N+c-1} u_j^{(\infty)}\right]}, \tag{9.8.13}$$

where

$$u_j^{(\infty)} = \frac{1}{\beta}\sum_{k=0}^{j} p_k^{(\infty)} \sum_{s=j-k+1}^{N+c-k} \beta_s.$$

A remarkable result is that for the case of a constant batch size Q with $Q \le N+1$ the approximation (9.8.13) is exact for both the $M^X/G/1/N+1$ queue with complete rejection and the $M^X/M/c/c+N$ queue with complete rejection; see Exercises 9.20 and 9.21. In these cases with a constant batch size Q it holds that $p_j \approx \gamma p_j^{(\infty)}$ for any $0 \le j \le N+c-Q$.

Table 9.8.3 The $M^X/G/1/N+1$ queue with complete rejection

		Geometric			Two-point		
c_S^2		$N=0$	$N=50$	$N=250$	$N=0$	$N=50$	$N=250$
0.1	app	8.99E-1	1.40E-2	1.62E-7	8.86E-1	8.88E-3	1.29E-8
	exa	9.40E-1	1.59E-2	1.82E-7	8.86E-1	9.01E-3	1.31E-8
10	app	8.99E-1	6.09E-2	2.64E-4	8.86E-1	5.58E-2	1.79E-4
	exa	9.40E-1	6.21E-2	2.68E-4	8.86E-1	5.55E-2	1.79E-4

Table 9.8.3 gives some numerical results for P_{rej} in the $M^X/G/1/N+1$ queue with complete rejection. For the batch size we consider both the two-point distribution $P\{X=1\}=P\{X=7\}=0.5$ and the geometric distribution $P\{X=j\}=(1/4)(3/4)^{j-1}$ for $j\geq 1$. In both cases the mean batch size $\beta=4$. The service-time distributions are the E_{10} distribution ($c_S^2=0.1$) and the H_2 distribution with the gamma normalization ($c_S^2=10$). The offered load ρ is taken equal to 0.8. The results in Table 9.8.3 indicate that the approximation (9.8.13) performs quite well for practical purposes.

Asymptotic expansion for P_{rej}

For larger values of the buffer capacity N, the calculation of P_{rej} can further be simplified when an asymptotic expansion for the tail probabilities in the infinite-buffer model is known. If $P_{rej}=(1-\rho)\sum_{j=N+c}^{\infty}\pi_j^{(\infty)}/[1-\rho\sum_{j=N+c}^{\infty}\pi_j^{(\infty)}]$ and an asymptotic expansion $\pi_j^{(\infty)}\sim\sigma\eta^j$ as $j\to\infty$ is known, then

$$P_{rej}\approx\frac{(1-\rho)\sigma\eta^{N+c}/(1-\eta)}{1-\rho\sigma\eta^{N+c}/(1-\eta)}\approx(1-\rho)\sigma\eta^{N+c}/(1-\eta)\quad\text{for large } N.$$

To illustrate this, consider the single-server $M^X/G/1/N+1$ queue with partial rejection. For the $M^X/G/1$ queue the asymptotic expansion $\pi_j^{(\infty)}\sim\sigma\eta^j$ as $j\to\infty$ holds when the service time is not heavy-tailed, where the constants σ and $\eta=1/\tau$ are determined by the relations (9.3.5) and (9.3.6). When using the asymptotic expansion one needs only to calculate the root of a non-linear equation in a single variable.

Two-moment approximation

The practical applicability of the formulas for P_{rej} stands or falls with the computation of the state probabilities $\pi_j^{(\infty)}$. In some queueing models it is computationally feasible to calculate these probabilities using embedded Markov chain analysis or continuous-time Markov chain analysis. However, in many queueing models the exact computation of the state probabilities $\pi_j^{(\infty)}$ is not practically feasible. This

is for instance the case in the $M^X/G/c$ queue with general service times. In such situations one might try to approximate the exact solution of the complex model through the exact solutions of simpler related models. In this chapter we have already seen several examples of such two-moment approximations. The rejection probability itself is not directly amenable to a two-moment approximation, but indirectly a two-moment approximation is possible through the 'percentile' $N(\alpha)$ defined by

$$N(\alpha) = \text{the minimal buffer size for which the rejection probability } P_{rej} \text{ does not exceed the value } \alpha.$$

This will be illustrated for the $M^X/G/c/c+N$ queue. Denoting by c_S^2 the squared coefficient of the service time of a customer, the two-moment approximation to $N(\alpha)$ is given by

$$N_{app}(\alpha) = (1 - c_S^2)N_{det}(\alpha) + c_S^2 N_{exp}(\alpha), \tag{9.8.14}$$

where $N_{det}(\alpha)$ and $N_{exp}(\alpha)$ are the (approximate) values of the minimal buffer size $N(\alpha)$ for the $M^X/D/c/c+N$ queue and the $M^X/M/c/c+N$ queue. The buffer sizes $N_{det}(\alpha)$ and $N_{exp}(\alpha)$ are computed by using the (approximate) formula for P_{rej} in the particular cases of deterministic services and exponential services. Relatively simple algorithms are available to compute the state probabilities $\pi_j^{(\infty)}$ in the $M^X/M/c$ queue and the $M^X/D/c$ queue; see Section 9.6.3. The two-moment approximation (9.8.14) is only recommended when c_S^2 is not too large (say, $0 \leq c_S^2 \leq 2$).

Table 9.8.4 illustrates the performance of the two-moment approximation (9.8.14) for the $M/G/c/c+N$ queue, where the number of servers has the two values $c = 1$ and $c = 10$. For both Erlang-2 services ($c_S^2 = 0.5$) and H_2 services with gamma normalization ($c_S^2 = 2$), the approximate and exact values of $N(\alpha)$ are given for several values of α. Any fractional value resulting from the interpolation formula (9.8.14) has been rounded up. The results in the table show an excellent performance of the two-moment approximation and also nicely demonstrate that $N(\alpha)$ increases logarithmically in α as α increases.

9.8.4 Discrete-Time Queueing Systems

Many practical queueing systems operate on a discrete-time basis. A discrete-time queueing system is characterized by time-slotted service. A new service can only start at the beginning of a time slot, and the service time is a multiple of time slots. In applications the discrete-time queueing systems typically have finite buffers to store incoming packets. Packets are the entities to be served. Let us assume that there are c service channels and a buffer of capacity N to store incoming packets. The buffer excludes any packet in service. Each service channel can handle only one packet at a time. A new service can only start at the beginning of a time slot. The service times of the packets are independent of each other. It is

Table 9.8.4 The minimal buffer size in the $M/G/c/c+N$ queue

		$\rho = 0.5$					$\rho = 0.8$				
	α	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}
c=1											
$c_S^2 = \frac{1}{2}$	exa	4	9	15	20	25	10	26	41	57	73
	app	4	10	15	20	25	10	26	41	57	73
$c_S^2 = 2$	exa	7	16	26	35	45	19	49	80	110	141
	app	7	17	25	36	46	19	50	80	111	140
c=10											
$c_S^2 = \frac{1}{2}$	exa	1	7	12	17	23	8	24	39	55	71
	app	1	7	13	17	23	8	24	39	55	71
$c_S^2 = 2$	exa	1	10	20	29	39	14	44	74	105	135
	app	1	10	20	29	39	14	45	75	106	135

assumed that the number of time slots needed to serve a packet has a *geometric distribution* $\{(1-r)^{n-1}r, n \geq 1\}$. The case of deterministic services is included as a special case ($r = 1$). In many telecommunication applications the service time of a packet is deterministic and equals one time slot. A served packet leaves the system at the end of the time slot in which the service is completed. The numbers of packets arriving in the system during consecutive time slots are independent non-negative random variables with the common probability distribution $\{a_n, n \geq 0\}$. It is assumed that the packets arrive *individually* during the time slots and that an arriving packet is rejected when it finds the buffer full upon arrival. It is no restriction to use the convention of individual arrivals provided that the partial rejection strategy applies when arrivals actually occur in batches. The load factor ρ is defined as

$$\rho = \frac{\lambda\mu}{c},$$

where $\lambda = \sum_{n=1}^{\infty} na_n$ is the arrival rate of new packets and $\mu = 1/r$ is the mean service time of a packet. Let

$$P_{rej} = \text{the long-run fraction of packets that are rejected.}$$

Under the assumption of $\rho < 1$ an approximation to P_{rej} can be given in terms of the state probabilities in the corresponding infinite-buffer model. Assuming that $\rho < 1$, define for the infinite-buffer model the probability $u_j^{(\infty)}$ by

$$u_j^{(\infty)} = \text{the long-run fraction of time slots at whose beginnings there are } j \text{ packets in the system}$$

for $j = 0, 1, \ldots$. By the assumption of geometrically distributed service times, the process describing the number of packets present at the beginning of a time slot is

a discrete-time Markov chain. This Markov chain was analysed in Example 3.4.1 for the particular case of deterministic services. Letting

$$U(z) = \sum_{j=0}^{\infty} u_j^{(\infty)} z^j, \quad |z| \leq 1,$$

a minor modification of the Markov-chain analysis in Example 3.4.1 yields

$$U(z) = \frac{A(z) \sum_{k=0}^{c-1} u_k^{(\infty)} [z^c (r + (1-r)z)^k - z^k (r + (1-r)z)^c]}{z^c - (r + (1-r)z)^c A(z)}, \tag{9.8.15}$$

where $A(z) = \sum_{n=0}^{\infty} a_n z^n$. This expression is well suited for numerical purposes. First the c unknowns $u_0^{(\infty)}, \ldots, u_{c-1}^{(\infty)}$ are determined by computing the complex roots of the denominator of (9.8.15); see Appendix G. Next the discrete FFT method can be applied to obtain the numerical values of the state probabilities $u_j^{(\infty)}$. In order to obtain the approximation to P_{rej} in the finite-buffer model, we need the tail probability $\sum_{j=N+c}^{\infty} \pi_j^{(\infty)}$ for the infinite-buffer model. In the infinite-buffer model the probability $\pi_j^{(\infty)}$ is defined as

$$\pi_j^{(\infty)} = \text{the long-run fraction of packets who find } j \text{ other packets present upon arrival.}$$

By the same arguments as used to obtain (9.8.10), we find

$$\pi_j^{(\infty)} = \frac{1}{\lambda} \sum_{k=0}^{j} u_k^{(\infty)} \sum_{s=j-k+1}^{\infty} a_s, \quad j = 0, 1, \ldots . \tag{9.8.16}$$

The proposed approximation to P_{rej} in the finite-buffer model is

$$P_{rej} \approx \frac{(1-\rho) \sum_{j=N+c}^{\infty} \pi_j^{(\infty)}}{1 - \rho \sum_{j=N+c}^{\infty} \pi_j^{(\infty)}}. \tag{9.8.17}$$

It has been shown in Gouweleeuw and Tijms (1998) that for the single-server case this approximation is asymptotically exact for large N (more precisely, the approximation (9.8.17) is exact for the single-server case when the probability of more than N arrivals during one time slot equals zero). In general it turns out that (9.8.17) provides an excellent approximation to the rejection probability. To illustrate this, Table 9.8.5 gives some numerical results for the case of deterministic service times. The number of servers is $c = 1$ and $c = 2$, while the Poisson

Table 9.8.5 Numerical results for the discrete-time queue

		$c=1$				$c=2$	
N		Poisson	Geometric	N		Poisson	Geometric
1	exa	3.406×10^{-1}	4.737×10^{-1}	2	exa	2.379×10^{-1}	4.133×10^{-1}
	app	3.024×10^{-1}	4.119×10^{-1}		app	1.879×10^{-1}	3.481×10^{-1}
5	exa	5.505×10^{-2}	1.260×10^{-1}	5	exa	6.595×10^{-2}	1.970×10^{-1}
	app	5.504×10^{-2}	1.254×10^{-1}		app	6.044×10^{-2}	1.859×10^{-1}
10	exa	1.481×10^{-2}	5.081×10^{-2}	10	exa	1.693×10^{-2}	9.054×10^{-2}
	app	1.481×10^{-2}	5.081×10^{-2}		app	1.592×10^{-2}	8.849×10^{-2}
50	exa	3.294×10^{-6}	5.178×10^{-4}	50	exa	3.702×10^{-6}	3.036×10^{-3}
	app	3.294×10^{-6}	5.178×10^{-4}		app	3.511×10^{-6}	3.001×10^{-3}
100	exa	1.046×10^{-10}	2.656×10^{-6}	100	exa	6.626×10^{-13}	1.476×10^{-5}
	app	1.046×10^{-10}	2.656×10^{-6}		app	6.283×10^{-13}	1.460×10^{-5}

distribution and the geometric distribution are considered for the distribution $\{a_n\}$ of the number of arrivals during one time slot. In all examples we take the load factor $\rho = 0.9$.

To conclude this section, it is noted that the approximation to P_{rej} can be extended to discrete-time queueing systems with correlated input. In many applications the input is not renewal but correlated. The switched-batch Bernoulli process is often used for modelling correlated input processes. In this model there is an underlying phase process that is alternately in the states 1 and 2, where the sojourn times in the successive states are independent random variables that have a discrete geometric distribution. The mean of the geometric sojourn time and the distribution of the number of arrivals in a time slot depend on the state of the phase process. Exercise 9.16 is to work out the approximation to P_{rej} in this useful model with correlated input.

EXERCISES

9.1 Consider the $M/G/1$ queue with exceptional first service. This model differs from the standard $M/G/1$ queue only in the service times of the customers reactivating the server after an idle period. Those customers have special service times with distribution function $B_0(t)$, while the other customers have ordinary service times with distribution function $B(t)$. Use the regenerative approach to verify that the state probabilities can be computed from the recursion scheme (9.2.1) in which $\lambda p_0 a_{j-1}$ is replaced by $\lambda p_0 \overline{a}_{j-1}$, where $\overline{a}_n$ is obtained by replacing $B(t)$ by $B_0(t)$ in the integral representation for a_n. Also, argue that p_0 satisfies $1 - p_0 = \lambda[p_0\mu_0 + (1-p_0)\mu_1]$, where μ_1 and μ_0 denote the means of the ordinary service times and the special service times.

9.2 Consider the $M/G/1$ queue with server vacations. In this variant of the $M/G/1$ queue a server vacation begins when the server becomes idle. During a server vacation the server performs other work and is not available for providing service. The length V of a server vacation has a general probability distribution function $V(x)$ with density $v(x)$. If upon return from a vacation the server finds the system empty, a new vacation period begins, otherwise

the server starts servicing. Denote by $p_{0j}(p_{1j})$ the time-average probability that j customers are present and the server is on vacation (available for service). Use the regenerative approach to verify the recursion scheme:

$$p_{0j} = \frac{1-\rho}{E(V)} \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} \{1 - V(t)\}\, dt, \quad j \geq 0,$$

$$p_{1j} = \frac{1-\rho}{E(V)} \sum_{k=1}^{j} \nu_k a_{j-k} + \lambda \sum_{k=1}^{j} (p_{0k} + p_{1k}) a_{j-k}, \quad j \geq 1,$$

where a_n is given in Theorem 9.2.1 and ν_k is the probability of k arrivals during a single vacation period. (*Hint*: take as cycle the time elapsed between two consecutive epochs at which either the server becomes idle or finds an empty system upon return from vacation.)

9.3 Consider an $M/G/1$ queueing system in which the service time of a customer depends on the queue size at the moment the customer enters service. The service time has a probability distribution function $B_1(x)$ when R or fewer customers are present at the moment the customer enters service; otherwise, the service time has probability distribution function $B_2(x)$. Denote by $p_{1j}(p_{2j})$ the time-average probability that j customers are in the system and service according to $B_1(B_2)$ is provided. Use the regenerative approach to verify the recursion scheme

$$p_{1j} = \lambda p_0 a_{j-1}^{(1)} + \lambda \sum_{k=1}^{\min(j,R)} p_{1k} a_{j-k}^{(1)}, \quad j = 1, 2, \ldots .$$

$$p_{2j} = \lambda \sum_{k=R+1}^{j} (p_{1k} + p_{2k}) a_{j-k}^{(2)}, \quad j > R,$$

where $a_n^{(i)}$ is the same as the constant a_n in Theorem 9.2.1 except that $B(t)$ is replaced by $B_i(t), i = 1, 2$. Also, argue that $1 - p_0 = \lambda\{\mu_1 \sum_{j=0}^{R} p_{1j} + \mu_2(1 - \sum_{j=0}^{R} p_{1j})\}$, where μ_i is the mean of the distribution function B_i. (*Hint*: note that the long-run fraction of service completions at which j customers are left behind equals the long-run fraction of customers finding j other customers present upon arrival.)

9.4 Consider the $M/G/1$ retrial queue from Exercise 2.33 again. Let $p_{0j}(p_{1j})$ denote the long-run fraction of time that the server is idle (busy) and j customers are in orbit for $j = 0, 1, \ldots$.

(a) Use the regenerative aproach to establish the recursions

$$j\nu p_{0j} = \lambda p_{1,j-1}, \quad j = 1, 2, \ldots ,$$

$$p_{1j} = \frac{\lambda a_j}{1 - \lambda a_0} p_{00} + \frac{1}{1 - \lambda a_0} \sum_{k=1}^{j} \left(\lambda a_{j-k+1} + \frac{\lambda^2}{k\nu} a_{j-k} \right) p_{1,k-1}, \quad j = 1, 2, \ldots ,$$

where $a_k = \int_0^\infty e^{-\lambda t} (\lambda t)^k (1/k!)\{1 - B(t)\}\, dt$ with $B(t)$ denoting the probability distribution function of the service time of a customer. (*Hint*: let $T_{0j}(T_{1j})$ denote the amount of time during one cycle that the server is idle (busy) and j customers are in orbit and let N_{0j} denote the number of service completions in one cycle at which j customers are left behind in orbit. Argue that $\lambda E(T_{1,j-1}) = E(N_{0j})$ for $j \geq 0$, $\lambda E(T_{1,j-1}) = j\nu E(T_{0j})$ for $j \geq 1$ and $E(T_{1j}) = \sum_{k=0}^{j+1} E(N_{0k}) \overline{A}_{kj}$ for $j \geq 0$, where $\overline{A}_{kj}$ is defined as the expected amount

of time that j customers are in orbit during a given service time when k customers were left behind in orbit at the completion of the previous service time.)

(b) Use generating functions to verify that

$$p_{00} = (1-\rho)\exp\left(-\frac{\lambda^2}{\nu}\int_0^1 \frac{\alpha(z)}{1-\lambda\alpha(z)}\,dz\right),$$

where $\alpha(z) = \int_0^\infty e^{-\lambda t(1-z)}\{1-B(t)\}\,dt$.

(c) Instead of the $M/G/1$ queue with a linear retrial rate, consider the $M/G/1$ queue with a constant retrial rate. That is, retrials occur according to a Poisson process with rate ν when the orbit is not empty. Modify the above results. This problem is based on De Kok (1984).

9.5 Consider the $M/G/1$ queue with exponential first service from Exercise 9.1 again. Assume that service is in order of arrival. Let $W_q(x)$ denote the limiting distribution function of the delay in queue of a customer.

(a) Verify that the generating function $P(z) = \sum_{j=0}^\infty p_j z^j$ is given by

$$P(z) = \frac{p_0[1-\lambda(\alpha(z) - z\alpha_0(z))]}{1-\lambda\alpha(z)},$$

where $\alpha(z) = \int_0^\infty e^{-\lambda(1-z)t}\{1-B(t)\}\,dt$ and $\alpha_0(z) = \int_0^\infty e^{-\lambda(1-z)t}\{1-B_0(t)\}\,dt$.

(b) Verify that the relation (2.5.14) also applies to the $M/G/1$ queue with server vacations, where $E(z^{L_q^{(\infty)}}) = p_0 + \frac{1}{z}[P(z) - p_0]$. Next prove that

$$\int_0^\infty e^{-sx}\{1-W_q(x)\}\,dx = \frac{1}{s}\left[1 - p_0 - \frac{\lambda p_0(1-b_0^*(s))}{s-\lambda+\lambda b^*(s)}\right],$$

where $b_0^*(s) = \int_0^\infty e^{-sx}b_0(x)\,dx$ is the Laplace transform of the density of the exceptional first service and $b^*(s) = \int_0^\infty e^{-sx}b(x)\,dx$ is the Laplace transform of the density of the ordinary service.

9.6 Consider again the $M/G/1$ queue with server vacations from Exercise 9.2. Assuming that service is in order of arrival, let $W_q(x)$ denote the limiting distribution function of the delay in queue of a customer.

(a) Letting $P_0(z) = \sum_{j=0}^\infty p_{0j}z^j$ and $P_1(z) = \sum_{j=1}^\infty p_{1j}z^j$, verify from the recursion scheme in Exercise 9.2 that

$$P_0(z) = \frac{1-\rho}{E(V)}\nu(z) \quad \text{and} \quad P_1(z) = zP_0(z)\frac{\lambda\alpha(z)}{1-\lambda\alpha(z)},$$

where $\nu(z) = \int_0^\infty e^{-\lambda(1-z)t}\{1-V(t)\}\,dt$ and $\alpha(z) = \int_0^\infty e^{-\lambda(1-z)t}\{1-B(t)\}\,dt$. Argue that relation (2.5.14) also applies to the $M/G/1$ queue with server vacations where $E(z^{L_q^{(\infty)}})$ is given by $P_0(z) + P_1(z)/z$.

(b) Verify that the Laplace transform of $1 - W_q(x)$ is given by

$$\int_0^\infty e^{-sx}\{1-W_q(x)\}\,dx = \frac{1-\eta^*(s)\xi^*(s)}{s}$$

where $\xi^*(s) = (1-\rho)s/[s-\lambda+\lambda b^*(s)]$ and $\eta^*(s)$ is the Laplace transform of the density $[1-V(x)]/E(V)$. Here $b^*(s)$ is the Laplace transform of the service-time density, $\xi^*(s)$ corresponds to $E(e^{-sD_\infty})$ in the standard $M/G/1$ queue without vacations and $\eta^*(s)$ is the

Laplace transform of the equilibrium excess density of the vacation time. Decomposition results of this type are discussed more generally in Fuhrmann and Cooper (1985).

9.7 Consider the (R, S) inventory model with limited order sizes. In this model the inventory position is reviewed every R periods. At each review the inventory position is ordered up to the level S provided that the order size does not exceed the constant Q; otherwise, the replenishment order is of size Q. It is assumed that $Q > \mu_R$, where μ_R is the mean demand between two reviews. The lead time of a replenishment order is negligible. The cumulative demands between successive reviews are independent random variables. Demand in excess of on-hand inventory is back ordered.

(a) Let the random variable ξ_k denote the cumulative demand between the kth and $(k+1)$th review and let Δ_i denote the difference between S, the order-up-to level, and the inventory position just after the ith review. Verify the Lindley equation

$$\Delta_i = \max(0, \Delta_{i-1} + \xi_{i-1} - Q).$$

(b) Use the results (9.5.5) and (9.5.17) for the $D/G/1$ queue to derive an explicit expression for the long-run fraction of demand that is back ordered when the demand variables ξ_k have a Coxian-2 distribution.

9.8 A certain product is produced at a constant rate of $r > 0$. The product is temporarily stored in a finite buffer with capacity K. The production is stopped when the buffer is full. A stopped production is resumed as soon as the stock level falls below K by a customer demand. Customers asking for the product arrive according to a Poisson process with rate λ. The demand of each customer is for a constant amount of D. The customer is satisfied with the amount in the buffer when the stock level is below D. It is assumed that $\lambda D < r$. One wishes to choose the buffer size K such that the long-run fraction of customers with partially unsatisfied demand is below a prespecified level α with α small. Use results from Section 9.4.1 to show that the required buffer $K(\alpha)$ is approximately given by

$$K(\alpha) \approx \frac{1}{\delta} \ln\left(\frac{\gamma\delta}{\lambda\alpha}\right),$$

where $\delta > 0$ is the unique solution of $e^{\delta D} = 1 + r\delta/\lambda$ and the constant γ is given by $\gamma = (1-\rho)/(\delta D - (1-\rho))$ with $\rho = \lambda D/r$.

9.9 A finite buffer storing a liquid material is emptied at a constant rate of $r > 0$. Customers bringing in the liquid material arrive according to a Poisson process with rate λ. The buffer has a finite capacity of $K > 0$. If a customer brings in an amount of work that is larger than the remaining room in the buffer, the whole amount of work of the customer is rejected. The amounts of work brought in by the customers are independent and identically distributed positive random variables. This queueing model is known as the $M/G/1$ queue with *bounded sojourn time*. Let $\pi_{rej}(K)$ be defined as the long-run fraction of customers who are rejected.

(a) For the case that the amount of work brought in by a customer is a constant D, argue that $\pi_{rej}(K)$ equals the loss probability in the $M/G/1$ queue with impatient customers from Section 9.4.2, where the service time equals D/r and the impatience time τ equals $(K-D)/r$. In particular, conclude that

$$\pi_{rej}(K) \sim \frac{(1-\rho)^2 e^{\delta D}}{\delta D - (1-\rho)} e^{-\delta K} \quad \text{as } K \to \infty$$

where $\rho = \lambda D/r < 1$ and $\delta > 0$ is the unique solution of $e^{\delta D} = 1 + r\delta/\lambda$. If the amount of work brought in by a customer has an exponential distribution with mean α, then it follows from results in Gavish and Schweitzer (1977) that

$$\pi_{rej}(K) \sim (1-\rho) e^{-\rho} e^{-(1-\rho)K/\alpha} \quad \text{as } K \to \infty$$

provided that $\rho = \lambda\alpha/r$ is smaller than 1. It is an open problem whether the asymptotically exponential expansion for $\pi_{rej}(K)$ holds when the amount of work brought in by a customer has a general distribution with a non-heavy tail.

(b) Let $K(\alpha)$ be the smallest value of K for which $\pi_{rej}(K) \leq \alpha$. Use the discretization method from Example 5.5.2 to investigate the performance of the two-moment approximation $K(\alpha) \approx (1 - c_S^2)K_{det}(\alpha) + c_S^2 K_{exp}(\alpha)$ for α small and $0 \leq c_S^2 \leq 2$, where $K_{det}(\alpha)$ and $K_{exp}(\alpha)$ are determined by the asymptotic expansions in (a). Here c_S^2 is the squared coefficient of variation of the amount of work brought in by a customer. This problem is based on De Kok and Tijms (1985).

9.10 Consider the $M/G/c$ queue with service in order of arrival. Prove that relation (2.5.14) remains valid. Derive from this relation that

$$E[L_q^{(\infty)}(L_q^{(\infty)} - 1)\cdots(L_q^{(\infty)} - k + 1)] = \lambda^k E(D_\infty^k), \quad k = 1, 2, \ldots .$$

9.11 Consider the $M/G/c$ queue with service in order of arrival. Let $V(x)$ denote the conditional waiting-time distribution function of a delayed customer. That means $V(x) = [W_q(x) - W_q(0)]/P_{delay}$. Denote by $v(x)$ the derivative of $V(x)$ for $x > 0$.

(a) Use relation (2.5.14) to verify that

$$\sum_{j=0}^{\infty} p_{c+j} z^j = P_{delay} \int_0^\infty e^{-\lambda(1-z)x} v(x)\, dx.$$

(b) Let p_j^{app} denote the approximation to p_j from Theorem 9.6.1 and let $v_{app}(x)$ be the corresponding approximation to $v(x)$. Use (9.6.21) and (9.6.23) to verify that the Laplace transform of $v_{app}(x)$ is given by

$$\int_0^\infty e^{-st} v_{app}(t)\, dt = \frac{(1-\rho)\alpha^*(s)}{1 - \rho\beta^*(s)},$$

where the Laplace transforms $\alpha^*(s)$ and $\beta^*(s)$ are given by

$$\alpha^*(s) = \frac{c}{\mu}\int_0^\infty e^{-st}\{1 - B_e(t)\}^{c-1}\{1 - B(t)\}\, dt, \quad \beta^*(s) = \frac{c}{\mu}\int_0^\infty e^{-st}\{1 - B(t)\}\, dt.$$

Here $B_e(t)$ is the excess equilibrium distribution function of the service time.

(c) Verify by inversion of the Laplace transform of $v_{app}(x)$ that

$$V_{app}(x) = (1-\rho)\{1 - (1 - B_e(x))^c\} + \lambda \int_0^x V_{app}(x-y)\{1 - B(cy)\}\, dy, \quad x \geq 0.$$

Assuming that the service-time distribution is not heavy-tailed, use the same arguments as in Section 8.4 to verify that

$$1 - V_{app}(x) \sim \frac{e^{-\delta x}\int_0^\infty e^{\delta y}[1 - \rho B_e(cy) - (1-\rho)\{1 - (1 - B_e(y))^c\}]\, dy}{\lambda \int_0^\infty y e^{\delta y}\{1 - B(cy)\}\, dy}$$

as $x \to \infty$, where $\delta > 0$ is the solution to $\lambda \int_0^\infty e^{\delta t}\{1 - B(ct)\}\, dt = 1$. This problem is based on Van Hoorn and Tijms (1982).

9.12 Use exact results from Section 9.5.3 to verify numerically that

$$P_{delay}^{app} = \frac{1 - B(D)}{\int_D^\infty e^{\delta(t-D)} b(t)\, dt} \quad \text{and} \quad W_q^{app} = \frac{\int_D^\infty (t - D) b(t)\, dt}{B(D) - 1 + \int_D^\infty e^{\delta(t-D)} b(t)\, dt}$$

are excellent approximations to P_{delay} and W_q in the $D/G/1$ queue. Here $B(t)$ and $b(t)$ are the probability distribution function and the probability density of the service time. The constant δ is the unique positive solution to $e^{-\delta D} \int_0^\infty e^{\delta y} b(y)\, dy = 1$. These approximations for the $D/G/1$ queue are due to Fredericks (1982).

9.13 Consider the machine-repair model from Exercise 5.2. Assume now that the service time S of a request has a general probability distribution function $B(x)$. Extend the approximate analysis of the $M/G/c$ queue in Section 9.6.2 to the machine-repair model. Verify that the resulting approximation to the limiting distribution $\{p_j\}$ of the number of service requests in the system is given by

$$p_j^{app} = \binom{N}{j} [\nu E(S)]^j p_0^{app}, \quad 0 \le j \le c - 1$$

$$p_j^{app} = (N - c + 1)\nu \alpha_{cj} p_{c-1}^{app} + \sum_{k=c}^{j} (N - k)\nu \beta_{kj} p_k^{app}, \quad c \le j \le N$$

with

$$\alpha_{cj} = \int_0^\infty \{1 - B_e(t)\}^{c-1} \{1 - B(t)\} \phi_{cj}(t)\, dt, \quad \beta_{kj} = \int_0^\infty \{1 - B(ct)\} \phi_{kj}(t)\, dt,$$

where $B_e(t)$ denotes the equilibrium excess distribution of $B(t)$ and $\phi_{kj}(t)$ is given by $\phi_{kj}(t) = \binom{N-k}{j-k} (1 - e^{-\nu t})^{j-k} e^{-\nu t(N-j)}$, $t > 0$ and $k \ge j \ge c$.

9.14 Consider the finite-capacity $M/D/c/c + N$ queue with deterministic services. It is assumed that the server utilization is less than 1. Let $W_q(x)$ be the limiting distribution of the delay in queue of an accepted customer. For $k = 1, \ldots, c$, let

$$U_k(x) = \sum_{j=k}^{c} \binom{c}{j} \left(\frac{x}{D}\right)^j \left(1 - \frac{x}{D}\right)^{c-j}, \quad 0 \le x \le D$$

be the probability distribution function of the kth order statistic of c independent random variables that are uniformly distributed on $(0, D)$. An approximation to $W_q(x)$ can be calculated by the following algorithm:

Step 0. Use the results of Theorem 9.8.1 to compute approximations p_j^{app} to the state probabilities p_j in the $M/D/c/c + N$ queue.

Step 1. Approximate $1 - W_q(x)$ by $\sum_{j=c}^{N+c-1} [p_j^{app}/(1 - p_{N+c}^{app})] V_j(x)$, where $V_{kc+r}(x)$ is given by $1 - U_{r+1}(x - kD)$ for $k \ge 0$ and $0 \le r \le c - 1$.

Use computer simulation to find out how well this approximation to $W_q(x)$ performs. Investigate the quality of the approximation to $W_q(x)$ which results by approximating p_j through γp_j^∞ for $0 \le j \le N + c - 1$ in accordance with (9.8.3), where $p_j^{(\infty)}$ is the state probability in the $M/D/c$ queue. Further, investigate how well the two-moment approximation (9.6.31) works for the conditional waiting-time percentiles in the $M/G/c/c + N$ queue (the computation of $W_q(x)$ in the $M/M/c/c + N$ queue is discussed in Exercise 5.1).

9.15 Consider a single-server queueing system in which the arrival process is the result of the superposition of m homogeneous on-off sources. Each source is alternately on and off, where the on-time has an exponential distribution with mean $1/\nu_{on}$ and the off-time has an exponential distribution with mean $1/\nu_{off}$. The sources act independently of each other.

Whenever a source is on, it generates service requests according to a Poisson process with rate δ. There is a buffer of capacity N for temporarily storing service requests which find the server busy upon arrival; an arriving service request finding the buffer full is rejected. The service time of a request is distributed as a mixture of Erlangian distributions with the same scale parameters. This queueing system is a special case of the so-called $MAP/G/1/N+1$ queue with a Markov modulated Poisson arrival process.

Develop a computer program to test the performance of formula (9.8.8) as an approximate formula for the rejection probabilities P_{rej}. Use the results from Exercise 5.27 to compute the customer-average probabilities $\pi_j^{(\infty)}$ in the infinite-buffer model. Check your computer program with the results that are given below for the case of E_2 services, $m = 25$ sources, $\nu_{on} = \nu_{off} = 0.1$ and the two values 0.2 and 0.8 for the system load ρ.

$\rho = 0.2$	exa	app	$\rho = 0.8$	exa	app
$N = 0$	1.708×10^{-1}	1.723×10^{-1}	$N = 0$	4.496×10^{-1}	4.616×10^{-1}
$N = 5$	1.766×10^{-5}	1.817×10^{-5}	$N = 5$	5.455×10^{-2}	5.825×10^{-2}
$N = 10$	1.972×10^{-9}	2.042×10^{-9}	$N = 50$	9.453×10^{-7}	1.002×10^{-6}
$N = 15$	2.437×10^{-13}	2.530×10^{-13}	$N = 100$	6.139×10^{-12}	6.511×10^{-12}

9.16 Consider the discrete-time $SBBP/D/c/c+N$ queueing system. In this model there is an underlying phase process that is alternately in the states 1 and 2. The sojourn times in the successive states are independent positive random variables that have a geometric distribution with mean $1/\omega_i$ in state i for $i = 1, 2$. If the phase process is in state i at the beginning of a time slot, then the number of packets arriving during that time slot has the discrete probability distribution $\{a_k^{(i)}, k \geq 0\}$ for $i = 1, 2$. This phase process is called a switched-batch Bernoulli process (SBBP). There is a buffer of capacity N to store incoming packets. Any arriving packet finding the buffer full is rejected. The transmission of a packet can only start at the beginning of a time slot. The transmission time of a packet is deterministic and equals any time slot. There are c service channels. Letting $\alpha_i = \sum_{k=1}^{\infty} k a_k^{(i)}$ for $i = 1, 2$, the system load ρ is defined by $\rho = \lambda/c$ with $\lambda = (\alpha_1/\omega_1 + \alpha_2/\omega_2)/(1/\omega_1 + 1/\omega_2)$ denoting the average arrival rate of packets. It is assumed that $\rho < 1$. For the infinite-buffer model, define $u_{n,i}^{(\infty)}$ as the long-run fraction of time slots at whose beginning n packets are present and the phase process is in state i. Let $U^{(i)}(z) = \sum_{n=0}^{\infty} u_{n,i}^{(\infty)} z^n$ for $i = 1, 2$.

(a) Use discrete-time Markov chain analysis to verify that

$$U^{(1)}(z) = \frac{\sum_{k=0}^{c-1} [A^{(1)}(z)\{\gamma_1 z^c - \gamma A^{(2)}(z)\} u_{k,1}^{(\infty)} + A^{(2)}(z)\omega_2 z^c u_{k,2}^{(\infty)}] \times (z^c - z^k)}{z^{2c} - [\gamma_1 A^{(1)}(z) + \gamma_2 A^{(2)}(z)] z^c + \gamma A^{(1)}(z) A^{(2)}(z)},$$

where $A^{(i)}(z) = \sum_{n=0}^{\infty} a_n^{(i)} z^n$ and $\gamma_i = 1 - \omega_i$ for $i = 1, 2$ and $\gamma = 1 - \omega_1 - \omega_2$. The expression for $U^{(2)}(z)$ is obtained by interchanging the roles of 1 and 2 in the expression for $U^{(1)}(z)$. Argue that

$$\sum_{i=1}^{2} \sum_{n=1}^{c-1} n u_{n,i}^{(\infty)} + c \left(1 - \sum_{i=1}^{2} \sum_{n=0}^{c-1} u_{n,i}^{(\infty)} \right) = c\rho$$

and argue that an additional $2c-1$ relations between the $2c$ unknowns $u_{n,i}^{(\infty)}$ for $0 \leq n \leq c-1$ and $i = 1, 2$ are obtained by noting that $U^{(1)}(z)$ and $U^{(2)}(z)$ are analytic for $|z| \leq 1$.

(b) For the infinite-buffer model, let $\pi_j^{(\infty)}$ be the long-run fraction of packets that find j other packets in the system upon arrival. Argue that

$$\pi_j^{(\infty)} = \frac{1}{\lambda} \sum_{k=0}^{j} \sum_{i=1}^{2} u_{k,i}^{(\infty)} \sum_{s=j-k+1}^{\infty} a_s^{(i)}, \quad j = 0, 1, \ldots .$$

(c) Develop a computer program for the discrete-time $SBBP/D/c/c+N$ queue. Check your computer program with the results below for the parameter values $c = 3, \omega_1 = 0.4, \omega_2 = 0.2, \alpha_1 = 1.4$ and $\alpha_2 = 2.0$. In case 1 a Poisson distribution is taken for each of the distributions $\{a_n^{(1)}\}$ and $\{a_n^{(2)}\}$; in case 2 a geometric distribution is taken for $\{a_n^{(1)}\}$ and a Poisson distribution for $\{a_n^{(2)}\}$.

		$N = 5$	$N = 10$	$N = 20$	$N = 30$
Case 1	exa	1.683×10^{-2}	2.194×10^{-4}	3.908×10^{-8}	6.965×10^{-12}
	app	1.085×10^{-2}	1.689×10^{-4}	3.024×10^{-8}	5.390×10^{-12}
Case 2	exa	3.781×10^{-2}	2.927×10^{-3}	3.101×10^{-5}	3.413×10^{-7}
	app	2.603×10^{-2}	2.245×10^{-3}	2.506×10^{-5}	2.816×10^{-7}

9.17 Consider the $D/M/c/c+N$ queue and the $M/M/c/c+N$ queue with the same average arrival rate and the same mean service time. For these two models, denote by $N_{det}(\alpha)$ and $N_{exp}(\alpha)$ the smallest value of N for which the rejection probability is below a prespecified level α. Verify experimentally that $N_{det}(\alpha) \approx \frac{1}{2} N_{exp}(\alpha)$.

9.18 Consider the finite-capacity variants of the $M/G/1$ queue with exceptional first service from Exercise 9.1, the $M/G/1$ queue with server vacations from Exercise 9.2 and the $M/G/1$ queue with variable service rate from Exercise 9.3. Verify that the structural form (9.8.4) for P_{rej} remains valid for these queueing models. Do the same for the finite-capacity variant of the $M/M/c$ queue with impatient customers from Exercise 5.3.

9.19 Consider the batch-arrival $M^X/G/c/N+c$ queue with complete rejection of a batch when an arriving batch of customers does not find enough room in the buffer for the whole batch. Let P_{rej} denote the long-run fraction of customers who are rejected.

(a) Argue that

$$P_{rej} = \frac{1}{\beta} \sum_{k=0}^{N+c} p_k \sum_{s>N+c-k} s\beta_s.$$

(b) Using the approximation assumption $p_j \approx \gamma p_j^{(\infty)}$ for $j = 0, 1, \ldots, N+c-1$, modify the proof of part (b) of Theorem 9.8.2 to obtain the approximation (9.8.13) to P_{rej}.

9.20 Consider the batch-arrival $M^X/G/c/c+N$ queue with complete rejection. Suppose that the batch-size distribution $\{\beta_j\}$ has the property that $\sum_{s=1}^{Q} \beta_s = 1$ for some $1 \le Q \le N+1$. Prove that $p_j = \gamma p_j^{(\infty)}$ for $0 \le j \le N+c-Q$ for both the $M^X/G/1/N+1$ queue and the $M^X/M/c/c+N$ queue. (*Hint:* define T, T_j, $T^{(\infty)}$, $T_j^{(\infty)}$ as in the proof of part (a) of Theorem 9.8.2 and let N_k and $N_k^{(\infty)}$ denote the number of service completions in one cycle at which k customers are left behind. Argue first that $E(N_k) = E(N_k^{(\infty)})$ for $0 \le k \le N+c-Q$. Next conclude that $E(T_j) = E(T_j^{(\infty)})$ for $0 \le j \le N+c-Q$, since $E(N_j) = \lambda E[T_j + \ldots + T_{j+1-Q}]$ for $0 \le j \le N+c-Q$.)

9.21 Consider the batch-arrival $M^X/G/c/c+N$ queue with complete rejection. Suppose that the batch size is a constant Q with $1 \leq Q \leq N+1$. Prove that the approximation (9.8.13) to P_{rej} is exact for both the $M^X/G/1/N+1$ queue and the $M^X/M/c/c+N$ queue.

BIBLIOGRAPHIC NOTES

The queueing theory literature is voluminous. A good account of the basic theory is provided by the books of Cooper (1991), Kleinrock (1975,1976) and Takács (1962). A book emphasizing the analysis of the transient behaviour of queues is Newell (1971). A thorough treatment of most of the background material in Section 9.1 can be found in the book of Wolff (1989). The regenerative approach used in Sections 9.2 and 9.3 to analyse single-server queues with Poisson input has its origin in the paper of Hordijk and Tijms (1976). This versatile approach was used in Tijms *et al*. (1981) and Tijms and Van Hoorn (1982) to give an approximate analysis of multi-server queues with state-dependent Poisson input; see also Van Hoorn (1984). For finite-capacity queues of the $M/G/1$ type the structural form for the rejection probability was noticed in the papers of Keilson and Servi (1989) and Tijms and Van Ommeren (1989). The papers of Sakasegawa *et al*. (1993) and Tijms (1992) provide theoretical and empirical support to this formula as an approximation to a broad class of queueing systems; see also Gouweleeuw (1996). The material on two-moment approximations for the minimal buffer size is based on De Kok and Tijms (1985) and Gouweleeuw and Tijms (1996).

REFERENCES

Abate, J. and Whitt, W. (1992) Solving probability transform functional equations for numerical inversion. *Operat. Res. Lett.*, **12**, 275–281.

Ackroyd, M.H. (1980) Computing the waiting-time distribution for the $G/G/1$ queue by signal processing methods. *IEEE Trans. Commun.*, **28**, 52–58.

Anick, D., Mitra, D. and Sondhi, M.M. (1982) Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, **61**, 1871–1894.

Boots, N.K. and Tijms, H.C. (1999) A multi-server queueing system with impatient customers. *Management Sci.*, **45**, 444–448.

Boxma, O.J., Cohen, J.W. and Huffels, N. (1979) Approximations of the mean waiting time in an $M/G/s$ queueing system. *Operat. Res.*, **27**, 1115–1127.

Cohen, J.W. (1982) *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

Cooper, R.B. (1991) *Introduction to Queueing Theory*, 2nd edn. North-Holland, Amsterdam.

Cosmetatos, G.P. (1976) Some approximate equilibrium results for the multi-server queue $M/G/r$. *Operat. Res. Quart.*, **27**, 615–620.

Crommelin, C.D. (1932) Delay probability formulas when the holding times are constant. *Post Office Electr. Engng. J.*, **25**, 41–50.

De Kok, A.G. (1984) Algorithmic methods for single server systems with repeated attempts. *Statistica Neerlandica*, **38**, 23–32.

De Kok, A.G. (1989) A moment-iterating method for approximating the waiting-time characteristics of the $GI/G/1$ queue. *Prob. Engng Inform. Sci.*, **3**, 273–288.

De Kok, A.G. and Tijms, H.C. (1985) A two-moment approximation for a buffer design problem requiring a small rejection probability. *Performance Evaluation*, **7**, 77–86.

Franx, G.J. (2001) A simple solution for the $M/D/c$ waiting-time distribution. *Operat. Res. Lett.*, **29**, 221–229.

Franx, G.J. (2002) The waiting-time distribution for the $M^X/D/c$ queue. *Prob. Engng Inform. Sci.*, submitted.

Fredericks, A.A. (1982) A class of approximations for the waiting time distribution in a $GI/G/1$ queueing system. *Bell System Techn. J.*, **61**, 295–325.

Fuhrmann, S.W. and Cooper, R.B. (1985) Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operat. Res.*, **33**, 1117–1129.

Gavish, B. and Schweitzer, P.J. (1977) The Markovian queue with bounded waiting time. *Management Sci.*, **23**, 1349–1357.

Gouweleeuw, F.N. (1996) *A General Approach to Computing Loss Probabilities in Finite-Buffer Queues*. Tinbergen Institute, Amsterdam.

Gouweleeuw, F.N. and Tijms, H.C. (1996) A simple heuristic for buffer design in finite-capacity queues. *Eur. J. Operat. Res.*, **88**, 592–598.

Gouweleeuw, F.N. and Tijms, H.C. (1998) Computing loss probabilities in discrete-time queues. *Operat. Res.*, **46**, 149–153.

Hooghiemstra, G. (1987) A path construction for the virtual waiting-time of an $M/G/1$ queue. *Statistica Neerlandica*, **45**, 175–181.

Hordijk, A. and Tijms, H.C. (1976) A simple proof of the equivalence of the limiting distributions of the continuous-time and the embedded process of the queue size in the $M/G/1$ queue. *Statistica Neerlandica*, **30**, 97–100.

Keilson, J. and Servi, L.D. (1989) Blocking probability for $M/G/1$ vacation systems with occupancy level dependent schedules. *Operat. Res.*, **37**, 134–140.

Kleinrock, L. (1975) *Queueing Systems,* Vol. I, *Theory.* John Wiley & Sons, Inc., New York.

Kleinrock, L. (1976) *Queueing Systems,* Vol. II, *Computer Applications.* John Wiley & Sons, Inc., New York.

Krämer, W. and Langenbach-Belz, M. (1976) Approximate formulas for the delay in the queueing system $GI/G/1$. In: *Proc. 8th International Teletraffic Congress,* Melbourne, paper 235, pp. 1–8. North-Holland, Amsterdam.

Newell, G.F. (1971) *Applications of Queueing Theory*. Chapman and Hall, London.

Page, E. (1972) *Queueing Theory in O.R.* Butterworth, London.

Sakasegawa, H., Miyazawa, M. and Yamazaki, G. (1993) Evaluating the overflow probability using the infinite queue. *Management Sci.*, **39**, 1238–1245.

Schwartz, M. (1996) *Broadband Integrated Networks*. Prentice Hall, Englewood Cliffs NJ.

Seelen, L.P., Tijms, H.C. and Van Hoorn, M.H. (1985) *Tables for Multi-Server Queues*. North-Holland, Amsterdam.

Takács, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, Oxford.

Takahashi, Y. (1981) Asymptotic exponentiality of the tail of the waiting time distribution in a $Ph/Ph/c$ queue. *Adv. Appl. Prob.*, **13**, 619–630.

Takahashi, Y. and Takami, Y. (1976) A numerical method for the steady-state probabilities of a $GI/G/c$ queueing system in a general class. *J. Operat. Res. Soc. Japan*, **19**, 147–157.

Tijms, H.C. (1992) Heuristics for finite-buffer queues. *Prob. Engng Inform. Sci.*, **6**, 277–285.

Tijms, H.C. and Van Hoorn, M.H. (1982) Computational methods for single-server and multi-server queues with Markovian input and general service times. In *Applied Probability Computer Sciences, The Interface*, edited by R.L. Disney and T.J. Ott, Vol. II, pp. 71–102. Birkhäuser, Boston.

Tijms, H.C., Van Hoorn, M.H. and Federgruen, A. (1981) Approximations for the steady-state probabilities in the $M/G/c$ queue. *Adv. Appl. Prob.*, **13**, 186–206.

Tijms, H.C. and Van Ommeren, J.W. (1989) Asymptotic analysis for buffer behavior in communication systems. *Prob. Engng Inform. Sci.*, **3**, 1–12.

Tran-Gia, P. (1986) Discrete-time analysis for the interdeparture distribution of $GI/G/1$ queues. In *Teletraffic Analysis and Computer Performance Evaluation*, edited by O.J. Boxma, J.W. Cohen and H.C. Tijms, pp. 341–357. North-Holland, Amsterdam.

Van Hoorn, M.H. (1984) *Algorithms and Approximations for Queueing Systems*. CWI, Amsterdam.

Van Hoorn, M.H. and Seelen, L.P. (1986) Approximations for the $GI/G/c$ queue. *J. Appl. Prob.*, **23**, 484–494.

Van Hoorn, M.H. and Tijms, H.C. (1982) Approximation for the waiting time distribution of the $M/G/c$ queue. *Performance Evaluation*, **2**, 22–28.

Van Ommeren, J.W. (1988) Exponential expansion for the tail of the waiting-time probabilities in the single-server queue with batch arrivals. *Adv. Appl. Prob.*, **20**, 880–895.

Wolff, R.W. (1989) *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs NJ.

Appendices

APPENDIX A. USEFUL TOOLS IN APPLIED PROBABILITY

This appendix summarizes some basic tools that can be found in most introductory texts on probability.

Law of total expectation

In many applied probability problems it is only possible to compute certain probabilities and expectations by using appropriate conditioning arguments. Since conditional expectations are based on additional information, they are often easier to compute than unconditional expectations. The *law of total expectation* states that, for any two random variables X and Y defined on the same probability space,

$$E(X) = \sum_y E(X \mid Y = y) P\{Y = y\} \tag{A.1}$$

when Y has a discrete distribution and

$$E(X) = \int_{-\infty}^{\infty} E(X \mid Y = y) f(y)\, dy \tag{A.2}$$

when Y has a continuous distribution with probability density $f(y)$. It is assumed that the relevant expectations exist. The *law of total probability* is a special case of the law of total expectation:

$$P\{X \leq x\} = \sum_y P\{X \leq x \mid Y = y\} P\{Y = y\} \tag{A.3}$$

when Y has a discrete distribution and

$$P\{X \leq x\} = \int_{-\infty}^{\infty} P\{X \leq x \mid Y = y\} f(y)\, dy \tag{A.4}$$

A First Course in Stochastic Models H.C. Tijms
 ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

when Y has a continuous distribution with probability density $f(y)$. The law of total expectation and the law of total probability will be frequently used in this book. We illustrate these laws by two examples.

Example A.1 Service with interruptions

A single unloader is available to unload ships. The unloading time U of a ship has a given probability density $f(t)$ with finite mean γ. The unloading process, however, is subject to interruptions. Those interruptions have exogenous causes and occur according to a Poisson process with rate λ. The durations of the interruptions are independent and identically distributed random variables with mean δ. After an interruption the unloading of the ship is resumed at the point it was stopped by the interruption. What is the expected amount of time needed to complete the unloading of the ship?

Letting the completion time C denote the total amount of time needed to complete the unloading of the ship, the answer to the above question is

$$E(C) = \gamma(1 + \lambda\delta). \tag{A.5}$$

To verify this, let N denote the number of interruptions during the unloading of the ship. By conditioning upon the unloading time U of the ship, it follows from the law of total probability that

$$\begin{aligned} P\{N = n\} &= \int_0^\infty P\{N = n \mid U = t\} f(t)\, dt \\ &= \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} f(t)\, dt, \quad n = 0, 1, \ldots . \end{aligned}$$

By conditioning on N and letting R_i denote the duration of the ith interruption, it follows from the law of total expectation that

$$\begin{aligned} E(C) &= \sum_{n=0}^\infty E(C \mid N = n) P\{N = n\} \\ &= \sum_{n=0}^\infty E(U + R_1 + \cdots + R_n \mid N = n) P\{N = n\} \\ &= \sum_{n=0}^\infty E(U \mid N = n) P\{N = n\} + \sum_{n=1}^\infty E(R_1 + \cdots + R_n) P\{N = n\} \end{aligned}$$

and so

$$E(C) = E(U) + \sum_{n=1}^\infty n E(R_1) P\{N = n\} = E(U) + E(R_1)E(N).$$

Since $E(N \mid U = t) = \lambda t$, we have $E(N) = \int_0^\infty \lambda t f(t)\, dt = \lambda\gamma$ and thus (A.5) follows.

Example A.2 The double-up strategy in roulette

In European roulette the wheel is divided in 37 sections, numbered as $1, \ldots, 36$ and 0. Of the sections numbered from 1 to 36, 18 are red and 18 are black. The section marked 0 is assumed to be winning for the house. You have decided to bet on 10 spins of the wheel and to use the double-up strategy. You bet each time on red. Your initial bet is €1. You double your bet each time red does not come up. If red appears, you start again with a bet of €1. You get paid twice your bet when red comes up and you lose your bet otherwise. What is the expected value of your loss after a playing round of 10 bets and what is the expected value of the total amount you bet during the playing round?

To answer these questions, note that the betting process starts anew each time red comes up except that fewer bets are left. Instead of considering 10 bets, consider a playing round of n bets under the double-up strategy and define the random variables L_n and A_n by

$$L_n = \text{the player's loss after a playing round of } n \text{ bets},$$

$$A_n = \text{the total amount the player bets in a playing round of } n \text{ bets}.$$

To compute the expected values of L_n and A_n, it is natural to condition on the random variable Y denoting the number of spins of the wheel until red comes up for the first time. Obviously, Y is geometrically distributed with parameter $p = \frac{18}{37}$. By conditioning on Y and noting that the player's profit is €1 each time red comes up, it follows that

$$\begin{aligned} E(L_n) &= \left[-1 + E(L_{n-1})\right] p + \left[-1 + E(L_{n-2})\right] (1-p)\, p + \cdots \\ &\quad + \left[-1 + E(L_1)\right] (1-p)^{n-2}\, p + \left[1 + 2 + \cdots + 2^{n-1}\right] (1-p)^n, \\ E(A_n) &= \left[1 + E(A_{n-1})\right] p + \left[1 + 2 + E(A_{n-2})\right] (1-p)\, p + \cdots + \left[1 + 2\right. \\ &\quad \left. + \cdots + 2^{n-2} + E(A_1)\right] (1-p)^{n-2} p + \left[1 + 2 + \cdots + 2^{n-1}\right] (1-p)^{n-1}. \end{aligned}$$

Since $1 + 2 + \cdots + 2^{k-1} = 2^k - 1$, we thus have the recursions

$$E(L_n) = \sum_{k=0}^{n-1} \left[-1 + E(L_{n-k-1})\right] (1-p)^k\, p + \left(2^n - 1\right) (1-p)^n,$$

$$E(A_n) = \sum_{k=0}^{n-2} \left[2^{k+1} - 1 + E(A_{n-k-1})\right] (1-p)^k\, p + \left(2^n - 1\right) (1-p)^{n-1}$$

for $n \geq 1$ with the boundary condition $E(L_0) = E(A_0) = 0$. These relations enable us to compute recursively the values of $E(L_n)$ and $E(A_n)$. In particular, $E(L_{10}) = 0.9421$ and $E(A_{10}) = 34.858$. To conclude, we remark that explicit expressions for $E(L_n)$ and $E(A_n)$ can be derived from the recursive relations by

using the generating-function technique to be discussed in Appendix C. Omitting the details, we state

$$E(L_n) = -pn + \frac{1-p}{1-2p}\left[(2\,(1-p))^n - 1\right], \quad n \geq 1,$$

$$E(A_n) = \frac{1}{1-2p}E(L_n), \quad n \geq 1.$$

Indeed $E(L_n)/E(A_n) = 1 - 2p = \frac{1}{37}$ in accordance with the fact that the house percentage in European roulette is 2.702%.

Convolution formula

Let X_1 and X_2 be two independent, non-negative random variables with respective probability distribution functions $F_1(x)$ and $F_2(x)$. For ease assume that $F_2(x)$ has a probability density $f_2(x)$. Then, by a direct application of the law of total probability, we have the convolution formula

$$P\{X_1 + X_2 \leq x\} = \int_0^x F_1(x-y)\, f_2(y)\, dy, \quad x \geq 0.$$

Moments of a non-negative random variable

Let N be a non-negative, integer-valued random variable. A useful formula is

$$E(N) = \sum_{k=0}^{\infty} P\{N > k\}. \tag{A.6}$$

To verify this result, write $\sum_{k=0}^{\infty} P\{N > k\} = \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} P\{N = j\}$ and interchange the order of summation. The relation (A.6) can be generalized. For any non-negative random variable X with probability distribution function $F(x)$,

$$E(X) = \int_0^{\infty} [1 - F(x)]\, dx. \tag{A.7}$$

A probabilistic proof of (A.7) is as follows. Imagine that X is the lifetime of a machine. Define the indicator variable $I(t)$ by $I(t) = 1$ if the machine is still working at time t and by $I(t) = 0$ otherwise. Then, by $E[I(t)] = P\{I(t) = 1\}$ and $P\{I(t) = 1\} = P\{X > t\}$, it follows that

$$E(X) = E\left[\int_0^{\infty} I(t)\, dt\right] = \int_0^{\infty} E\,[I(t)]\, dt = \int_0^{\infty} P\{X > t\}\, dt,$$

which proves (A.7). The interchange of the order of expectation and integration is justified by the non-negativity of $I(t)$. The result (A.7) can be extended to

$$E(X^k) = k \int_0^{\infty} x^{k-1}\, [1 - F(x)]\, dx, \quad k = 1, 2, \ldots\,. \tag{A.8}$$

To see this, note that (A.7) implies

$$E(X^k) = \int_0^\infty P\{X^k > t\}\, dt = \int_0^\infty P\{X > t^{1/k}\}\, dt$$

and next use the change of variable $t = x^k$.

Mean and variance of a random sum of random variables

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables whose first two moments are finite. Also, let N be a non-negative and integer-valued random variable having finite first two moments. If the random variable N is independent of the random variables $X_1, X_2, \ldots$, then

$$E\left(\sum_{k=1}^{N} X_k\right) = E(N)E(X_1), \tag{A.9}$$

$$\text{var}\left(\sum_{k=1}^{N} X_k\right) = E(N)\text{var}(X_1) + \text{var}(N)E^2(X_1), \tag{A.10}$$

where $E^2(X_1)$ is the shorthand notation for $[E(X_1)]^2$. The proof uses the law of total expectation. By conditioning on N, we find

$$\begin{aligned} E\left(\sum_{k=1}^{N} X_k\right) &= \sum_{n=0}^{\infty} E\left(\sum_{k=1}^{N} X_k \mid N = n\right) P\{N = n\} \\ &= \sum_{n=0}^{\infty} E\left(\sum_{k=1}^{n} X_k\right) P\{N = n\} = \sum_{n=0}^{\infty} nE(X_1)P\{N = n\}, \end{aligned}$$

which verifies (A.9). Note that the second equality uses that the random variables $X_1, \ldots, X_n$ are independent of the event $\{N = n\}$. Similarly,

$$\begin{aligned} E\left[\left(\sum_{k=1}^{N} X_k\right)^2\right] &= \sum_{n=0}^{\infty} E\left[\left(\sum_{k=1}^{N} X_k\right)^2 \mid N = n\right] P\{N = n\} \\ &= \sum_{n=0}^{\infty} [nE(X_1^2) + n(n-1)E^2(X_1)]P\{N = n\} \\ &= E(N)E(X_1^2) + E[N(N-1)]E^2(X_1). \end{aligned} \tag{A.11}$$

Using $\sigma^2(S) = E(S^2) - E^2(S)$, we obtain (A.10) from (A.9) and (A.11).

Wald's equation

The result (A.9) remains valid when the assumption that the random variable N is independent of the sequence $X_1, X_2, \ldots$ is somewhat weakened. Suppose that the following conditions are satisfied:

(i) $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with finite mean,

(ii) N is a non-negative, integer-valued random variable with $E(N) < \infty$,

(iii) the event $\{N = n\}$ is independent of $X_{n+1}, X_{n+2}, \ldots$ for each $n \geq 1$.

Then it holds that

$$E\left(\sum_{k=1}^{N} X_k\right) = E(X_1)E(N). \tag{A.12}$$

This equation is known as Wald's equation. It is a very useful result in applied probability. To prove (A.12), let us first assume that the X_i are non-negative. The following trick is used. For $n = 1, 2, \ldots$, define the random variable I_k by

$$I_k = \begin{cases} 1 & \text{if } N \geq k, \\ 0 & \text{if } N < k. \end{cases}$$

Then $\sum_{k=1}^{N} X_k = \sum_{k=1}^{\infty} X_k I_k$ and so

$$E\left(\sum_{k=1}^{N} X_k\right) = E\left(\sum_{k=1}^{\infty} X_k I_k\right) = \sum_{k=1}^{\infty} E(X_k I_k),$$

where the interchange of the order of expectation and summation is justified by the non-negativity of the random variables involved. The random variable I_k can take on only the two values 0 and 1. The outcome of I_k is completely determined by the event $\{N \leq k-1\}$. This event depends on $X_1, \ldots, X_{k-1}$, but not on $X_k, X_{k+1}, \ldots$. This implies that I_k is independent of X_k. Consequently, $E(X_k I_k) = E(X_k)E(I_k)$ for all $k \geq 1$. Since $E(I_k) = P\{I_k = 1\}$ and $P\{I_k = 1\} = P\{N \geq k\}$, we obtain (A.9) from (A.6) and

$$E\left(\sum_{k=1}^{N} X_k\right) = \sum_{k=1}^{\infty} E(X_1)P\{N \geq k\}.$$

For the general case, treat separately the positive and negative parts of the X_i.

The assumption $E(N) < \infty$ is essential in Wald's equation. To illustrate this, consider the symmetric random walk $\{S_n, n \geq 0\}$ with $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$, where $X_1, X_2, \ldots$ is a sequence of independent random variables with $P\{X_i = 1\} = P\{X_i = -1\} = \frac{1}{2}$ for all i. Define the random variable N as $N = \min\{n \geq 1 \mid S_n = -1\}$, that is, N is the epoch of the first visit of the random walk to the level -1. Then $E(X_1 + \cdots + X_N) = -1$. Noting that $E(X_i) = 0$, we

have, however, that $E(X_1 + \cdots + X_N)$ is not equal to $E(N)E(X_1)$. The reason is that $E(N) = \infty$.

Example A.3 A reliability problem

To illustrate Wald's equation, consider the following reliability problem. An electronic system has a built-in redundancy in the form of a standby unit to support an operating unit. The two units are identical. When the operating unit fails, its tasks are immediately taken over by the standby unit if available. A failed unit immediately enters repair. The system goes down when the operating unit fails while the other unit is still in repair. The lifetime L of an operating unit is assumed to have a continuous probability distribution $F(x)$ with finite mean μ. The repair time of a failed unit is a constant $\alpha > 0$. The successive lifetimes of the operating unit are independent of each other. A repaired unit is as good as new. Both units are in perfect condition at time 0. What is the expected time until the system goes down for the first time?

To solve this problem, denote by L_0 the lifetime of the operating unit installed at time 0 and denote by $L_1, L_2, \ldots$ the lifetimes of the subsequent operating units. Then the time until the first system failure is distributed as $L_0 + L_1 + \cdots + L_N$, where the random variable N denotes the first $n \geq 1$ for which L_n is less than the nth repair time. The random variables $L_1, \ldots, L_n$ and the event $\{N = n\}$ are mutually dependent, but the event $\{N = n\}$ is independent of $L_{n+1}, L_{n+2}, \ldots$ for each $n \geq 1$. Hence we can apply Wald's equation. This gives

$$\begin{aligned} E(\text{time until the first system failure}) &= E(L_0) + E(L_1)E(N) \\ &= \mu\,[1 + E(N)]\,. \end{aligned}$$

To find $E(N)$, note that N has a geometric distribution with parameter $p = P\{L < \alpha\}$. Hence $E(N) = 1/F(\alpha)$ and so

$$E(\text{time until the first system failure}) = \mu\left[1 + \frac{1}{F(\alpha)}\right].$$

In practical applications the mean lifetime will be much larger than the mean repair time. In other words, the occurrence of a system failure is a *rare event*. For those situations there is a deep but extremely useful result stating that the time until the first system failure is approximately *exponentially distributed*; see also Example 2.2.4.

Coefficient of variation

Let X be a positive random variable with finite mean $E(X)$ and finite standard deviation $\sigma(X)$. The *coefficient of variation* of X is defined by

$$c_X = \frac{\sigma(X)}{E(X)}.$$

Since this quantity is dimensionless, it is a very useful measure for the variability of the random variable X. Usually one works with the squared coefficient of variation c_X^2 rather than with c_X. For example, the deterministic distribution has $c_X^2 = 0$, the exponential distribution has $c_X^2 = 1$ and the Erlang distribution with shape parameter k has the intermediate value $c_X^2 = 1/k$.

Failure rate function

Let X be a positive random variable with a probability distribution function $F(t)$ and a continuous probability density $f(t)$. For example, the random variable X represents the lifetime of some item. The *failure, or hazard, rate function* of the random variable X is defined by

$$r(t) = \frac{f(t)}{1 - F(t)}$$

for those values of t with $F(t) < 1$. The failure rate has a useful probabilistic interpretation. Think of the random variable X as the lifetime of an item. The probability that an item of age t will fail in the next Δt time units is given by

$$\begin{aligned} P\{t < X \leq t + \Delta t \mid X > t\} &= \frac{P\{t < X \leq t + \Delta t\}}{P\{X > t\}} \\ &= \frac{f(t)\Delta t}{1 - F(t)} + o(\Delta t) \text{ as } \Delta t \to 0. \end{aligned}$$

Hence $r(t)\Delta t$ gives approximately the probability that an item of age t will fail in the next Δt time units when Δt is small. Hence the name 'failure rate'. Noting that $-r(t)$ is the derivative of the function $\ln[1 - F(t)]$, it follows that the failure rate function $r(t)$ determines uniquely the corresponding lifetime distribution function $F(t)$ by

$$1 - F(t) = \exp\left\{-\int_0^t r(x)\,dx\right\}, \quad t \geq 0.$$

As a consequence, the case of a *constant* failure rate $r(x) = \lambda$ for all x corresponds to the *exponential* distribution function $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$. In other words, an item in use is as good as new when its lifetime is exponentially distributed. Other important cases are the case of an *increasing* failure rate (the older, the worse) and the case of a *decreasing* failure rate (the older, the better). A random variable with an increasing (decreasing) failure rate can be shown to have the property that its coefficient of variation is smaller (larger) than 1. The failure rate is a concept that enables us to discriminate between distributions on physical considerations.

Convergence theorems

To conclude this appendix, we state a number of basic convergence theorems that will be used in this book. These theorems can be found in any textbook on real analysis, e.g. Rudin (1964).

Theorem A.1 *Let $a_{nm}, n, m = 0, 1, \ldots$ be real numbers. If all the numbers a_{nm} are non-negative or if $\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} |a_{nm}| < \infty$, then*

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a_{nm} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{nm}.$$

This theorem is a special case of what is known as Fubini's theorem in analysis.

Theorem A.2 *Let $\{p_m, m = 0, 1, \ldots\}$ be a sequence of non-negative numbers. Suppose that the numbers a_{nm}, $n, m = 0, 1, \ldots$ are such that*

$$\lim_{n \to \infty} a_{nm} = a_m$$

exists for all $m = 0, 1, \ldots$.

(a) If all numbers a_{nm} are non-negative, then

$$\lim_{n \to \infty} \inf \sum_{n=0}^{\infty} a_{nm} p_m \geq \sum_{m=0}^{\infty} a_m p_m.$$

(b) If there is a finite constant $M > 0$ such that $|a_{nm}| \leq M$ for all n, m and if $\sum_{m=0}^{\infty} p_m < \infty$, then

$$\lim_{n \to \infty} \sum_{m=0}^{\infty} a_{nm} p_m = \sum_{m=0}^{\infty} a_m p_m.$$

The first part of the theorem is a special case of *Fatou's lemma* and the second part of the theorem is a special case of the *bounded convergence theorem.*

The above theorems can be stated in greater generality. For example, a more general version of the bounded convergence theorem is as follows. Let $\{X_n\}$ be a sequence of random variables that converge with probability 1 to a random variable X. Then

$$\lim_{n \to \infty} E(X_n) = E(X)$$

provided that $|X_n| \leq Y$, $n \geq 1$, for some random variable Y with $E(Y) < \infty$. Recall that convergence with probability 1 means that

$$P\{\omega \in \Omega: \lim_{n \to \infty} X_n(\omega) = X(\omega)\} = 1,$$

where Ω is the common sample space of the random variables X_n, $n \geq 1$, and the random variable X. Often one uses the term 'almost sure convergence' instead of the term 'convergence with probability 1'.

Finally, we mention the important concept of the Cesaro limit. A sequence $\{a_n, n \geq 1\}$ of real numbers is said to have a *Cesaro limit* if $\lim_{n \to \infty} (1/n) \sum_{k=1}^{n} a_k$ exists. A sequence $\{a_n\}$ may have a Cesaro limit while the ordinary limit does

not exist. For example, suppose that $a_n = 1$ for n even and $a_n = 0$ for n odd. Then $\lim_{n\to\infty} a_n$ does not exist, while $\lim_{n\to\infty}(1/n)\sum_{k=1}^{n} a_k = 1/2$. However, if the ordinary limit exists then the Cesaro limit exists as well and is equal to the ordinary limit.

APPENDIX B. USEFUL PROBABILITY DISTRIBUTIONS

This appendix discusses a number of important distributions which have been found useful for describing random variables in inventory, reliability and queueing applications. In particular, attention is paid to the practical problem of fitting a tractable distribution to the first two moments of a positive random variable.

The exponential distribution

A positive random variable X is said to be exponentially distributed with parameter $\lambda > 0$ when it has the probability density

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

The corresponding probability distribution function $F(t)$ is given by

$$F(t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

Its mean and squared coefficient of variation are given by

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad c_X^2 = 1.$$

The exponential distribution is of extreme importance in applied probability. The main reason for this is its memoryless property and its intimate relation with the Poisson process. The *memoryless property* states that

$$P\{X > t + x \mid X > t\} = e^{-\lambda x}, \quad x \geq 0,$$

independently of t. In other words, imagining that X represents the lifetime of an item, the residual life of the item has the *same* exponential distribution as the original lifetime, regardless of how long the item has already been in use. The memoryless property is in agreement with the constant failure rate property of the exponential distribution.

The following well-known results for the exponential distribution are very useful. If X_1 and X_2 are two independent random variables that are exponentially distributed with respective means $1/\lambda_1$ and $1/\lambda_2$, then, for any $t \geq 0$,

$$P\{\min(X_1, X_2) \leq t\} = 1 - e^{-(\lambda_1+\lambda_2)t} \quad \text{and} \quad P\{X_1 < X_2\} = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \quad \text{(B.1)}$$

In other words, the minimum of the two exponentially distributed lifetimes X_1 and X_2 is exponentially distributed with mean $1/(\lambda_1 + \lambda_2)$ and the probability that the lifetime X_1 expires earlier than the lifetime X_2 is $\lambda_1/(\lambda_1 + \lambda_2)$.

Example B.1 A first-passage time problem

An electronic system has two crucial components, 1 and 2, that operate independently of each other. The lifetime of component i has an exponential distribution with mean $1/\alpha_i$ for $i = 1, 2$. If a component breaks down, it is replaced by a new one. The time needed to replace component i by a new one is exponentially distributed with mean $1/\beta_i$ for $i = 1, 2$. The system continues to operate as long as one of the components is functioning, but it fails when none of the two components works. Both components are in perfect condition at time 0. What is the expected time until the first system failure?

Let us say that the system is in state 1 (2) if only component 1 (2) is functioning and it is in state 3 when both components are functioning. In view of the memoryless property of the exponential distribution, we can define the random variable T_i as the time until the first system failure when the current state of the system is state i. We wish to compute $E(T_3)$. To do so, we derive a system of linear equations in $E(T_i)$ for $i = 1, 2, 3$. By conditioning on the next state and using the results in (B.1), it follows that

$$E(T_1) = \frac{1}{\alpha_1 + \beta_2} + \frac{\beta_2}{\alpha_1 + \beta_2} E(T_3), \quad E(T_2) = \frac{1}{\alpha_2 + \beta_1} + \frac{\beta_1}{\alpha_2 + \beta_1} E(T_3),$$

$$E(T_3) = \frac{1}{\alpha_1 + \alpha_2} + \frac{\alpha_2}{\alpha_1 + \alpha_2} E(T_1) + \frac{\alpha_1}{\alpha_1 + \alpha_2} E(T_2).$$

These equations are easily solved for $E(T_3)$.

The gamma distribution

A positive random variable X is said to be gamma (α, λ) distributed when it has the probability density

$$f(t) = \frac{\lambda^\alpha t^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda t}, \quad t \geq 0,$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. The symbol $\Gamma(\alpha)$ denotes the *complete gamma function* which is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, dt, \quad \alpha > 0.$$

This function has the property that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$. In particular, $\Gamma(\alpha) = (\alpha - 1)!$ if α is a positive integer. The probability distribution function

$F(t)$ of a gamma (α, λ) distributed random variable X is given by

$$F(t) = \frac{1}{\Gamma(\alpha)} \int_0^{\lambda t} e^{-u} u^{\alpha-1}\, du, \quad t \geq 0.$$

The latter integral is known as the *incomplete gamma function.* The class of gamma distributions is closed in the following sense. If X_1 and X_2 are two independent random variables that are gamma (α_1, λ) and gamma (α_2, λ) distributed, then $X_1 + X_2$ has a gamma $(\alpha_1 + \alpha_2, \lambda)$ distribution (an easy way to prove this is to use Laplace transforms; see Appendix E). In particular, the sum of n independent random variables each having the same gamma (α, λ) distribution is gamma $(n\alpha, \lambda)$ distributed. In queueing applications the gamma distribution is often used to model service-time distributions and in inventory applications to model demand distributions. The numerical evaluation of the gamma distribution function is hardly more difficult than that of the standard normal distribution function. Fast numerical procedures for the computation of the incomplete gamma function are widely available; see for example Press *et al.* (1992).

The mean and the squared coefficient of variation of a gamma (α, λ) distributed random variable X are given by

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad c_X^2 = \frac{1}{\alpha}.$$

This result shows that a unique gamma distribution can be fitted to each positive random variable with given first two moments. To characterize the shape and the failure rate of the gamma density, we distinguish between the cases $c_X^2 < 1$ $(\alpha > 1)$ and $c_X^2 \geq 1$ $(\alpha \leq 1)$. The gamma density is always *unimodal*; that is, the density has only one maximum. For the case $c_X^2 < 1$ the density first increases to the maximum at $t = (\alpha - 1)/\lambda > 0$ and next decreases to zero as $t \to \infty$, whereas for the case $c_X^2 \geq 1$ the density has its maximum at $t = 0$ and thus decreases from $t = 0$ onwards. The failure rate function is increasing from zero to λ if $c_X^2 < 1$ and is decreasing from infinity to zero if $c_X^2 > 1$. The exponential distribution $(c_X^2 = 1)$ has a constant failure rate λ and is a natural boundary between the cases $c_X^2 < 1$ and $c_X^2 > 1$.

The Erlang distribution

The Erlang (E_k) distribution is a special case of the gamma distribution. For a positive integer k, the Erlang (k, λ) distribution is nothing else than the gamma (α, λ) distribution with $\alpha = k$. The probability density and the probability distribution function of an Erlang (k, λ) distributed random variable X are

$$f(t) = \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda t} \quad \text{and} \quad F(t) = 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \quad t \geq 0.$$

The Erlang (k, λ) distribution has a very useful interpretation. An Erlang (k, λ) distributed random variable X can be decomposed as the sum of k independent

random variables each having an exponential distribution with the same mean $1/\lambda$; see also Appendix E. The Erlang probability distribution function can be numerically evaluated without using a general code for the incomplete gamma integral. For fixed t, the Poisson probabilities $p_j(t) = e^{-\lambda t}(\lambda t)^j/j!$ can be recursively calculated from $p_0(t) = e^{-\lambda t}$ and $p_j(t) = (\lambda t/j)p_{j-1}(t)$ for $j = 1, 2, \ldots$. However, exponent underflow may occur in the calculation of $p_0(t)$ when λt is very large. There is a simple trick to avoid the exponent underflow. Define $q_j(t) = \ln\left[p_j(t)\right]$. The recursion scheme $q_0(t) = -\lambda t$ and $q_j(t) = \ln(\lambda t/j) + q_{j-1}(t)$ for $j \geq 1$ offers no numerical difficulties at all. Any desired $p_j(t)$ is calculated from $p_j(t) = \exp[q_j(t)]$ if $q_j(t) \geq -100$ (say) and $p_j(t) = 0$ otherwise. The trick of working with logarithms is one of the most useful tricks to avoid underflow in numerical analysis. Logarithms enable us to reduce the manipulation with extremely large (small) numbers to the manipulation with moderately sized numbers.

The lognormal distribution

A positive random variable X is said to be lognormally distributed when it has the probability density

$$f(t) = \frac{1}{\alpha t\sqrt{2\pi}}\exp\left[-\frac{1}{2}[\ln(t) - \lambda]^2/\alpha^2\right], \quad t > 0,$$

where the shape parameter α is positive and the scale parameter λ may assume each real value. The probability density function $F(t)$ equals

$$F(t) = \Phi\left(\frac{\ln(t) - \lambda}{\alpha}\right), \quad t > 0,$$

where $\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x}\exp(-u^2/2)\,du$ is the standard normal probability distribution function. The mean and the squared coefficient of variation of the lognormal distribution are given by

$$E(X) = \exp\left(\lambda + \frac{1}{2}\alpha^2\right) \quad \text{and} \quad c_X^2 = \exp(\alpha^2) - 1.$$

Thus a unique lognormal distribution can be fitted to each positive random variable with given first two moments. The lognormal density is always unimodal with a maximum at $t = \exp(\lambda - \alpha^2)$. The failure rate function first increases and next decreases to zero as $t \to \infty$ and thus the failure rate is only decreasing in the long-life range.

The Weibull distribution

A positive random variable X is said to be Weibull distributed when it has the probability density

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1}\exp[-(\lambda t)^\alpha], \quad t > 0,$$

with the shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. The corresponding probability distribution function $F(t)$ is given by

$$F(t) = 1 - \exp[-(\lambda t)^{\alpha}], \quad t \geq 0.$$

The mean and the squared coefficient of variation of the Weibull random variable X are

$$E(X) = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \text{and} \quad c_X^2 = \frac{\Gamma(1 + 2/\alpha)}{[\Gamma(1 + 1/\alpha)]^2} - 1.$$

A unique Weibull distribution can be fitted to each positive random variable with given first two moments. For that purpose a non-linear equation in α must be numerically solved. The Weibull density is always unimodal with a maximum at $t = \lambda^{-1}(1 - 1/\alpha)^{1/\alpha}$ if $c_X^2 < 1$ ($\alpha > 1$), and at $t = 0$ if $c_X^2 \geq 1$ ($\alpha \leq 1$). The failure rate function is increasing from 0 to infinity if $c_X^2 < 1$ and is decreasing from infinity to zero if $c_X^2 > 1$.

The gamma and Weibull densities are similar in shape, and for $c_X^2 < 1$ the lognormal density takes on shapes similar to the gamma and Weibull densities. For the case $c_X^2 \geq 1$ the gamma and Weibull densities have their maximum value at $t = 0$; most outcomes tend to be small and very large outcomes occur only occasionally. The lognormal density goes to zero as $t \to 0$ faster than any power of t, and thus the lognormal distribution will typically produce fewer small outcomes than the other two distributions. This explains the popular use of the lognormal distribution in actuarial studies. The differences between the gamma, Weibull and lognormal densities become most significant in their tail behaviour. The densities for large t go down like $\exp[-\lambda t]$, $\exp[-(\lambda t)^{\alpha}]$ and $\exp[-\frac{1}{2}[\ln(t) - \lambda]^2/\alpha^2]$. Thus, for given values of the mean and the coefficient of variation, the lognormal density always has the longest tail. The gamma density has the second longest tail only if $\alpha > 1$; that is, only if its coefficient of variation is less than one. In Figure B.1 we illustrate these facts by drawing the gamma, Weibull and lognormal densities for $c_X^2 = 0.25$, where $E(X)$ is taken to be 1. To conclude this appendix, we discuss several useful generalizations of exponential and Erlangian distributions. In many queueing and inventory applications there is a very substantial (numerical) advantage in using the generalized distributions rather than other distributions.

Generalized Erlangian distributions

An Erlang-k (E_k) distributed random variable can be represented as the sum of k independent exponentially distributed random variables with the same means. A generalized Erlangian distribution is one built out of a random sum of exponentially distributed components. A particularly convenient distribution arises when these components have the same means. In fact, such a distribution can be used to approximate arbitrarily closely any distribution having its mass on the positive

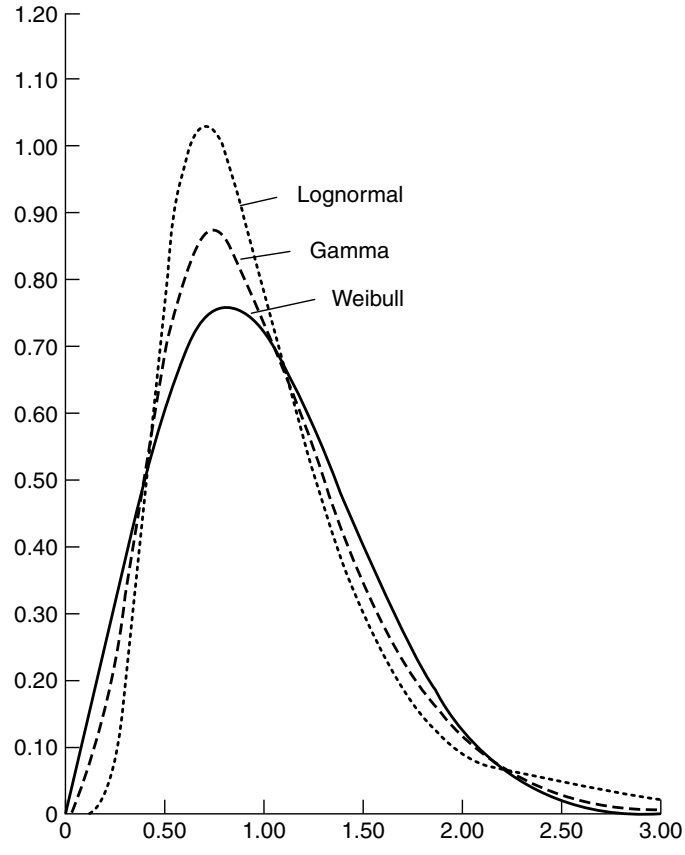


Figure B.1 The gamma, lognormal and Weibull densities

half-axis; see also Section 5.5. We discuss two special cases of mixtures of Erlangian distributions with the same scale parameters. First, we consider the $E_{k-1,k}$ distribution which is defined as a mixture of E_{k-1} and E_k distributions with the same scale parameters. The probability density of an $E_{k-1,k}$ distribution has the following form:

$$f(t) = p\mu^{k-1}\frac{t^{k-2}}{(k-2)!}e^{-\mu t} + (1-p)\mu^k\frac{t^{k-1}}{(k-1)!}e^{-\mu t}, \quad t \geq 0,$$

where $0 \leq p \leq 1$. In other words, a random variable having this density is with respective probabilities p and $1-p$ distributed as the sum of $k-1$ and k independent exponentials with common mean $1/\mu$. By choosing the parameters p and μ as

$$p = \frac{1}{1+c_X^2}\left[kc_X^2 - \sqrt{k(1+c_X^2) - k^2c_X^2}\right] \quad \text{and} \quad \mu = \frac{k-p}{E(X)},$$

the associated $E_{k-1,k}$ distribution fits the first two moments of a positive random variable X provided that

$$\frac{1}{k} \leq c_X^2 \leq \frac{1}{k-1}.$$

We note that only coefficients of variation between 0 and 1 can be achieved by mixtures of the $E_{k-1,k}$ type. Also, it is noteworthy that the $E_{k-1,k}$ density can be shown to have an increasing failure rate.

Next we consider the $E_{1,k}$ distribution, which is defined as a mixture of E_1 and E_k distributions with the same scale parameters. The density of the $E_{1,k}$ distribution has the form

$$f(t) = p\mu e^{-\mu t} + (1-p)\mu^k \frac{t^{k-1}}{(k-1)!} e^{-\mu t}, \quad t \geq 0,$$

where $0 \leq p \leq 1$. By choosing

$$p = \frac{2kc_X^2 + k - 2 - \sqrt{k^2 + 4 - 4kc_X^2}}{2(k-1)(1+c_X^2)} \quad \text{and} \quad \mu = \frac{p + k(1-p)}{E(X)},$$

the associated $E_{1,k}$ distribution fits the first two moments of a positive random variable X provided that

$$\frac{1}{k} \leq c_X^2 \leq \frac{k^2+4}{4k}.$$

Hence the $E_{1,k}$ distribution can also achieve values of c_X^2 with $c_X^2 > 1$.

For use in applications the $E_{k-1,k}$ density is generally better suited than the $E_{1,k}$ density since the $E_{k-1,k}$ density is always unimodal and has a shape similar to the frequently occurring gamma density. The $E_{1,k}$ density may be useful in sensitivity analysis. For both theoretical and practical purposes it is often easier to work with mixtures of Erlangian distributions than with gamma distributions, since mixtures of Erlangian distributions with the same scale parameters allow for the probabilistic interpretation that they represent a random sum of independent exponentials with the same means.

Hyperexponential distribution

A commonly used representation of a positive random variable with a coefficient of variation greater than 1 is a mixture of two exponentials with different means. The distribution of such a mixture is called a *hyperexponential* distribution of order 2, an H_2 distribution. The density of the H_2 distribution has the form

$$f(t) = p_1\mu_1 e^{-\mu_1 t} + p_2\mu_2 e^{-\mu_2 t}, \quad t \geq 0,$$

where $0 \leq p_1,\ p_2 \leq 1$. Note that always $p_1 + p_2 = 1$, since the density $f(t)$ represents a probability mass of 1. In words, a random variable having the H_2 density is distributed with probability p_1 (p_2) as an exponential variable with mean

$1/\mu_1$ $(1/\mu_2)$. The hyperexponential density always has a coefficient of variation of at least 1 and is unimodal with a maximum at $t = 0$. The failure rate function of the hyperexponential distribution is decreasing. The H_2 density has three parameters and is therefore not uniquely determined by its first two moments. For a two-moment fit, the H_2 density with *balanced means* is often used; that is, the normalization $p_1/\mu_1 = p_2/\mu_2$ is used. The parameters of the H_2 density having balanced means and fitting the first two moments of a positive random variable X with $c_X^2 \geq 1$ are

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\right), \quad p_2 = 1 - p_1, \quad \mu_1 = \frac{2p_1}{E(X)}, \quad \mu_2 = \frac{2p_2}{E(X)}.$$

In the context of a Coxian-2 distribution we give below another normalization we believe to be a more natural one. A three-moment fit by an H_2 density is not always possible, but it is unique whenever it exists. An H_2 density can only be fitted to the first three moments m_1, m_2 and m_3 of a positive random variable X with $c_X^2 > 1$ when the requirement $m_1 m_3 \geq \frac{3}{2}m_2^2$ is satisfied; see Whitt (1982). If $m_1 m_3 = \frac{3}{2}m_2^2$ then the H_2 fit is the exponential density, otherwise the parameters of the three-moment fit are given by

$$\mu_{1,2} = \frac{1}{2}\left\{a_1 \pm \sqrt{a_1^2 - 4a_2}\right\}, \quad p_1 = \frac{\mu_1(1 - \mu_2 m_1)}{\mu_1 - \mu_2}, \quad p_2 = 1 - p_1,$$

where $a_2 = (6m_1^2 - 3m_2)/(\frac{3}{2}m_2^2 - m_1 m_3)$ and $a_1 = (1 + \frac{1}{2}m_2 a_2)/m_1$. The requirement $m_1 m_3 \geq \frac{3}{2}m_2^2$ holds for both a gamma distributed and a lognormal distributed random variable X with $c_X^2 > 1$.

Coxian-2 distribution

The hyperexponential density requires that the weights p_1 and p_2 are non-negative. However, in order that $p_1\mu_1\exp(-\mu_1 t) + p_2\mu_2\exp(-\mu_2 t)$ represents a probability density, it is not necessary to require that p_1 and p_2 are both non-negative. The class of H_2 distributions can be shown to be a subclass of the class of so-called *Coxian-2* (C_2) distributions. A random variable X is said to be Coxian-2 distributed if X can be represented as

$$X = \begin{cases} X_1 + X_2 & \text{with probability } b, \\ X_1 & \text{with probability } 1 - b, \end{cases}$$

where X_1 and X_2 are independent random variables having exponential distributions with respective means $1/\mu_1$ and $1/\mu_2$. In words, the lifetime X first goes through an exponential phase X_1 and then through a second exponential phase X_2 with probability b or it goes out with probability $1 - b$; see Figure B.2. It can be assumed without loss of generality that $\mu_1 \geq \mu_2$.

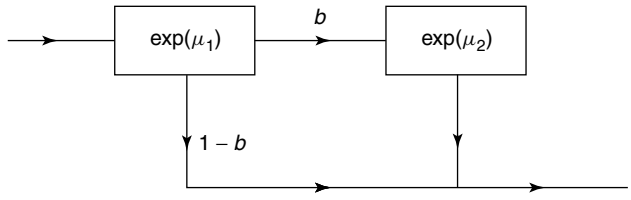


Figure B.2 The Coxian distribution with two phases

A Coxian-2 distribution having parameters (b, μ_1, μ_2) with $\mu_1 < \mu_2$ can be shown to have the same probability density as the Coxian-2 distribution having parameters (b^*, μ_1^*, μ_2^*) with $\mu_1^* = \mu_2$, $\mu_2^* = \mu_1$ and $b^* = 1 - (1-b)\mu_1/\mu_2$. Assuming that $\mu_1 \geq \mu_2$, the Coxian-2 distributed random variable X has the density

$$f(t) = \begin{cases} p_1\mu_1 e^{-\mu_1 t} + (1-p_1)\mu_2 e^{-\mu_2 t} & \text{if } \mu_1 \neq \mu_2, \\ p_1\mu_1 e^{-\mu_1 t} + (1-p_1)\mu_1^2 t e^{-\mu_1 t} & \text{if } \mu_1 = \mu_2, \end{cases}$$

where $p_1 = 1 - b\mu_1/(\mu_1 - \mu_2)$ if $\mu_1 \neq \mu_2$ and $p_1 = 1 - b$ if $\mu_1 = \mu_2$. Thus the class of H_2 densities is contained in the class of Coxian-2 densities. Note that the H_2 distribution allows for two different but equivalent probabilistic interpretations. The H_2 distribution can be interpreted in terms of exponential phases in parallel and in terms of exponential phases in series.

The density of a Coxian-2 distributed random variable X always has a unimodal shape. Moreover, it holds that

$$c_X^2 \geq \frac{1}{2},$$

where $c_X^2 \geq 1$ only if the density has the form $p_1\mu_1\exp(-\mu_1 t) + p_2\mu_2\exp(-\mu_2 t)$ for non-negative p_1 and p_2. The Coxian-2 density has three parameters (b, μ_1, μ_2). Hence an infinite number of Coxian-2 densities can in principle be used for a two-moment fit to a random variable X with $c_X^2 > \frac{1}{2}$ (the E_2 density is the only possible choice when $c_X^2 = \frac{1}{2}$). A particularly useful choice for a two-moment match is the Coxian-2 density with parameters

$$\mu_1 = \frac{2}{E(X)}\left(1 + \sqrt{\frac{c_X^2 - \frac{1}{2}}{c_X^2 + 1}}\right), \quad \mu_2 = \frac{4}{E(X)} - \mu_1, \quad b = \frac{\mu_2}{\mu_1}\{\mu_1 E(X) - 1\}.$$

This particular Coxian-2 density has the remarkable property that its third moment is the same as that of the gamma density with mean $E(X)$ and squared coefficient of variation c_X^2. The unique Coxian-2 density having this property will therefore be called the Coxian-2 density with *gamma normalization.* This normalization is a natural one in many applications.

A two-stage process with negative probabilities

For $c_X^2 < \frac{1}{2}$ it is not possible to fit a Coxian-2 distribution to the first two moments of the positive random variable X. A fit using an $E_{k,k-1}$ distribution requires many stages when c_X^2 is close to zero and thus might be unattractive in (queueing) applications. A remarkable alternative involving two exponential stages was proposed in Nojo and Watanabe (1987). The positive random variable X is approximated through a two-stage process. The process starts in stage 1. It stays in stage 1 for an exponentially distributed time with mean $1/\gamma$. Upon completion of the sojourn time in stage 1, the process expires with probability p_1 and moves to stage 2 with probability $1 - p_1$. The sojourn time in stage 2 is also exponentially distributed with the same mean $1/\gamma$. Upon completion of the sojourn time in stage 2, the process expires with probability p_2 and returns to stage 1 with probability $1 - p_2$. In stage 1 the process starts anew. The idea is to approximate the random variable X by the time until the process expires. Using results from Appendix E, it is not difficult to verify that the Laplace transform of this lifetime is given by

$$f^*(s) = \frac{\gamma p_1 s + \gamma^2 (p_1 + p_2 - p_1 p_2)}{s^2 + 2\gamma s + \gamma^2 (p_1 + p_2 - p_1 p_2)}.$$

The moments of the lifetime are directly obtained from the Laplace transform $f^*(s)$; see (E.2) in AppendixE. If $c_X^2 < \frac{1}{2}$ and the first three moments m_1, m_2 and m_3 satisfy $m_1 m_3 < \frac{3}{2} m_2^2$, it is nearly always possible to match the first three moments of $f^*(s)$ with the first three moments of X by allowing for *negative* values of p_1 and p_2 but requiring that $\gamma > 0$. This is particularly true when $c_X^2 = 0$. A surprising finding is that in many (queueing) applications excellent approximations are obtained by replacing the random variable X through the two-stage process and treating p_1 and p_2 as if they were probabilities.

APPENDIX C. GENERATING FUNCTIONS

The generating function (or z-transform) of a discrete probability distribution $\{p_k, k = 0, 1, \dots\}$ is defined by

$$P(z) = \sum_{k=0}^{\infty} p_k z^k, \quad |z| \leq 1.$$

The variable z is usually taken as a real-valued variable, but in certain applications it may be convenient to treat z as a complex-valued variable. It is easily verified that the probability distribution $\{p_k, k = 0, 1, \dots\}$ can be recovered analytically from the compressed function $P(z)$ by

$$p_k = \frac{1}{k!} \frac{d^k P(z)}{dz^k}\bigg|_{z=0}, \quad k = 0, 1, \dots . \tag{C.1}$$

The result (C.1) shows that a discrete probability distribution is uniquely determined by its generating function. Also, the moments of the probability distribution $\{p_k\}$

are readily obtained from $P(z)$. For example, the first two moments are obtained from the relations

$$\sum_{k=0}^{\infty} kp_k = P'(1) \quad \text{and} \quad \sum_{k=1}^{\infty} k(k-1)p_k = P''(1).$$

In general the relation (C.1) is only of theoretical value. It is often possible to obtain an explicit expression for $P(z)$ when the probabilities p_k are unknown (e.g. from a difference equation for the p_k). In Appendix D we discuss the discrete Fast Fourier Transform algorithm to recover the p_k numerically when an explicit expression for $P(z)$ is available. Usually it is not possible to analytically recover the p_k from (C.1).

A useful probabilistic interpretation can be given to $P(z)$. If the random variable N is distributed according to $\{p_k\}$, then

$$P(z) = E(z^N). \tag{C.2}$$

A direct consequence of this relation is that the generating function of the convolution of two discrete probability distributions is the product of the generating functions of these two probability distributions. More specifically, suppose that the random variable $N = X + Y$, where X and Y are two independent discrete random variables with respective probability distributions $\{a_k, k = 0, 1, \dots\}$ and $\{b_k, k = 0, 1, \dots\}$. Let $p_k = P\{N = k\}$, $k = 0, 1, \dots$. Then the generating function $P(z)$ of the distribution $\{p_k\}$ is given by

$$P(z) = A(z)B(z), \tag{C.3}$$

where $A(z)$ and $B(z)$ are the generating functions of the probability distributions $\{a_k\}$ and $\{b_k\}$. This follows from $E(z^{X+Y}) = E(z^X)E(z^Y)$. In practice it is usually faster to compute the p_j by applying the discrete Fast Fourier Transform method rather than using the convolution formula $p_j = \sum_{k=0}^{j} a_{j-k}b_k$ for $j \geq 0$.

Example C.1 The coupon-collecting problem

Suppose there are r different types of coupons and each time we obtain a coupon it is equally likely to be any one of the r types. How do we compute the probability distribution of the number of coupons we need to collect for a complete set of coupons? Denote this number by the random variable X. The random variable X can be written as

$$X = Y_1 + \cdots + Y_r,$$

where Y_i is the number of additional coupons that need to be collected to increase the number of different coupons in the collection from $i - 1$ to i. The random variables $Y_1, \dots, Y_r$ are independent of each other and Y_i has a geometric distribution

with parameter $\alpha_i = 1 - (i-1)/r$. The generating function of Y_i is

$$P_i(z) = \sum_{k=1}^{\infty} \alpha_i (1-\alpha_i)^{k-1} z^k = \frac{\alpha_i z}{1-(1-\alpha_i)z}.$$

Noting that $\alpha_1 = 1$ and letting $\beta_i = 1 - \alpha_i = (i-1)/r$, it follows from (C.3) that the generating function $P(z) = \sum_{k=1}^{\infty} P\{X = k\} z^k$ is given by

$$P(z) = P_1(z) \cdots P_r(z) = \frac{\alpha_2 \cdots \alpha_r z^r}{(1-\beta_2 z) \cdots (1-\beta_r z)}.$$

Using partial-fraction expansion, we next find

$$P(z) = \alpha_2 \cdots \alpha_r z^r \left[\frac{\gamma_2}{1-\beta_2 z} + \cdots + \frac{\gamma_r}{1-\beta_r z} \right],$$

where the residue γ_i is given by

$$\gamma_i = \prod_{\substack{\ell=2 \\ \ell \neq i}}^{r} \left(\frac{1}{1-\beta_\ell/\beta_i} \right), \quad i = 2, \ldots, r.$$

Noting that $\sum_{j=1}^{\infty} (1-p) p^{j-1} z^j = (1-p)z/(1-(1-p)z)$, we can invert the final expression for $P(z)$ to obtain

$$P\{X = k\} = \alpha_2 \cdots \alpha_r \left[\gamma_2 \beta_2^{k-r} + \cdots + \gamma_r \beta_r^{k-r} \right], \quad k \geq r. \tag{C.4}$$

Example C.2 Success runs

Another illustration of the usefulness of the generating function approach is the analysis of success runs in independent Bernoulli trials. How do we compute the probability that in n independent Bernoulli trials with success probability p there is *some* sequence of s consecutive successes? For fixed s, denote this probability by P_n for $n \geq 0$. The probability P_n can be written as $P_n = \sum_{j=0}^{n} p_j$ for $n = 0, 1, \ldots$, where the probability p_j is defined as

$$p_j = \text{the probability that for the first time a sequence of } s \text{ consecutive successes occurs at the } j\text{th trial.}$$

Note that $\{p_j,\ j = 0, 1, \ldots\}$ is a probability distribution with $\sum_{j=0}^{\infty} p_j = 1$. Obviously $p_j = 0$ for $j < s$ and $p_s = p^s$. For $j > s$, we have the recursion

$$p_j = \sum_{k=1}^{s} p^{k-1}(1-p) p_{j-k}, \quad j = s+1, s+2, \ldots .$$

To prove this, fix $j > s$ and denote by A the event that a sequence of s consecutive successes occurs for the first time at the jth trial. The event A can only occur

if one of the mutually exclusive events $B_1, \ldots, B_s$ occurs, where B_k is the event that each of the first $k-1$ trials have success as outcome but the kth trial does not. Noting that $P(A) = p_j$, $P(B_k) = p^{k-1}(1-p)$ and $P(A \mid B_k) = p_{j-k}$, the recursion follows by applying the law of conditional probabilities. As an alternative to the recursion scheme, the probabilities p_j can also be numerically obtained by numerical inversion of the generating function. Multiplying both sides of the above recursion for p_j by z^j and summing over j, it follows that the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$ satisfies

$$\begin{aligned} P(z) &= p_s z^s + \sum_{j=s+1}^{\infty} z^j \sum_{k=1}^{s} p^{k-1}(1-p) p_{j-k} \\ &= p_s z^s + \sum_{k=1}^{s} z^k p^{k-1}(1-p) \sum_{j=s+1}^{\infty} p_{j-k} z^{j-k} \\ &= p_s z^s + P(z) \sum_{k=1}^{s} z^k p^{k-1}(1-p). \end{aligned}$$

This gives

$$P(z) = \frac{p^s z^s}{1 - \sum_{k=1}^{s} p^{k-1}(1-p) z^k}. \tag{C.5}$$

Hence an *explicit* expression has been obtained for the generating function $P(z)$ of the *unknown* probabilities $\{p_j\}$. Using this expression the unknown probabilities p_j can also be numerically obtained by applying the discrete Fast Fourier Transform method from Appendix D. A simple but extremely useful method to compute p_j for large j is to use an asymptotic expansion. This approach will be discussed below in a general setting. To do so, some basic concepts from complex function theory are needed such as the concept of an analytic function. In a nutshell, a function on a domain in the complex plane is called *analytic* when the function is differentiable infinitely often on that domain. A fundamental theorem from complex function theory states that a function $f(z)$ is analytic in the complex region $|z| < R$ if and only if $f(z)$ allows for the power series representation $f(z) = \sum_{n=0}^{\infty} f_n z^n$ for $|z| < R$.

Asymptotic expansion

Suppose that the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$ of an (unknown) probability distribution $\{p_j,\ j = 0, 1, \ldots\}$ has the form

$$P(z) = \frac{N(z)}{D(z)}. \tag{C.6}$$

The generating function $P(z)$ is defined for $|z| \leq 1$, but assume that $N(z)$ and $D(z)$ are analytic functions whose domains of definition can be extended to a

region $|z| < R$ in the complex plane for some $R > 1$. It is essential that the radius R is larger than 1. Note that the generating function (C.5) is indeed of the form (C.6), where the numerator and denominator are analytic functions on the whole complex plane ($R = \infty$). It is no restriction to assume that $N(z)$ and $D(z)$ have no common zeros; otherwise, cancel out common zeros. Let us further assume that the following regularity conditions are satisfied:

C1 The equation $D(z) = 0$ has a *real* root z_0 on the interval $(1,R)$.

C2 The function $D(z)$ has no zeros in the domain $1 < |z| < z_0$ of the complex plane.

C3 The zero $z = z_0$ of $D(z)$ is of multiplicity 1 and is the only zero of $D(z)$ on the circle $|z| = z_0$.

The following theorem is of utmost importance. The insightful proof of the theorem is included for completeness. Recall that $f(x) \sim g(x)$ as $x \to \infty$ means that $f(x)/g(x) \to 1$ as $x \to \infty$.

Theorem C.1 *Under the conditions C1 to C3,*

$$p_j \sim \gamma_0 z_0^{-j} \quad as\ j \to \infty, \tag{C.7}$$

where the constant γ_0 is given by

$$\gamma_0 = -\frac{1}{z_0}\frac{N(z_0)}{D'(z_0)}. \tag{C.8}$$

Here $D'(z_0)$ denotes the derivative of $D(x)$ at $x = z_0$.

Proof We first mention the following basic facts from complex function theory. The most important fact is that a function $f(z)$ is analytic at a point $z = a$ if and only if $f(z)$ can be expanded in a power series $f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n$ in $|z-a| < \rho$ for some $\rho > 0$. The coefficient a_n of the Taylor series is the nth derivative of $f(z)$ at $z = a$ divided by $n!$. The analytic function $f(z)$ is said to have a zero of multiplicity k in $z = a$ if $a_0 = \cdots = a_{k-1} = 0$ and $a_k \neq 0$. Another basic result is the following. The Taylor series $\sum_{n=0}^{\infty} a_n(z-a)^n$ of a function $f(z)$ at the point $z = a$ coincides with the function $f(z)$ in the interior of the largest circle whose interior lies wholly within the domain on which $f(z)$ is analytic.

The proof of (C.7) now proceeds as follows. The conditions C1 to C3 imply that there is a circle around $z = 0$ with radius R_0 larger than z_0 such that $P(z)$ is analytic in $|z| < R_0$ except for the isolated point $z = z_0$. Since $D(z)$ has a zero of multiplicity 1 at $z = z_0$, it follows from the Taylor series that $D(z) = (z - z_0)\phi(z)$ in $|z| < R_0$, where $\phi(z)$ is an analytic function with $\phi(z_0) \neq 0$. Thus we can write $P(z)$ as $P(z) = H(z)/(z - z_0)$ for some analytic function $H(z)$ in $|z| < R_0$ with $H(z_0) \neq 0$. Using a Taylor expansion $H(z) = H(z_0) + (z - z_0)U(z)$, we next find

that $P(z)$ can be represented as

$$P(z) = \frac{r_0}{z - z_0} + U(z) \tag{C.9}$$

in $|z| < R_0$, $z \neq z_0$. Here $U(z)$ is an analytic function in the domain $|z| < R_0$ and the residue $r_0 = H(z_0)$ is given by

$$r_0 = \lim_{z \to z_0} (z - z_0) P(z) = N(z_0)/D'(z_0).$$

The remainder of the proof is simple. Since $U(z)$ is analytic for $|z| < R_0$ we have the power series representation $U(z) = \sum_{j=0}^{\infty} u_j z^j$ for $|z| < R_0$. Let R_1 be any number with $z_0 < R_1 < R_0$. Then, for some constant b, $|u_j| \leq b R_1^{-j}$ for all $j \geq 0$. This follows from the fact that the series $\sum_{j=0}^{\infty} u_j z^j$ is convergent for $z = R_1$. Using the power series representation of $U(z)$ and the fact that the power series representation $P(z) = \sum_{j=0}^{\infty} p_j z^j$ extends to $|z| < z_0$, it follows from (C.9) that

$$\sum_{j=0}^{\infty} p_j z^j = \frac{-r_0}{z_0} \sum_{j=0}^{\infty} (z/z_0)^j + \sum_{j=0}^{\infty} u_j z^j, \quad |z| < z_0.$$

Equating coefficients yields

$$p_j = -r_0 z_0^{-j-1} + u_j, \quad j \geq 0.$$

Since $|u_j| \leq b R_1^{-j}$ for some constant b and $R_1 > z_0$, the coefficient u_j tends to zero faster than z_0^{-j}. Hence we can conclude the asymptotic expansion (C.7).

It is noted that Theorem C.1 does not require that $\{p_j\}$ is a probability distribution. The theorem applies to any sequence $\{p_j,\ j = 0, 1, \ldots\}$ with $p_j \geq 0$ for all j and $\sum_{j=0}^{\infty} p_j < \infty$. The asymptotic expansion (C.7) is very useful for both theoretical and computational purposes. It appears that in many applications the asymptotic expansion for p_j can be used for relatively small values of j. To illustrate this, consider the generating function (C.5) for the problem of success runs. This generating function $P(z)$ is the ratio of the two analytic functions $N(z) = p^s z^s$ and $D(z) = 1 - \sum_{k=1}^{s} p^{k-1}(1-p) z^k$ whose domains of definition can be extended to the whole complex plane ($R = \infty$). It is readily verified that the equation

$$1 - \sum_{k=1}^{s} p^{k-1}(1-p) x^k = 0$$

has a unique root z_0 on the interval $(1, \infty)$. Hence condition C1 is satisfied. The verification of the technical conditions C2 and C3 is omitted and is left to the interested reader. The unique root z_0 of the above equation must be numerically

Table C.1 The exact and approximate values for Q_n

	$s = 2$			$s = 5$	
n	exact	approximate	n	exact	approximate
5	0.50000	0.50156	15	0.831543	0.831541
10	0.173828	0.173824	50	0.4558865475	0.4558865475
15	0.06024170	0.06024171	100	0.1931794513	0.1931794513
25	0.0072355866	0.0072355866	200	0.0346871989	0.0346871989

calculated. A safe and fast method to compute z_0 is the bisection method. Once z_0 is computed, we can approximately calculate p_j from

$$p_j \approx \frac{p(pz_0)^s}{(1-p)\sum_{k=1}^{s} k(pz_0)^k} z_0^{-j} \quad \text{for } j \text{ large enough.}$$

Denoting by $Q_n = \sum_{j=n}^{\infty} p_j$ the probability that it takes n or more Bernoulli trials to obtain a sequence of s consecutive successes, we give in Table C.1 the exact and approximate values of Q_n for several values of n. We take $p = 0.5$ and $s = 2$ and $s = 5$. The numerical results in Table C.1 confirm the finding that the asymptotic expansion (C.7) is remarkably accurate and already applies for relatively small values of j. This finding is very important for practical purposes.

APPENDIX D. THE DISCRETE FAST FOURIER TRANSFORM

The discrete Fast Fourier Transform (FFT) method is a very powerful method to recover numerically the values of unknown probabilities p_k, $k = 0, 1, \ldots$ when an explicit expression is available for the generating function $P(z) = \sum_{k=0}^{\infty} p_k z^k$. The FFT method has many other applications. Another applied probability problem for which the discrete FFT method may be very useful is the calculation of the convolution of two or more discrete probability distributions. The discrete FFT method represents a breakthrough in numerical analysis.

Before stating the discrete FFT method for the numerical inversion of a generating function, here are some basic facts from discrete Fourier analysis. The discrete Fourier transform takes n numbers $f_0, \ldots, f_{n-1}$ into n coefficients $c_0, \ldots, c_{n-1}$ such that there is a one-to-one correspondence between $\{f_k\}$ and $\{c_k\}$. The f_k are real or complex numbers and the c_k are complex numbers. A finite Fourier series

$$\sum_{k=0}^{n-1} c_k e^{ikx}$$

is sought that agrees with f at n equally spaced points $x_l = 2\pi\ell/n$ between 0 and 2π. More specifically, we look for complex numbers $c_0, \ldots, c_{n-1}$ such that

$$\sum_{k=0}^{n-1} c_k\, e^{ik(2\pi\ell/n)} = f_\ell, \quad \ell = 0, \ldots, n-1. \tag{D.1}$$

It is convenient to write these linear equations in matrix notation as

$$Fc = f.$$

Here F is a complex-valued matrix whose (ℓ, k)th element $(F)_{\ell k}$ is given by

$$(F)_{\ell k} = w^{k\ell}, \quad \ell, k = 0, \ldots, n-1,$$

where the complex number w is defined by

$$w = e^{2\pi i/n}.$$

Let $\overline{F}$ be the matrix whose elements are the complex conjugates of the elements of the matrix F. The matrix F has the nice property that

$$\overline{F}F = F\overline{F} = nI, \tag{D.2}$$

where I is the identity matrix (the column vectors of the symmetric matrix F form an orthogonal system). To verify this, let $\overline{w} = e^{-2\pi i/n}$ denote the complex conjugate of w. The inproduct of the rth row of $\overline{F}$ and the sth column of F is given by

$$\gamma_{rs} = \overline{w}^0 w^0 + \overline{w}^r w^s + \overline{w}^{2r} w^{2s} + \cdots + \overline{w}^{(n-1)r} w^{(n-1)s}.$$

For $r = s$ each term equals $e^0 = 1$ and so the sum γ_{rs} is n. For $r \neq s$ the sum γ_{rs} can be written as $1 + \alpha + \cdots + \alpha^{n-1} = (1-\alpha^n)/(1-\alpha)$ with $\alpha = \overline{w}^r w^s (\neq 1)$. Since $w^n = e^{2\pi i} = 1$ and $\overline{w}^n = e^{-2\pi i} = 1$, we have $\alpha^n = 1$ and so $\gamma_{rs} = 0$ for $r \neq s$. This gives (D.2). By (D.2), we have $F^{-1} = (1/n)\overline{F}$. It now follows that the vector c of Fourier coefficients is given by $c = (1/n)\overline{F}f$. Componentwise, we have

$$c_k = \frac{1}{n}\sum_{\ell=0}^{n-1} f_\ell\, e^{-2\pi i\ell k/n}, \quad k = 0, \ldots, n-1. \tag{D.3}$$

This inversion formula parallels the formula $c_k = (2\pi)^{-1}\int_{-\pi}^{\pi} f(x)\, e^{-ikx}\, dx$ in continuous Fourier analysis. Notice that (D.3) inherits the structure of (D.1).

In many applications, however, we proceed in reverse order: we know the Fourier coefficients c_k and wish to calculate the original coefficients f_j. By the formula (D.1) we can transform c back into f. The matrix multiplications in (D.1) would normally require n^2 multiplications. However, the discrete FFT method performs the multiplications in an extremely fast and ingenious way that requires only $n\log_2(n)$ multiplications instead of n^2. The key to the method is the simple observation that the discrete Fourier transform of length n (n even) can be written

as the sum of two discrete Fourier transforms, each of length $n/2$. Suppose we know the c_k and wish to compute the f_ℓ from (D.1). It holds that

$$\sum_{k=0}^{n-1} c_k\, e^{2\pi i k\ell/n} = \sum_{k=0}^{\frac{1}{2}n-1} c_{2k}\, e^{2\pi i \ell(2k)/n} + \sum_{k=0}^{\frac{1}{2}n-1} c_{2k+1}\, e^{2\pi i \ell(2k+1)/n}$$

$$= \sum_{k=0}^{\frac{1}{2}n-1} c_{2k}\, e^{2\pi i k\ell/(n/2)} + w^\ell \sum_{k=0}^{\frac{1}{2}n-1} c_{2k+1}\, e^{2\pi i k\ell/(n/2)}. \quad \text{(D.4)}$$

The discrete Fourier transform of length n can thus be written as the sum of two discrete Fourier transforms each of length $n/2$. This beautiful trick can be applied recursively. For the implementation of the recursive discrete FFT procedure it is convenient to choose

$$n = 2^m$$

for some positive m (if necessary, zeros can be added to the sequence $f_0, \ldots, f_{n-1}$ in order to achieve that $n = 2^m$ for some m). The discrete FFT method is numerically very stable (it is a fast and accurate method even for values of n with an order of magnitude of a hundred thousand). The discrete FFT method that calculates the original coefficients f_j from the Fourier coefficients c_k is usually called the *inverse discrete FFT method*. Ready-to-use codes for the discrete FFT method are widely available. The discrete FFT method is a basic tool that should be part of the toolbox of any applied probabilist. It is noted that the discrete FFT method can be extended to a complex function defined over a multidimensional grid.

Numerical inversion of the generating function

Suppose an *explicit* expression is available for the generating function

$$P(z) = \sum_{\ell=0}^{\infty} p_\ell z^\ell, \quad |z| \le 1.$$

How do we obtain the unknown probabilities p_ℓ? Choose an integer $n = 2^m$ such that

$$\sum_{j=n}^{\infty} p_j \le \epsilon$$

for some prespecified accuracy number ϵ, say $\epsilon = 10^{-12}$ (often one can find a known distribution $\{a_j\}$ such that $\sum_{j=k}^{\infty} p_j \le \sum_{j=k}^{\infty} a_j$ for all k; otherwise, the truncation integer n has to be found by trial and error). Then calculate the complex

numbers

$$c_k = \frac{1}{n} P(e^{-2\pi ik/n}), \quad k = 0, \ldots, n-1 \tag{D.5}$$

from the explicit expression for $P(z)$. Note that each of the points $z_k = e^{-2\pi ik/n}$, $k = 0, \ldots, n-1$ satisfies $z^n = 1$ and thus lies on the unit circle $|z| = 1$. By the power series representation of $P(z)$ and the choice of the integer n, we have

$$c_k \approx \frac{1}{n} \sum_{\ell=0}^{n-1} p_\ell \, e^{-2\pi i \ell k/n}, \quad k = 0, \ldots, n-1.$$

This relation is of the same form as (D.3). Thus the unknown probabilities p_ℓ can be calculated by applying the inverse discrete FFT method to the known vector $(c_0, \ldots, c_{n-1})$.

Example D.1 The $M/D/1$ queue

Consider the $M/D/1$ queue with deterministic services. In Section 2.5 it was shown that the generating function of the limiting distribution $\{p_k\}$ of the number of customers present is given by

$$P(z) = \frac{(1-\lambda D)(1-z)}{1 - ze^{\lambda D(1-z)}}, \quad |z| \le 1,$$

where λ is the arrival rate of customers and D is the fixed service time of a customer with $\lambda D < 1$. Hence the state probabilities $\{p_k\}$ can be calculated by applying the discrete FFT method. In the specific problem of the $M/D/1$ queue, the explicit expression for the generating function $P(z)$ is of the form $Q(z)/R(z)$. In such a situation one should verify whether or not $R(z)$ has zeros on the unit circle $|z| = 1$ (each zero of $R(z)$ on the unit circle must also be a zero of $Q(z)$ since $P(z) = \sum_{k=0}^{\infty} p_k z^k$ is analytic for $|z| \le 1$). If a point $z_k = e^{-2\pi ik/n}$ is a zero of $R(z)$, the corresponding Fourier coefficient c_k cannot be calculated directly from (D.5) but can be found by applying L'Hospital's rule to $Q(z)/R(z)$ at the point $z = z_k$ (often $z_0 = 1$ is a zero as is the case in the $M/D/1$ problem).

APPENDIX E. LAPLACE TRANSFORM THEORY

This appendix gives a brief outline of some results from Laplace transform theory that are useful in applied probability problems. Suppose that $f(x)$ is a continuous real-valued function in $x \ge 0$ such that $|f(x)| \le Ae^{Bx}$, $x \ge 0$, for some constants A and B. The *Laplace transform* of $f(x)$ is defined by the integral

$$f^*(s) = \int_0^\infty e^{-sx} f(x)\, dx$$

as a function of the complex variable s with $\mathrm{Re}(s) > B$. The integral always exists when $\mathrm{Re}(s) > B$. If $f(x)$ is the probability density of a random variable X, the Laplace transform $f^*(s)$ is defined for all s with $\mathrm{Re}(s) > 0$ and can be interpreted as

$$f^*(s) = E(e^{-sX}). \tag{E.1}$$

Moreover, we then have

$$E(X^k) = (-1)^k \lim_{s \to 0} \frac{d^k f^*(s)}{ds^k}, \quad k = 1, 2, \ldots . \tag{E.2}$$

The results (a) to (c) below can easily be verified from the definition of Laplace transform. In the statements it is assumed that the various integrals exist.

(a) If the function $f(x) = ag(x) + bh(x)$ is a linear combination of the functions $g(x)$ and $h(x)$ with Laplace transforms $g^*(s)$ and $h^*(s)$, then

$$f^*(s) = ag^*(s) + bh^*(s). \tag{E.3}$$

(b) If $F(x) = \int_0^x f(y)\,dy$, then

$$\int_0^\infty e^{-sx} F(x)\,dx = \frac{f^*(s)}{s}. \tag{E.4}$$

If $f(x)$ has a continuous derivative $f'(x)$ then

$$\int_0^\infty e^{-sx} f'(x)\,dx = sf^*(s) - f(0). \tag{E.5}$$

(c) If the function $f(x)$ is given by the convolution

$$f(x) = \int_0^x g(x - y)h(y)\,dy$$

of two functions $g(x)$ and $h(x)$ with Laplace transforms $g^*(s)$ and $h^*(s)$, then

$$f^*(s) = g^*(s)h^*(s). \tag{E.6}$$

In addition to these results, we mention without proof the following useful Abelian theorem. If $\int_0^\infty e^{-sx} f(x)\,dx$ is convergent for $\mathrm{Re}(s) > 0$ and $\lim_{x \to \infty} f(x)$ exists, then

$$\lim_{x \to \infty} f(x) = \lim_{s \to 0} s \int_0^\infty e^{-sx} f(x)\,dx. \tag{E.7}$$

In applied probability problems one often encounters the situation of a non-negative random variable X that has a positive mass at $x = 0$ and a density on $(0, \infty)$. Then

$$\int_0^\infty e^{-sx} P\{X > x\}\,dx = \frac{1 - E(e^{-sX})}{s}. \tag{E.8}$$

Using that $E(e^{-sX}) = P\{X = 0\} + \int_0^\infty e^{-sx}\pi(x)\,dx$ and $P\{X > x\} = 1 - P\{X = 0\} - \int_0^x \pi(y)\,dy$ with $\pi(x)$ denoting the derivative of $P\{X \le x\}$ for $x > 0$, the relation (E.8) follows directly from (E.3) and (E.4). Of course the result (E.8) also holds when X has a zero mass at $x = 0$.

In specific applications requiring the determination of some unknown function $f(x)$ it is often possible to obtain the Laplace transform $f^*(s)$ of $f(x)$. A very useful result is that a continuous function $f(x)$ is uniquely determined by its Laplace transform $f^*(s)$. In principle the function $f(x)$ can be obtained by inversion of its Laplace transform. Extensive tables are available for the inverse of basic forms of $f^*(s)$; see for example Abramowitz and Stegun (1965). An inversion formula that is sometimes helpful in applications is the Heaviside formula. Suppose that

$$f^*(s) = \frac{P(s)}{Q(s)},$$

where $P(s)$ and $Q(s)$ are polynomials in s such that the degree of $P(s)$ is smaller than that of $Q(s)$. It is no restriction to assume that $P(s)$ and $Q(s)$ have no zeros in common. Let $s_1, \ldots, s_k$ be the distinct zeros of $Q(s)$ in the complex plane. For ease of presentation, assume that each root s_j is simple (i.e. has multiplicity 1). Then it is known from algebra that $P(s)/Q(s)$ admits the partial fraction expansion

$$\frac{P(s)}{Q(s)} = \frac{r_1}{s - s_1} + \frac{r_2}{s - s_2} + \cdots + \frac{r_k}{s - s_k},$$

where $r_j = \lim_{s \to s_j}(s - s_j)P(s)/Q(s)$ and so $r_j = P(s_j)/Q'(s_j)$, $1 \le j \le k$. The inverse of the Laplace transform $f^*(s) = P(s)/Q(s)$ is now given by

$$f(x) = \sum_{j=1}^{k} \frac{P(s_j)}{Q'(s_j)} e^{s_j x}, \tag{E.9}$$

as can be verified by taking the Laplace transform of both sides of this equation. This result is readily extended to the case in which some of the roots of $Q(s) = 0$ are not simple. For example, the inverse of the Laplace transform

$$f^*(s) = \frac{P(s)}{(s - a)^m},$$

where $P(s)$ is a polynomial of degree lower than m, is given by

$$f(x) = e^{ax} \sum_{j=1}^{m} \frac{P^{(m-j)}(a)x^{j-1}}{(m - j)!(j - 1)!}. \tag{E.10}$$

Here $P^{(n)}(a)$ denotes the nth derivative of $P(x)$ at $x = a$ with $P^{(0)}(a) = P(a)$.

It is usually not possible to give an explicit expression for the inverse of a given Laplace transform. In those situations the unknown function $f(x)$ may be

obtained by numerical inversion of its Laplace transform $f^*(s)$. Numerical inversion methods that perform well for probability functions $f(x)$ are discussed in Appendix F.

Example E.1 The Erlang distribution

Suppose that $X_1, \ldots, X_n$ are independent random variables having a common exponential distribution with mean $1/\mu$. Then $X_1 + \cdots + X_n$ has the probability density

$$\frac{\mu^n x^{n-1} e^{-\mu x}}{(n-1)!}, \quad x \geq 0,$$

that is, $X_1 + \cdots + X_n$ is Erlang (n, μ) distributed. To prove this, note that the Laplace transform of the probability density $f_n(x)$ of $X_1 + \cdots + X_n$ is given by

$$\begin{aligned} f_n^*(s) &= E[e^{-s(X_1+\cdots+X_n)}] \\ &= E(e^{-sX_1}) \cdots E(e^{-sX_n}). \end{aligned}$$

Noting that $E(e^{-sX_i}) = \int_0^\infty e^{-sx} \mu e^{-\mu x}\, dx = \mu/(s+\mu)$ for all s with $\mathrm{Re}(s) > -\mu$, it follows that

$$f_n^*(s) = \frac{\mu^n}{(s+\mu)^n}.$$

Using (E.10), the inversion of $f_n^*(s)$ shows that $f_n(x)$ is indeed given by the Erlang (n, μ) density.

Example E.2 The renewal function

Consider a renewal process for which the probability distribution function $B(x)$ of the interoccurrence times of the events has a probability density $b(x)$. The renewal function $M(x)$ is defined by

$$M(x) = \sum_{n=1}^{\infty} B_n(x), \tag{E.11}$$

where $B_n(x)$ is the probability distribution function of $X_1 + \cdots + X_n$. That is, $B_n(x)$ is the n-fold convolution of $B(x)$ with itself. The distribution function $B_n(x)$ has a probability density $b_n(x)$. Since $b_n(x)$ is the density of $X_1 + \cdots + X_n$,

$$\int_0^\infty e^{-sx} b_n(x)\, dx = E\left[e^{-s(X_1+\cdots+X_n)}\right] = \left[b^*(s)\right]^n,$$

where $b^*(s) = \int_0^\infty e^{-sx} b(x)\, dx$. By (E.4),

$$\int_0^\infty e^{-sx} B_n(x)\, dx = \frac{\left[b^*(s)\right]^n}{s}. \tag{E.12}$$

Thus we find $M^*(s) = \sum_{n=1}^{\infty} s^{-1}[b^*(s)]^n$, which yields the general formula

$$M^*(s) = \frac{b^*(s)}{s[1 - b^*(s)]}. \tag{E.13}$$

This expression can be inverted for the Erlang density. As an illustration, consider the case of $b(x) = \lambda^2 x e^{-\lambda x}$. Then $b^*(s) = [\lambda/(\lambda + s)]^2$ and so

$$M^*(s) = \frac{\lambda^2}{s^2(s + 2\lambda)}.$$

Partial-fraction expansion gives

$$M^*(s) = \frac{-\frac{1}{4}s + \frac{1}{2}\lambda}{s^2} + \frac{\frac{1}{4}}{s + 2\lambda}.$$

By applying (E.3), (E.9) and (E.10), we obtain

$$M(t) = \frac{1}{2}\lambda t - \frac{1}{4} + \frac{1}{4}e^{-2\lambda t}, \quad t \geq 0.$$

APPENDIX F. NUMERICAL LAPLACE INVERSION

For a long time numerical Laplace inversion had the reputation of being difficult and numerically unreliable. However, contrary to previous impressions, it is nowadays not difficult to compute probabilities and other quantities of interest in probability models by using reliable Laplace inversion methods. This appendix briefly discusses two effective Laplace inversion algorithms. These algorithms involve complex calculations. There is nothing magic about doing calculations with complex numbers. These calculations can be reduced to operations with real numbers by dealing separately with the real part and the imaginary part of the complex numbers. Simple facts such as the relation $e^{ix} = \cos(x) + i\sin(x)$ for any real x and the representation $z = re^{i\theta}$ for any complex number z are typically used in the calculations in addition to the basic rules for adding and multiplying two complex numbers. Here i denotes the complex number with $i^2 = -1$. Certain computer languages such as the language C++ have automatic provision for doing complex calculations. In many applied probability problems it is possible to derive an expression for the Laplace transform of some unknown function. Let the real-valued function $f(t)$ be an unknown function in the variable $t \geq 0$. Suppose its Laplace transform

$$f^*(s) = \int_0^{\infty} e^{-st} f(t)\, dt$$

in the complex variable s is known. Assume that the function $f(t)$ satisfies the following conditions:

1. $f(t)$ is of bounded variation on any finite interval.

2. $f(t)$ is continuous for $t \geq 0$.
3. For any $b > 0$ the function $e^{-bt} f(t)$ is monotone for $t \geq t_0(b)$ for some number $t_0(b)$.
4. $\int_0^\infty e^{-bt} |f(t)| \, dt < \infty$ for any $b > 0$.

In probability applications the function $f(t)$ is often the complementary cumulative probability distribution function of a continuous random variable. In this case the conditions 1 to 4 are automatically satisfied. A basic result from analysis is that a real-valued function $f(t)$ is of bounded variation if and only if it can be written as the difference of two monotone functions. Under the above conditions the following version of the Poisson summation formula from Fourier analysis holds:

$$\sum_{n=-\infty}^{\infty} f\left(t + \frac{2\pi n}{h}\right) e^{-b(t+2\pi n/h)} = \frac{h}{2\pi} \sum_{n=-\infty}^{\infty} f^*(b + inh) e^{inht}$$

for any constants $h, b > 0$. This Poisson summation formula is the basis for the following algorithm of Abate and Whitt (1992).

Inversion algorithm of Abate and Whitt

In Abate and Whitt (1992) it was shown that

$$f(t) = \frac{e^{\frac{1}{2}a}}{2t} f^*\left(\frac{a}{2t}\right) + \frac{e^{\frac{1}{2}a}}{t} \sum_{k=1}^{\infty} (-1)^k \mathrm{Re}\left(f^*\left(\frac{a + 2k\pi i}{2t}\right)\right) - \epsilon(a, t) \qquad \text{(F.1)}$$

for any constant $a > 0$, where the error term $\epsilon(a, t)$ is given by

$$\epsilon(a, t) = \sum_{n=1}^{\infty} e^{-na} f((2n + 1)t).$$

To calculate $f(t)$ from (F.1) for a given value of t, we need $f^*(s)$ for the sequence $\{(a + 2k\pi)/2t, k = 0, 1, \ldots\}$ of complex numbers. In calculating $f(t)$ through the representation (F.1) there are three possible sources of error. First the discretization error associated with $\epsilon(a, t)$. Second, the truncation error associated with approximately calculating the infinite series in (F.1). Third, the round-off error associated with subtracting positive numbers that are close to each other. The discretization error can be controlled by choosing the constant a sufficiently large. Assuming that the function $f(t)$ is bounded by 1, as typically holds in probability applications, it follows from the inequality

$$|\epsilon(a, t)| \leq \frac{e^{-a}}{1 - e^{-a}}$$

that the discretization error can be limited to 10^{-8} by choosing $a = 19.1$ and to 10^{-13} by choosing $a = 28.3$. However, the constant a should not be chosen unnecessarily large. The risk of losing significant digits when calculating the infinite series in (F.1) increases when the constant a gets too large. A useful method of summation for the infinite series in (F.1) is the classical Euler summation method. This method proves to be quite effective in accelerating the convergence of the alternating infinite series in (F.1). Also, the method decreases the risk of losing significant digits in the calculations. In Euler summation the infinite series $\sum_{k=0}^{\infty}(-1)^k a_k$ in (F.1) is approximated by the Euler sum

$$E(m,n) = \sum_{k=0}^{m} \binom{m}{k} 2^{-m} S_{n+k}$$

for appropriately chosen values of m and n, where

$$S_j = \sum_{k=0}^{j} (-1)^k a_k.$$

Numerical experience shows that the Euler sum $E(m,n)$ computes the infinite series $\sum_{k=0}^{\infty}(-1)^k a_k$ in (F.1) usually with an error of 10^{-13} or less when $n = 38$ and $m = 11$ are taken (this requires the computation of only 50 terms). For more details the interested reader is referred to Abate and Whitt (1992). The Abate–Whitt algorithm gives excellent results for functions $f(t)$ that are sufficiently smooth (say, twice continuously differentiable). However, the inversion algorithm performs less satisfactorily for points at which the function $f(t)$ or its derivative is not differentiable.

Inversion algorithm of Den Iseger

Another simple algorithm to invert Laplace transforms was given in Den Iseger (2002). In general this algorithm outperforms the Abate–Whitt algorithm in stability and accuracy. The strength of the Den Iseger algorithm is the fact that in essence it boils down to an application of the discrete FFT algorithm. The Den Iseger algorithm has the additional advantage of inverting the Laplace transform simultaneously at several points. Suppose you wish to calculate $f(t)$ for a number of points in the interval $[0, t_0]$. For appropriately chosen values of $\Delta > 0$ and $M > 1$ with $(M-1)\Delta = t_0$, the algorithm calculates the function values $f(\ell\Delta)$ for $\ell = 0, 1, \ldots, M-1$. The algorithm is based on the representation

$$f(\ell\Delta) \approx \frac{e^{b\ell}}{\Delta} \sum_{j=1}^{n} \alpha_j \int_{-1}^{1} \mathrm{Re}\left[f^*\left(\frac{b + i\lambda_j + i\pi t}{\Delta} \right) \right] \cos(\pi \ell (t+1))\, dt \qquad \text{(F.2)}$$

for appropriately chosen values of b and n, where the abscissae λ_j and the weights α_j for $j = 1, \ldots, n$ are given numbers that depend only on n. The error in (F.2) converges very fast to zero as n gets larger. For practical purposes it suffices to

Table F.1 The constants α_j and λ_j for $n = 8, 16$

$\alpha_j (n = 8)$	λ_j
2.00000000000000000000E+00	3.14159265358979323846E+00
2.00000000000009194165E+00	9.42477796076939341796E+00
2.00000030233693694331E+00	1.57079633498486685135E+01
2.00163683400961269435E+00	2.19918840702852034226E+01
2.19160665410378500033E+00	2.84288098692614839228E+01
4.01375304677448905244E+00	3.74385643171158002866E+01
1.18855502586988811981E+01	5.93141454252504427542E+01
1.09907452904076203170E+02	1.73674723843715552399E+02
$\alpha_j (n = 16)$	λ_j
2.00000000000000000000E+00	3.14159265358979323846E+00
2.00000000000000000000E+00	9.42477796076937971539E+00
2.00000000000000000000E+00	1.57079632679489661923E+01
2.00000000000000000000E+00	2.19911485751285526692E+01
2.00000000000000025539E+00	2.82743338823081392079E+01
2.00000000001790585116E+00	3.45575191894933477513E+01
2.00000009630928117646E+00	4.08407045355964511919E+01
2.00006881371091937456E+00	4.71239261219868564304E+01
2.00840809734614010315E+00	5.34131955661131603664E+01
2.18638923693363504375E+00	5.99000285454941069650E+01
3.03057284932114460466E+00	6.78685456453781178352E+01
4.82641532934280440182E+00	7.99199036559694718061E+01
8.33376254184457094255E+00	9.99196221424608443952E+01
1.67554002625922470539E+01	1.37139145843604237972E+02
4.72109360166038325036E+01	2.25669154692295029965E+02
4.27648046755977518689E+02	6.72791727521303673697E+02

take n as large as 8 or 16 to achieve a very high precision. In Table F.1 we give both for $n = 8$ and $n = 16$ the abscissae λ_j and the weights α_j for $j = 1, \ldots, n$.

It is convenient to rewrite (F.2) as

$$f(\ell\Delta) \approx \frac{e^{b\ell}}{\Delta} \sum_{j=1}^{n} \alpha_j \int_0^2 \operatorname{Re}\left[f^*\left(\frac{b + i\lambda_j + i\pi(t-1)}{\Delta} \right) \right] \cos(\pi \ell t)\, dt.$$

Put for abbreviation $g_\ell = \frac{1}{2} \sum_{j=1}^{n} \alpha_j \int_0^2 \operatorname{Re}\left[f^*\left(\frac{b+i\lambda_j+i\pi(t-1)}{\Delta} \right) \right] \cos(\pi \ell t)\, dt$. Then $f(\ell\Delta) \approx (2e^{b\ell}/\Delta) g_\ell$. The integral in g_ℓ is calculated by using the trapezoidal rule approximation with a division of the integration interval $(0, 2)$ into $2m$ subintervals of length $1/m$ for an appropriately chosen value of m. It is recommended to take $m = 4M$. This gives

$$g_\ell \approx \frac{1}{2m} \sum_{p=1}^{2m-1} f_p^* \cos\left(\frac{\pi \ell p}{m} \right) + \frac{f_0^* + f_{2m}^*}{2}, \tag{F.3}$$

where f_p^* is defined by

$$f_p^* = \sum_{j=1}^{n} \alpha_j \operatorname{Re}\left[f^*\left(\frac{b + i\lambda_j + i\pi(p/m - 1)}{\Delta} \right) \right], \quad p = 0, 1, \ldots, 2m.$$

The approximation of $(2e^{b\ell}/\Delta)g_\ell$ to $f(\ell\Delta)$ is extraordinarily accurate. Rather than calculating from (F.3) the constants g_ℓ for $\ell = 0, 1, \ldots, M-1$ by direct summation, it is much better to use the discrete Fast Fourier Transform method to calculate the constants g_ℓ for $\ell = 0, 1, \ldots, 2m-1$. More important than speeding up the calculations, the discrete FFT method has the advantage of its numerical stability. To see how to apply the discrete FFT method to (F.3), define $\widehat{g}_k$ by

$$\widehat{g}_k = \begin{cases} \frac{1}{2}(f_0^* + f_{2m}^*), & k = 0, \\ f_k^*, & k = 1, \ldots, 2m-1. \end{cases}$$

Then, we can rewrite the expression (F.3) for g_ℓ as

$$g_\ell \approx \frac{1}{2m} \operatorname{Re}\left[\sum_{k=0}^{2m-1} \widehat{g}_k e^{2\pi i \ell k/2m} \right] \tag{F.4}$$

for $\ell = 0, 1, \ldots, 2m-1$. The discrete FFT method can be applied to this representation. Applying the inverse discrete FFT method to the vector $(\widehat{g}_0, \ldots, \widehat{g}_{2m-1})$ yields the sought vector $(g_0, \ldots, g_{2m-1})$. Here is a summary of the algorithm:

Input: M, Δ, b, n and m.

Output: $f(k\Delta)$ for $k = 0, 1, \ldots, M-1$.

Step 1: Calculate for $p = 0, 1, \ldots, 2m$ and $1 \le j \le n$,

$$f_{jp}^* = \operatorname{Re}\left[f^*\left(\frac{b + i\lambda_j + i\pi(p/m - 1)}{\Delta} \right) \right].$$

Next calculate $f_p^* = \sum_{j=1}^{n} \alpha_j f_{jp}^*$ for $p = 0, 1, \ldots, 2m$. Let $\widehat{g}_0 = \frac{1}{2}(f_0^* + f_{2m}^*)$ and $\widehat{g}_k = f_k^*$ for $k = 1, \ldots, 2m-1$.

Step 2: Apply the inverse discrete FFT method to the vector $(\widehat{g}_0, \ldots, \widehat{g}_{2m-1})$ in order to obtain the desired vector $(g_0, \ldots, g_{2m-1})$.

Step 3: Let $f(\ell\Delta) = (2e^{b\ell}/\Delta)g_\ell$ for $0 \le \ell \le M-1$.

In step 3 of the algorithm g_ℓ is multiplied by $e^{b\ell}$. In order to avoid numerical instability, it is important to choose b not too large. Assuming that the ratio m/M is large enough, say 4, numerical experiments indicate that $b = 22/m$ gives results that are almost of machine accuracy $2E - 16$ (in general, it is best to choose b somewhat larger than $-\ln(\xi)/(2m)$ where ξ is the machine precision). If f is sufficiently smooth, it usually suffices to take $n = 8$, otherwise $n = 16$ is

recommended. The parameter M is taken as a power of 2 (say, $M = 32$ or $M = 64$) while the parameter m is chosen equal to $4M$. The choices of M and Δ are not particularly relevant when f is smooth enough (theoretically, the accuracy increases when Δ gets smaller). In practice it is advisable to apply the algorithm for Δ and $\frac{1}{2}\Delta$ to see whether or not the results are affected by the choice of Δ.

Non-smooth functions

The Den Iseger algorithm may also perform unsatisfactorily when f or its derivative has discontinuities. In such cases the numerical difficulties may be circumvented by using a simple modification of the algorithm. To do this, assume that $f^*(s)$ can be represented as

$$f^*(s) = v(s, e^{x_0 s}) \tag{F.5}$$

for some real scalar x_0 and some function $v(s, u)$ with the property that for any fixed u the function $v(s, u)$ is the Laplace transform of a *smooth* function. As an example, consider the complementary waiting-time distribution $f(t) = P\{W_q > t\}$ in the $M/D/1$ queue with deterministic service times D and service in order of arrival; see Chapter 9. This function $f(t)$ is continuous but is not differentiable at the points $t = D, 2D, \ldots$. The Laplace transform $f^*(s)$ of $f(t)$ is given by

$$f^*(s) = \frac{\rho s - \lambda + \lambda e^{-sD}}{s[s - \lambda + \lambda e^{-sD}]}, \tag{F.6}$$

where λ is the average arrival rate and $\rho = \lambda D < 1$. Then (F.5) applies with

$$x_0 = -D \quad \text{and} \quad v(s, u) = \frac{\rho s - \lambda + \lambda u}{s(s - \lambda + \lambda u)}.$$

In this example we have indeed that for any fixed u the function $v(s, u)$ is the Laplace transform of an analytic (and hence smooth) function.

In the modified Den Iseger algorithm the basic relation (F.2) should be modified as

$$f(\ell\Delta) \approx \frac{e^{b\ell}}{\Delta} \sum_{j=1}^{n} \alpha_j \int_{-1}^{1} v^j(t) \cos(\pi \ell (t+1))\, dt \tag{F.7}$$

with

$$v^j(t) = \operatorname{Re}\left[v\left(\frac{b + i\lambda_j + i\pi t}{\Delta},\ \exp\left(\frac{i\pi x_0}{\Delta} - \frac{b + i\pi t}{\Delta} \right) \right) \right].$$

It is essential that in (F.7) the constant $\Delta > 0$ is chosen such that $|x_0|$ is a multiple of Δ, where x_0 comes from (F.5). As before, the integral in (F.7) can be approximated

Table F.2 The waiting-time probabilities

t	$P\{W_q > t\}$
1	0.554891814301507
5	0.100497238246398
10	0.011657108265013
25	0.00001819302497
50	3.820E-10

by the composite trapezoidal rule. In (F.3) the quantity f_p^* should now be read as

$$f_p^* = \sum_{j=1}^{n} \alpha_j \times \mathrm{Re}\left[v\left(\frac{b + i\lambda_j + i\pi(p/m - 1)}{\Delta}, \exp\left(\frac{i\pi x_0}{\Delta} - \frac{b + i\pi(p/m - 1)}{\Delta} \right) \right) \right].$$

The modification (F.7) gives excellent results (for continuous non-analytic functions one usually has an accuracy two or three figures less than machine precision). To illustrate this, we apply the modified approach to the Laplace transform (F.6) for the $M/D/1$ queue with service time $D = 1$ and traffic intensity $\rho = 0.8$. In Table F.2 the values of $f(t) = P\{W_q > t\}$ are given for $t = 1, 5, 10, 25$ and 50. The results in Table F.2 are accurate in all displayed decimals (13 to 15 decimals). The calculations were done with $M = 64$, $\Delta = 1$, $m = 4M$, $b = 22/m$ and $n = 8$. The inverse discrete FFT method was used to compute the g_ℓ from (F.4).

In sharp contrast with the accuracy of the modified approach (F.7), I found for the $M/D/1$ example the values 0.55607 and 0.55527 for $P\{W_q > t\}$ with $t = 1$ when using the unmodified Den Iseger inversion algorithm and the Abate–Whitt algorithm. These values give accuracy to only three decimal places. In the Abate–Whitt algorithm I took $a = 19.1$, $m = 11$ and $n = 38$ (I had to increase n to 5500 to get the value 0.5548948 accurate to five decimal places). The $M/D/1$ example shows convincingly how useful is the modification (F.7).

A scaling procedure

In applied probability problems one is often interested in calculating very small probabilities, e.g. probabilities in the range of 10^{-12} or smaller. In many cases asymptotic expansions are very useful for this purpose, but it may also be possible to use Laplace inversion with a scaling procedure. Such a scaling procedure was proposed in Choudhury and Whitt (1997). The idea of the procedure is very simple. Suppose that the function $f(t)$ is non-negative and that the (very small) function value $f(t_0)$ is required at the point $t_0 > 0$. The idea is to transform $f(t)$ into the scaled function

$$f_{a_0,a_1}(t) = a_0 e^{-a_1 t} f(t), \quad t \geq 0$$

for appropriately chosen constants a_0 and a_1 such that $f_{a_0,a_1}(t)$ is a probability density with mean t_0. The choice of the parameters a_0 and a_1 is intended to make $f_{a_0,a_1}(t)$ not too small. The unknown value $f_{a_0,a_1}(t_0)$ is computed by numerically inverting its Laplace transform $f^*_{a_0,a_1}(s)$, which is given by

$$f^*_{a_0,a_1}(s) = a_0 f^*(s + a_1).$$

Once $f_{a_0,a_1}(t_0)$ is computed the desired value $f(t_0)$ is easily obtained. The computation of the constants a_0 and a_1 is as follows:

1. Determine the smallest real number s^* such that $\int_0^\infty e^{-sx} f(x)\,dx$ is convergent for all s with $\mathrm{Re}(s) > s^*$ (possibly $s^* = -\infty$).
2. Try to find the real root a_1 of the equation

$$\frac{df^*(s)/ds}{f^*(s)} + t_0 = 0$$

on the interval (s^*, ∞). Since the function $-[1/f^*(s)]\,df^*(s)/ds$ can be shown to be decreasing on the interval (s^*, ∞), this equation has at most one root.

3. Determine $a_0 = 1/f^*(a_1)$.

In many applications this procedure works surprisingly well. We used the modified Den Iseger algorithm in combination with the scaling procedure to compute $P\{W_q > t\}$ for $t = 75$, 100 and 125 in the $M/D/1$ example discussed above. The respective values 8.022E − 15, 1.685E − 19 and 3.537E − 24 were calculated. Those values were exactly the same as the values obtained from the asymptotic expansion for $P\{W_q > t\}$ for t large.

Analytically intractable Laplace transforms

Sometimes the Laplace transform $f^*(s)$ of the unknown function $f(t)$ is not given in an explicit form, but contains an analytically intractable expression. To illustrate this, consider the Laplace transform $M^*(s)$ of the renewal function $M(t)$ for a renewal process. As shown by formula (E.12) in Appendix E, the Laplace transform $M^*(s)$ is given by

$$M^*(s) = \frac{b^*(s)}{s[1 - b^*(s)]},$$

where $b^*(s)$ is the Laplace transform of the interoccurrence-time density $b(t)$. Suppose now that this density is given by a lognormal density. In this particular case it is not possible to give an explicit expression for $b^*(s)$ and one has to handle an analytically intractable integral. How do we handle this? Suppose we wish to compute $M(t)$ for a number of t-values in the interval $[0, t_0]$. The key observation is that, by the representation (E.11), the renewal function $M(t)$ for $0 \le t \le t_0$ uses the interoccurrence-time density $b(t)$ only for $0 \le t \le t_0$. The same is true

for the waiting-time distribution function $W_q(t)$ in the $M/G/1$ queue with service in order of arrival. Then it follows from the representation (8.2.10) that $W_q(t)$ for $0 \leq t \leq t_0$ requires the service-time density $b(t)$ only for $0 \leq t \leq t_0$. If the Laplace transform $b^*(s)$ of the density $b(t)$ is analytically intractable, the idea is to approximate the density $b(t)$ by a polynomial $P(t)$ on the interval $[0, t_0]$ and by zero outside this interval. Consequently, the intractable Laplace transform $b^*(s)$ is approximated by a tractable expression

$$b^*_{app}(s) = \int_0^{t_0} e^{-st} P(t)\, dt.$$

A naive approach uses a single polynomial approximation $P(t)$ for the whole interval $[0, t_0]$. A polynomial approximation that is easy to handle is the Chebyshev approximating polynomial. Gauss–Legendre integration is then recommended to evaluate the required function values of $b^*_{app}(s)$. A code to compute the function values of the Chebyshev approximating polynomial at the points used in the numerical integration procedure can be found in the sourcebook by Press *et al.* (1992). One has a smooth function $P(t)$ when using a single Chebyshev polynomial approximation $P(t)$ for the whole interval $[0, t_0]$. However, a better accuracy is obtained by a more refined approach in which the function $b(t)$ on the interval $[0, t_0]$ is replaced by a piecewise polynomial approximation on each of the subintervals of length Δ with Δ as in (F.2). Den Iseger (2002) suggests approximating $b(t)$ on each of the subintervals $[k\Delta, (k+1)\Delta)$ by a linear combination of Legendre polynomials of degrees $0, 1, \ldots, 2n-1$ with n as in (F.2). This leads to an approximating function with discontinuities at the points $k\Delta$. However, this difficulty can be resolved by the modification (F.7) for non-smooth functions. Details can be found in Den Iseger (2002). A simpler approach seems possible when the analytically intractable Laplace transform $b^*(s)$ is given by $b^*(s) = E(e^{-sX})$ for a continuous random variable X with a strictly increasing probability distribution function $F(x)$. Then $b^*(s) = E[g(U, s)]$ for a uniform $(0, 1)$ random variable U, where $g(u, s) = \exp(-sF^{-1}(u))$. The (complex) integral $\int_0^1 g(u, s)\, du$ can be evaluated by Gauss–Legendre integration. The required numerical values of the inverse function $F^{-1}(u)$ may be obtained by using bisection.

APPENDIX G. THE ROOT-FINDING PROBLEM

The analysis of many queueing problems can be simplified by computing first the roots of a certain function inside or on the unit circle in the complex plane. It is a myth that the method of finding roots in the complex plane is difficult to use for practical purposes. In this appendix we address the problem of finding the roots of the equation

$$1 - z^c e^{\lambda D\{1-\beta(z)\}} = 0 \tag{G.1}$$

inside or on the unit circle. Here c is a positive integer, $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ is the generating function of a discrete probability distribution $\{\beta_j,\ j \geq 1\}$ and the real

numbers λ and D are positive constants such that $\lambda D\beta/c < 1$ with $\beta = \sum_{j=1}^{\infty} j\beta_j$. This root-finding problem arises in the analysis of the multi-server $M^X/D/c$ queue with batch arrivals. The equation (G.1) has c roots inside or on the unit circle. The proof is not given here, but is standard in complex analysis and uses the so-called Rouché theorem; see for example Chaudry and Templeton (1983). Moreover, all the c roots of (G.1) are distinct. This follows from the following general result in Dukhovny (1994): if $K(z)$ is the generating function of a non-negative, integer-valued random variable such that $K'(1) < c$ and $\left|zK'(z)\right| \leq K'(1)\,|K(z)|$ for $|z| \leq 1$, then all the roots of the equation $z^c = K(z)$ in the region $|z| \leq 1$ are distinct. Apply this result with $K(z) = e^{-\lambda D\{1-\beta(z)\}}$ and note that $K(z)$ is the generating function of the total number of arrivals in a compound Poisson arrival process; see Section 1.2.

To obtain the roots of (G.1) it is not recommended to directly apply Newton–Raphson iteration to (G.1). In this procedure numerical difficulties arise when roots are close together. This difficulty can be circumvented by a simple idea. The key to the numerical solution of equation (G.1) is the observation that it can be written as

$$z^c e^{\lambda D\{1-\beta(z)\}} = e^{2\pi ik} \tag{G.2}$$

where k is any integer. The next step is to use logarithms. The general logarithmic function of a complex variable is defined as the inverse of the exponential function and is therefore a many-valued function (as a consequence of $e^{z+2\pi i} = e^z$). It suffices to consider the principal branch of the logarithmic function. This principal branch is denoted by $\ln(z)$ and adds to each complex number $z \neq 0$ the unique complex number w in the infinite strip $-\pi < \mathrm{Im}(w) \leq \pi$ such that $e^w = z$. The principal branch of the logarithmic function of a complex variable can be expressed in terms of elementary functions by

$$\ln(z) = \ln(r) + i\theta$$

using the representation $z = re^{i\theta}$ with $r = |z|$ and $-\pi < \theta \leq \pi$. Since $e^{\ln(z)} = z$ for any $z \neq 0$, we can write (G.2) as

$$e^{c\ln(z)+\lambda D\{1-\beta(z)\}} = e^{2\pi ik}$$

with k is any integer. This suggests we should consider the equation

$$c\ln(z) + \lambda D\{1 - \beta(z)\} = 2\pi ik \tag{G.3}$$

where k is any integer. If for fixed k the equation (G.3) has a solution z_k, then this solution also satisfies (G.2) and so z_k is a solution of (G.1). The question is to find the values of k for which the equation (G.3) has a solution in the region $|z| \leq 1$. It turns out that the c distinct solutions of (G.1) are obtained by solving (G.3) for the c consecutive values of k satisfying $-\pi < 2\pi k/c \leq \pi$. These values of k are $k = -\lfloor (c-1)/2 \rfloor, \ldots, \lfloor c/2 \rfloor$, where $\lfloor a \rfloor$ is the largest integer smaller than or equal to a. In solving (G.3) for these values of k, we can halve the amount of computational work by letting k run only from 0 to $\lfloor c/2 \rfloor$. To see this, note that the

complex conjugates of $\ln(z)$ and $\beta(z)$ are given by $\ln(\overline{z})$ and $\beta(\overline{z})$ (use that $\beta(z)$ is a power series in z with real coefficients). Thus, if z_ℓ is a solution to (G.3) with $k = \ell$, then the complex conjugate $\overline{z}_\ell$ is a solution to (G.3) with $k = -\ell$. Hence it suffices to let k run only from 0 to $\lfloor c/2 \rfloor$. Further, note that the solution of (G.3) with $k = 0$ is given by $z_0 = 1$. For each k with $1 \leq k \leq \lfloor c/2 \rfloor$ the equation (G.3) can be solved by using the well-known Newton–Raphson method. This powerful method uses the iteration

$$z^{(n+1)} = z^{(n)} - \frac{h(z^{(n)})}{h'(z^{(n)})}$$

when the equation $h(z) = 0$ has to be solved. Applied to the equation (G.3), the iterative scheme becomes

$$z_k^{(n+1)} = z_k^{(n)} \times \frac{1 - (\lambda D/c)[1 + z_k^{(n)}\beta'(z_k^{(n)}) - \beta(z_k^{(n)})] - \ln(z_k^{(n)}) + 2\pi ik/c}{1 - (\lambda D/c) z_k^{(n)} \beta'(z_k^{(n)})},$$

where $\beta'(z)$ is the derivative of $\beta(z)$. The starting value $z_k^{(0)}$ for the Newton–Raphson iteration has to be chosen properly. To make an appropriate choice for $z_k^{(0)}$, we have a closer look at the equation (G.3). Let us rewrite this equation as $\ln(z) = (\lambda D/c)\{\beta(z) - 1\} + 2\pi ik/c$ and analyse it for the case of light traffic with $\lambda \to 0$. Then the solution of the equation tends to $e^{2\pi ik/c}$. Inserting $z = e^{2\pi ik/c}$ on the right-hand side of the equation for $\ln(z)$ yields

$$z_k^{(0)} = \exp\left[(\lambda D/c)\{\beta(e^{2\pi ik/c}) - 1\} + 2\pi ik/c\right].$$

We empirically verified that this is an excellent choice for the starting value of the Newton–Raphson scheme. In the above approach the roots of (G.1) are calculated by solving (G.3) *separately* for each value of k. If some roots are close together, Newton–Raphson iteration may converge each time to the same root when this procedure is directly applied to (G.1). However, this numerical difficulty is eliminated when (G.3) is used as an intermediary.

The above approach for solving $1 - z^c e^{\lambda D\{1-\beta(z)\}} = 0$ can be modified to find the roots of the equation

$$z^c - A(z) = 0$$

inside or on the unit circle when $A(z)$ is the generating function of a non-negative, integer-valued random variable. Assuming that $A(0) \neq 0$ (otherwise, $z = 0$ is a root), the equation $z^c - A(z) = 0$ can be transformed into the equation

$$c\,\ln(z) - \ln(A(z)) = 2\pi ik$$

where k is any integer. In general it is recommended to solve this equation by the modified Newton–Raphson method; see Stoer and Bulirsch (1980). In the modified Newton–Raphson method the step size is adjusted at each iteration in order to ensure convergence. In the special case that $z^c - A(z)$ is a polynomial in z, the

equation $z^c - A(z) = 0$ can be also solved as an eigenvalue problem. Solving the nth degree polynomial equation $z^n - c_1 z^{n-1} - \cdots - c_{n-1} z - c_n = 0$ with $c_n \neq 0$ is equivalent to finding the eigenvalues of the matrix

$$\mathbf{A} = \begin{pmatrix} c_1 & c_2 & c_3 & . & . & . & c_{n-1} & c_n \\ 1 & 0 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & 0 & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 1 & 0 \end{pmatrix}.$$

Fast and reliable codes for computing eigenvalues are widely available.

Finally, we discuss the computation of the (complex) roots of the equation

$$(\alpha - s)^m - e^{-sD}\alpha^{m-1}(\alpha - ps) = 0 \tag{G.4}$$

in the right half-plane $\{s \mid \mathrm{Re}(s) > 0\}$, where $m > 0$ is a given integer and $\alpha > 0$, $D > 0$ and $0 \leq p < 1$ are given numbers. This equation appears in the analysis of the $Ph/D/1$ queue and the $D/Ph/1$ queue; see Section 9.5. The computation of the roots of equation (G.4) is more subtle than the computation of the roots of (G.1). The reason is that equation (G.4) has $m - 1$ roots when $m - p > \alpha D$ and m roots when $m - p < \alpha D$. To handle this subtlety, Newton–Raphson iteration should be used in combination with Smale's homotopy method. To explain this, we first rewrite (G.4) as

$$u^m - e^{-\alpha D(1-u)}(1 - p + pu) = 0 \tag{G.5}$$

by the change of variable $u = 1 - s/\alpha$. The roots of this equation have to be found in the region $\{u \mid \mathrm{Re}(u) < 1\}$ of the complex plane. In this region the equation (G.5) always has $m - 1$ (complex) roots. If $m - p < \alpha D$ then the equation has an additional root on (0, 1). This real root is most easily found by repeated substitution:

$$u_{\ell+1} = \left[e^{-\alpha D(1-u_\ell)}(1 - p + pu_\ell)\right]^{1/m}, \quad \ell = 0, 1, \ldots,$$

starting with $u_0 = 1 - 1/m$. Next we discuss the computation of the $m-1$ complex roots of (G.5). Put for abbreviation $\gamma = \alpha D/m$. In the same way as in the analysis of (G.1), we transform (G.5) into

$$\ln(u) = -(1 - u)\gamma + \frac{1}{m}\ln(1 - p + pu) + 2\pi i \frac{k}{m} \tag{G.6}$$

for $k = 1, 2, \ldots, m - 1$. To solve (G.6) for fixed k, we use Smale's continuation process in which parameters $\overline{\gamma}$ and $\overline{p}$ are continued from $\overline{\gamma} = 0$, $\overline{p} = 0$ onwards to $\overline{\gamma} = \gamma$, $\overline{p} = p$. For fixed k and given step size N_{step}, the equation

$$\ln(u) = -(1 - u)\gamma_j + \frac{1}{m}\ln(1 - p_j + p_j u) + 2\pi i \frac{k}{m} \tag{G.7}$$

is solved by Newton–Raphson iteration successively for $j = 1, \ldots, N_{step}$ with

$$\gamma_j = \frac{j}{N_{step}}\gamma \quad \text{and} \quad p_j = j\frac{p}{N_{step}}.$$

The Newton–Raphson iteration solving (G.7) for a given value of j starts with $u_0 = u^{(j-1)}$ with $u^{(j-1)}$ denoting the solution of (G.7) with $j-1$ instead of j. For $j = 1$ we take the starting value $u_0 = e^{2\pi ik/m}$, being the solution of $\ln(u) = 2\pi ik/m$. The procedure is very robust against the choice of N_{step}.

REFERENCES

Abate, J. and Whitt, W. (1992) The Fourier series method for inverting transforms of probability distributions. *Queueing Systems*, **10**, 5–88.

Abramowitz, M. and Stegun, I. (1965) *Handbook of Mathematical Functions*. Dover, New York.

Chaudry, M.L. and Templeton, J.G.C. (1983) *A First Course in Bulk Queues*. John Wiley & Sons, Inc., New York.

Choudhury, G.L. and Whitt, W. (1997) Probabilistic scaling for the numerical inversion of nonprobability transforms. *Informs J. on Computing*, **9**, 175–184.

Den Iseger, P. (2002) Numerical inversion of Laplace transforms using a Gaussian quadrature for the Poisson summation formula. *Prob. Engng. Inform. Sci.*, submitted.

Dukhovny, A. (1994) Multiple roots of some equations of queueing theory. *Stochastic Models*, **10**, 519–524.

Nojo, S. and Watanabe, H. (1987) A new stage method getting arbitrary coefficient of variation through two stages. *Trans. IECIE*, **E-70**, 33–36.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992) *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn. Cambridge University Press, New York.

Rudin, W. (1964) *Principles of Mathematical Analysis*, 2nd edn. McGraw-Hill, New York.

Stoer, J. and Bulirsh, R. (1980) *Introduction to Numerical Analysis*. Springer-Verlag, Berlin.

Whitt, W. (1982) Approximating a point process by a renewal process: I, two basic methods. *Operat. Res.*, **30**, 125–147.

Index

A First Course in Stochastic Models H.C. Tijms