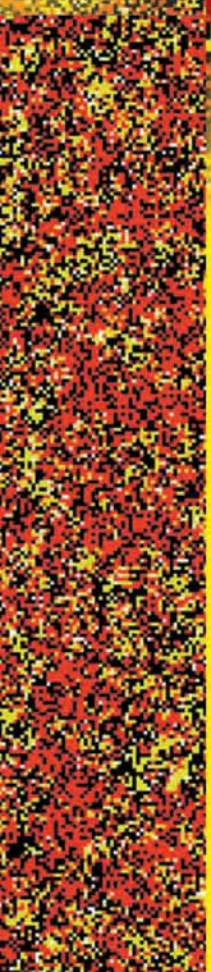
STOCHASTIC MODELLING AND APPLIED PROBABILITY

57

Søren Asmussen
Peter W. Glynn

Stochastic Simulation

Algorithms and Analysis



Springer

Stochastic Mechanics

Random Media

Signal Processing and Image Synthesis

Mathematical Economics and Finance

Stochastic Optimization

Stochastic Control

Stochastic Models in Life Sciences

Stochastic Modelling and Applied Probability

(Formerly:
Applications of Mathematics)

57

Stochastic Modelling and Applied Probability
formerly: Applications of Mathematics

1 Fleming/Rishel, **Deterministic and Stochastic Optimal Control** (1975)
2 Marchuk, **Methods of Numerical Mathematics** (1975, 2nd. ed. 1982)
3 Balakrishnan, **Applied Functional Analysis** (1976, 2nd. ed. 1981)
4 Borovkov, **Stochastic Processes in Queueing Theory** (1976)
5 Liptser/Shiryaev, **Statistics of Random Processes I: General Theory** (1977, 2nd. ed. 2001)
6 Liptser/Shiryaev, **Statistics of Random Processes II: Applications** (1978, 2nd. ed. 2001)
7 Vorob'ev, **Game Theory: Lectures for Economists and Systems Scientists** (1977)
8 Shiryaev, **Optimal Stopping Rules** (1978)
9 Ibragimov/Rozanov, **Gaussian Random Processes** (1978)
10 Wonham, **Linear Multivariable Control: A Geometric Approach** (1979, 2nd. ed. 1985)
11 Hida, **Brownian Motion** (1980)
12 Hestenes, **Conjugate Direction Methods in Optimization** (1980)
13 Kallianpur, **Stochastic Filtering Theory** (1980)
14 Krylov, **Controlled Diffusion Processes** (1980)
15 Prabhu, **Stochastic Storage Processes: Queues, Insurance Risk, and Dams** (1980)
16 Ibragimov/Has'minskii, **Statistical Estimation: Asymptotic Theory** (1981)
17 Cesari, **Optimization: Theory and Applications** (1982)
18 Elliott, **Stochastic Calculus and Applications** (1982)
19 Marchuk/Shaidourov, **Difference Methods and Their Extrapolations** (1983)
20 Hijab, **Stabilization of Control Systems** (1986)
21 Protter, **Stochastic Integration and Differential Equations** (1990)
22 Benveniste/Métivier/Priouret, **Adaptive Algorithms and Stochastic Approximations** (1990)
23 Kloeden/Platen, **Numerical Solution of Stochastic Differential Equations** (1992, corr. 3rd printing 1999)
24 Kushner/Dupuis, **Numerical Methods for Stochastic Control Problems in Continuous Time** (1992)
25 Fleming/Soner, **Controlled Markov Processes and Viscosity Solutions** (1993)
26 Baccelli/Brémaud, **Elements of Queueing Theory** (1994, 2nd. ed. 2003)
27 Winkler, **Image Analysis, Random Fields and Dynamic Monte Carlo Methods** (1995, 2nd. ed. 2003)
28 Kalpazidou, **Cycle Representations of Markov Processes** (1995)
29 Elliott/Aggoun/Moore, **Hidden Markov Models: Estimation and Control** (1995)
30 Hernández-Lerma/Lasserre, **Discrete-Time Markov Control Processes** (1995)
31 Devroye/Györfi/Lugosi, **A Probabilistic Theory of Pattern Recognition** (1996)
32 Maitra/Sudderth, **Discrete Gambling and Stochastic Games** (1996)
33 Embrechts/Klüppelberg/Mikosch, **Modelling Extremal Events for Insurance and Finance** (1997, corr. 4th printing 2003)
34 Duflo, **Random Iterative Models** (1997)
35 Kushner/Yin, **Stochastic Approximation Algorithms and Applications** (1997)
36 Musiela/Rutkowski, **Martingale Methods in Financial Modelling** (1997, 2nd. ed. 2005)
37 Yin, **Continuous-Time Markov Chains and Applications** (1998)
38 Dembo/Zeitouni, **Large Deviations Techniques and Applications** (1998)
39 Karatzas, **Methods of Mathematical Finance** (1998)
40 Fayolle/Iasnogorodski/Malyshev, **Random Walks in the Quarter-Plane** (1999)
41 Aven/Jensen, **Stochastic Models in Reliability** (1999)
42 Hernandez-Lerma/Lasserre, **Further Topics on Discrete-Time Markov Control Processes** (1999)
43 Yong/Zhou, **Stochastic Controls. Hamiltonian Systems and HJB Equations** (1999)
44 Serfozo, **Introduction to Stochastic Networks** (1999)
45 Steele, **Stochastic Calculus and Financial Applications** (2001)
46 Chen/Yao, **Fundamentals of Queuing Networks: Performance, Asymptotics, and Optimization** (2001)
47 Kushner, **Heavy Traffic Analysis of Controlled Queueing and Communications Networks** (2001)
48 Fernholz, **Stochastic Portfolio Theory** (2002)
49 Kabanov/Pergamenshchikov, **Two-Scale Stochastic Systems** (2003)
50 Han, **Information-Spectrum Methods in Information Theory** (2003)

(continued after index)

Søren Asmussen Peter W. Glynn

Stochastic Simulation: Algorithms and Analysis

Springer

Authors

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

Peter W. Glynn
Department of Management Science
and Engineering
Institute for Computational and
Mathematical Engineering
Stanford University
Stanford, CA 94305–4026
glynn@stanford.edu

Mathematics Subject Classification (2000): 65C05, 60-08, 62-01, 68-01

Library of Congress Control Number: 2007926471

ISSN: 0172-4568
ISBN-13: 978-0-387-30679-7
e-ISBN-13: 978-0-387-69033-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

Sampling-based computational methods have become a fundamental part of the numerical toolset of practitioners and researchers across an enormous number of different applied domains and academic disciplines. *This book is intended to provide a broad treatment of the basic ideas and algorithms associated with sampling-based methods, often also referred to as* Monte Carlo algorithms *or as* stochastic simulation. The reach of these ideas is illustrated here by discussing a wide range of different applications. Our goal is to provide coverage that reflects the richness of both the applications and the models that have found wide usage.

Of course, the models that are used differ widely from one discipline to another. Some methods apply across the entire simulation spectrum, whereas certain models raise particular computational challenges specific to those model formulations. As a consequence, the first part of the book focuses on general methods, whereas the second half discusses model-specific algorithms. The mathematical level is intended to accommodate the reader, so that for models for which even the model formulation demands some sophistication on the part of the reader (e.g., stochastic differential equations), the mathematical discussion will be at a different level from that presented elsewhere. While we deliver an honest discussion of the basic mathematical issues that arise in both describing and analyzing algorithms, we have chosen not to be too fussy with regard to providing precise conditions and assumptions guaranteeing validity of the stated results. For example, some theorem statements may omit conditions (such as moment hypotheses) that, while necessary mathematically, are not key to

understanding the practical domain of applicability of the result. Likewise, in some arguments, we have provided an outline of the key mathematical steps necessary to understand (for example) a rate of convergence issue, without giving all the mathematical details that would serve to provide a complete and rigorous proof.

As a result, we believe that this book can be a useful simulation resource to readers with backgrounds ranging from an exposure to introductory probability to a much more advanced knowledge of the area. Given the wide range of examples and application areas addressed, our expectation is that students, practitioners, and researchers in statistics, probability, operations research, economics, finance, engineering, biology, chemistry, and physics will find the book to be of value. In addition to providing a development of the area pertinent to each reader's specific interests, our hope is that the book also serves to broaden our audience's view of both Monte Carlo and stochastic modeling, in general.

There exists an extensive number of texts on simulation and Monte Carlo methods. Classical general references in the areas covered by this book are (in chronological order) Hammersley & Handscombe [173], Rubinstein [313], Ripley [300], and Fishman [118]. A number of further ones can be found in the list of references; many of them contain much practically oriented discussion not at all covered by this book. There are further a number of books dealing with special subareas, for example Gilks et al. [129] on Markov chain Monte Carlo methods, Newman & Barkema [276] on applications to statistical physics, Glasserman [133] on applications to mathematical finance, and Rubinstein & Kroese [318] on the cross-entropy method.

In addition to standard journals in statistics and applied probability, the reader interested in pursuing the literature should be aware of journals like *ACM TOMACS* (*ACM Transactions of Modeling and Computer Simulation*), *Management Science*, and the *IEEE* journals. Of course, today systematic scans of journals are to a large extent replaced by searches on the web. At the end of the book after the References section, we give some selected web links, being fully aware that such a list is likely to be outdated soon. These links also point to some important recurrent conferences on simulation, see in particular [w^3.14], [w^3.16], [w^3.17], [w^3.20].

The book is designed as a potential teaching and learning vehicle for use in a wide variety of courses. Our expectation is that the appropriate selection of material will be highly discipline-dependent, typically covering a large portion of the material in Part A on general methods and using those special topics chapters in Part B that reflect the models most widely used within that discipline. In teaching this material, we view some assignment of computer exercises as being essential to gaining an understanding and intuition for the material. In teaching graduate students from this book, one of us (SA) assigns a computer lab of three hours per week to complement lectures of two hours per week. Exercises labeled (A) are designed for such

a computer lab (although whether three hours is sufficient will depend on the students, and certainly some home preparation is needed). We have also deliberately chosen to not focus the book on a specific simulation language or software environment. Given the broad range of models covered, no single programming environment would provide a good universal fit. We prefer to let the user or teacher make the software choice herself. Finally, as a matter of teaching philosophy, we do not believe that programming should take a central role in a course taught from this book. Rather, the focus should be on understanding the intuition underlying the algorithms described here, as well as their strengths and weaknesses. In fact, to avoid a focus on the programming per se, we often hand out pieces of code for parts that are tedious to program but do not involve advanced ideas. Exercises marked (TP) are theoretical problems, highly varying in difficulty.

Since the first slow start of the writing of this book in 1999, we have received a large number of useful comments, suggestions, and corrections on earlier version of the manuscript. Thanks go first of all to the large number of students who have endured coping with these early versions. It would go too far to mention all the colleagues who have helped in one way or another. However, for a detailed reading of larger parts it is a pleasure to thank Hansjörg Albrecher, Morten Fenger-Grøn, Pierre L'Ecuyer, Thomas Mikosch, Leonardo Rojas-Nandayapa, and Jan Rosiński. At the technical level, Lars Madsen helped with many problems that were beyond our LaTeX ability.

A list of typos will be kept at [w^3.1], and we are greatful to be informed of misprints as well as of more serious mistakes and omissions.

Aarhus and Stanford — Søren Asmussen
February 2007 — Peter W. Glynn

Contents

Notation

Internal Reference System

The chapter number is specified only if it is not the current one. As examples, Proposition 1.3, formula (5.7) or Section 5 of Chapter IV are referred to as IV.1.3, IV.(5.7) and IV.5, respectively, in all chapters other than IV where we write Proposition 1.3, formula (5.7) (or just (5.7)) and Section 5.

Special Typeface

d	differential like in $\mathrm{d}x$, $\mathrm{d}t$, $F(\mathrm{d}x)$; to be distinguished from a variable or constant d, a function $d(x)$ etc.
e	the base $2.71\ldots$ of the natural logarithm; to be distinguished from e which can be a variable or a different constant.
i	the imaginary unit $\sqrt{-1}$; to be distinguished from a variable i (typically an index).
$\mathbb{1}$	the indicator function, for example $\mathbb{1}_A$, $\mathbb{1}_{x\in A}$, $\mathbb{1}\{x \in A\}$, $\mathbb{1}\{X(t) > 0 \text{ for some } t \in [0,1]\}$.
O, o	the Landau symbols. That is, $f(x) \,.=\, \mathrm{O}\big(g(x)\big)$ means that $f(x)/g(x)$ stays bounded in some limit, say $x \to \infty$ or $x \to 0$, whereas $f(x) = \mathrm{o}\big(g(x)\big)$ means $f(x)/g(x) \to 0$.
π	$3.1416\ldots$; to be distinguished from π which is often used for a stationary distribution or other.

$\mathcal{N}(\mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2.

Probability, expectation, variance, covariance are denoted $\mathbb{P}$, $\mathbb{E}$, $\mathbb{V}ar$, $\mathbb{C}ov$. The standard sets are $\mathbb{R}$ (the real line $(-\infty, \infty)$), the complex numbers $\mathbb{C}$, the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$, the integers $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$. Matrices and vectors are most often denoted by bold typeface, $\boldsymbol{C}$, $\boldsymbol{\Sigma}$, $\boldsymbol{x}$, $\boldsymbol{\alpha}$ etc., though exceptions occur. The transpose of $\boldsymbol{A}$ is denoted $\boldsymbol{A}^{\mathsf{T}}$.

Miscellaneous Mathematical Notation

$\stackrel{\text{def}}{=}$	a defining equality.
$\stackrel{\text{a.s.}}{\to}$	a.s. convergence
$\stackrel{\mathbb{P}}{\to}$	convergence in probability
$\stackrel{\mathscr{D}}{\to}$	convergence in distribution
$\stackrel{\mathscr{D}}{=}$	equality in distribution
$\longleftarrow$	an assignment in an algorithm (not used throughout)
$\lvert \cdot \rvert$	in addition to absolute value, also used for the number of elements (cardinality) $\lvert S \rvert$ of a set S, or its Lebesgue measure $\lvert S \rvert$.
$\mathbb{E}[X;\, A]$	$\mathbb{E}[X \mathbb{1}_A]$.
$\sim$	usually, $a(x) \sim b(x)$ means $a(x)/b(x) \to 1$ in some limit like $x \to 0$ or $x \to \infty$, but occassionally, other posssibilities occur. E.g. $X \sim \mathcal{N}(\mu, \sigma^2)$ specifies X to have a $\mathcal{N}(\mu, \sigma^2)$ distribution.
$\approx$	a different type of asymptotics, often just at the heuristical level.
$\stackrel{\mathscr{D}}{\approx}$	approximate equality in distribution.
$\propto$	proportional to.
$\widehat{F}[\cdot]$	the m.g.f. of a distribution F. Thus $\widehat{F}[\mathrm{i}s]$ is the characteristic function at s. Sometimes $\widehat{F}[\cdot]$ is also used for the probability generating function of a discrete r.v.

The letter U is usually reserved for a uniform$(0, 1)$ r.v., and the letter z for a quantity to be estimated by simulation, Z for a r.v. with $\mathbb{E}Z = z$. As is standard, Φ is used for the c.d.f. of $\mathcal{N}(0, 1)$ and $\varphi(x) \stackrel{\text{def}}{=} \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ for the density.. z_α often denotes the α-quantile of $\mathcal{N}(0, 1)$. A standard Brownian motion is denoted B and one with possibly drift $\mu \neq 0$ and/or variance σ^2 by W. Exceptions to all of this occur occasionally.

Conventions for a few selected standard distributions are given in A1.

Abbreviations

A-R	acceptance-rejection
BM	Brownian motion
c.g.f.	cumulant generating function (the log of the m.g.f.)
c.d.f.	cumulative distribution function, like $F(x) = \mathbb{P}(X \leq x)$
CIR	Cox-Ingersoll-Ross
CLT	central limit theorem
CMC	crude Monte Carlo
ECM	exponential change of measure
fBM	fractional Brownian motion
FD	finite differences
FIFO	First-in-first-out
GBM	geometric Brownian motion
GSMP	generalized semi-Markov process
GW	Galton-Watson
i.i.d.	independent identically distributed
i.o.	infinitely often
IPA	infinitesimal perturbation analysis
l.h.s.	left hand side
LLN	law of large numbers
LR	likelihood ratio
MAP	Markov additive process
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
m.g.f.	moment generating function
MSE	mean square error
NIG	normal inverse Gaussian
ODE	ordinary differential equation
O-U	Ornstein-Uhlenbeck
PDE	partial differential equation
QMC	quasi Monte Carlo
RBM	reflected Brownian motion
r.h.s.	right hand side
r.v.	random variable
s.c.v.	squared coefficient of variation
SDE	stochastic differential equation
TAVC	time average variance constant
VaR	Value-at-Risk
w.l.o.g.	without loss of generality
w.p.	with probability

Chapter I
What This Book Is About

1 An Illustrative Example: The Single-Server Queue

We start by introducing one of the classical models of applied probability, namely the single-server queue. Queuing models are widely used across an enormous variety of application areas, and arise naturally when resource contention among multiple users creates congestion effects. We shall use the single-server queue as a vehicle for introducing some of the key issues that a simulator may need to confront when using simulation as a numerical tool; a parallel area illustrative for this purpose, option pricing, will be introduced in Section 3.

Consider a single-server queue possessing an infinite capacity waiting room and processing customers according to a "first-in–first out" (FIFO) queue discipline. Let A_n, D_n, and W_n be the arrival time, departure time, and waiting time (exclusive of service) for the nth customer to enter the queue. The FIFO discipline then clearly implies that

$$W_{n+1} \;=\; \big[D_n - A_{n+1}\big]^+ ,$$

where $[x]^+ \stackrel{\text{def}}{=} \max(0,x)$ ($\stackrel{\text{def}}{=}$ means a defining equality). Also, it is evident that $D_n = A_n + W_n + V_n$, where V_n is the service time of customer n, and hence

$$W_{n+1} \;=\; \big[W_n + V_n - T_n\big]^+ \tag{1.1}$$

(the *Lindley recursion*), where $T_n \stackrel{\text{def}}{=} A_{n+1} - A_n$ is the time between the arrivals of customers n and $n+1$, $n = 0, 1, 2, \ldots$ Suppose that $\{V_n\}_{n\geq 0}$

and $\{T_n\}_{n\geq 0}$ are independent sequences of independent and identically distributed (i.i.d.) random variables (r.v.'s). Then the single-server queue model that we have described is known as the GI/G/1 queue.

Despite the simplicity of this model, it presents significant mathematical and computational challenges; in fact, thousands of papers have been devoted to the GI/G/1 queue and its applications. For example, consider computing the distribution of W_n. Even when $X_n \stackrel{\text{def}}{=} V_{n-1} - T_{n-1}$ has a distribution that can be computed explicitly (e.g., X_n is Gaussian or, more generally, infinitely divisible), the distribution of W_n can typically not be computed in closed form.

It follows from (1.5) below that

$$\mathbb{P}(W_n > x) \;=\; \int_{B_n(x)} \prod_{k=0}^{n-1} \mathbb{P}(V_k \in \mathrm{d}v_k)\mathbb{P}(T_k \in \mathrm{d}t_k)\,, \tag{1.2}$$

where

$$B_n(x) \;\stackrel{\text{def}}{=}\; \Big\{(v_0,t_0),\ldots,(v_{n-1},t_{n-1}) : \max_{k=0,\ldots,n-1} \sum_{j=k}^{n} (v_j - t_j) \;>\; x\Big\}\,,$$

so that $\mathbb{P}(W_n > x)$ can be computed as a $2n$-dimensional integral. Because of the high dimensionality, such a $2n$-dimensional numerical integration presents a significant challenge from a computational point of view. We shall return to this point in Chapter IX.

The distribution of W_n is an example of a *transient* characteristic, as opposed to *steady-state* or *stationary* characteristics, which are defined by taking the limit as $n \to \infty$. For example, an r.v. W_∞ having the limit distribution of W_n as $n \to \infty$ (provided such a limit exists as a probability measure on $\mathbb{R}$) is said to be the *steady-state waiting time*. Note that $\{W_n\}_{n\in\mathbb{N}}$ is a Markov chain with state space $[0,\infty)$. Therefore the theory of Markov chains with a discrete (i.e., finite or countable) state space suggests that under conditions corresponding to positive recurrence and aperiodicity,[1] W_∞ will exists, and that the Markov chain $\{W_n\}_{n\in\mathbb{N}}$ itself will obey the law of large numbers (LLN)

$$\frac{1}{N}\sum_{n=0}^{N-1} f(W_n) \;\stackrel{\text{a.s.}}{\to}\; \mathbb{E}f(W_\infty)\,, \quad N \to \infty. \tag{1.3}$$

This relation is one of the main reasons, if not the main one, for the interest in steady-state characteristics. Say we are interested in the average delay $N^{-1}\sum_0^{N-1} W_n$ of the first N customers. If N is large, (1.3) then asserts that $\mathbb{E}W_\infty$ should be a good approximation.

[1] The condition required for (1.3) and $W_\infty < \infty$ is that the load be strictly smaller than the offered service. This is expressed as $\rho < 1$, where $\rho = \mathbb{E}V/\mathbb{E}T$ is the *traffic intensity*; see [16].

Further transient characteristics of interest are first passage time quantities such as the time $\inf\{n : W_n > x\}$ until a customer experiences a long delay x, the number $\sigma \stackrel{\text{def}}{=} \inf\{n > 0 : W_n = 0\}$ of customers served in a busy period (recall $W_0 = 0$), and the total length $V_0 + \cdots + V_{\sigma-1}$ of the busy period. For both transient and steady-state characteristics, it is also of obvious interest to consider other stochastic processes associated with the system, such as the number $Q(t)$ of customers in system at time t (including the one being presently served), and the workload $V(t)$ (time to empty the system provided no new arrivals occur).

One of the key properties of the single-server queue is its close connection to random walk theory, which as a nice specific feature allows representations of steady-state distributions (equation (1.6) below) as well as transient characteristics (equation (1.5) below) in terms of an associated random walk. To make this connection precise, write as above $X_k = V_{k-1} - T_{k-1}$ and $S_n \stackrel{\text{def}}{=} X_1 + \cdots + X_n$ (with $S_0 = 0$), and note that if customer 0 enters an empty queue at time 0, then by (1.1),

$$\begin{aligned} W_1 &= \max(X_1, 0) = \max\big(S_1 - S_0, S_1 - S_1\big), \\ W_2 &= \max(W_1 + X_2, 0) \\ &= \max\big(\max(S_1 - S_0, S_1 - S_1) + S_2 - S_1, S_2 - S_2\big) \\ &= \max\big(S_2 - S_0, S_2 - S_1, S_2 - S_2\big) = S_2 - \min\big(S_0, S_1, S_2\big), \end{aligned}$$

and in general,

$$W_n = S_n - \min_{k=0,\ldots,n} S_k = \max_{k=0,\ldots,n} [S_n - S_k]. \tag{1.4}$$

Under our basic assumption that $\{V_n\}_{n\geq 0}$ and $\{T_n\}_{n\geq 1}$ are independent sequences of i.i.d. r.v.'s, $\{S_n\}_{n\geq 0}$ is a classical random walk, and (1.4) makes clear the connection between the GI/G/1 queue and the random walk.

Whereas (1.4) is a sample-path relation, a time-reversion argument translates (1.4) into a distributional relation of a simpler form. Indeed, using that

$$\begin{aligned} &\big(S_n - S_n, S_n - S_{n-1}, S_n - S_{n-2}, \ldots, S_n - S_1, S_n - S_0\big) \\ &\quad = \big(0, X_n, X_n + X_{n-1}, \ldots, X_n + \cdots + X_2, X_n + \cdots + X_1\big) \\ &\quad \stackrel{\mathscr{D}}{=} \big(0, X_1, X_1 + X_2, \ldots, X_1 + \cdots + X_{n-1}, X_1 + \cdots + X_n\big) \\ &\quad = \big(S_0, S_1, S_2, \ldots, S_{n-1}, S_n\big), \end{aligned}$$

where $\stackrel{\mathscr{D}}{=}$ denotes equality in distribution, we get

$$W_n = \max_{k=0,\ldots,n} [S_n - S_k] \stackrel{\mathscr{D}}{=} \max_{k=0,\ldots,n} S_k \stackrel{\text{def}}{=} M_n. \tag{1.5}$$

As a consequence, $W_n \stackrel{\mathscr{D}}{\to} W_\infty$ as $n \to \infty$, where

$$W_\infty \stackrel{\mathscr{D}}{=} M \stackrel{\text{def}}{=} \max_{k\geq 0} S_k\,. \tag{1.6}$$

It follows that if $\rho = \mathbb{E}V_0/\mathbb{E}T_0 < 1$ (i.e., the mean arrival rate $1/\mathbb{E}T_0$ is smaller than the service rate $1/\mathbb{E}V_0$), then W_∞ is a proper r.v. (so that the steady state is well defined), and

$$\mathbb{P}(W_\infty > x) \;=\; \mathbb{P}(M > x) \;=\; \mathbb{P}(\tau(x) < \infty)\,, \tag{1.7}$$

where $\tau(x) \stackrel{\text{def}}{=} \min\{n > 0 :\, S_n > x\}$.

It is easily seen that $\mathbb{P}(W_\infty > x)$ satisfies the integral equation

$$\mathbb{P}(W_\infty > x) \;=\; \int_0^\infty \mathbb{P}(W_\infty \in \mathrm{d}y)\mathbb{P}(X_1 > x - y) \tag{1.8}$$

(the analogue of the stationarity equation for a Markov chain). One possible means of computing the distribution of W_∞ is therefore to numerically solve (1.8). However, rewriting (1.8) in the equivalent form

$$\mathbb{P}(W_\infty \leq x) \;=\; \int_0^\infty \mathbb{P}(W_\infty \leq x - y)\mathbb{P}(X_1 \in \mathrm{d}y)$$

shows that (1.8) is of Wiener–Hopf type, and such equations are known to be numerically challenging.

The analytically most tractable special case of the GI/G/1 queue is the M/M/1 queue, where both the interarrival time and the service time distribution are exponential,[2] say with rates (inverse means) λ, μ. Then $\rho = \lambda/\mu$, and the distribution of W_∞ is explicitly available,

$$\mathbb{P}(W_\infty \leq x) \;=\; 1 - \rho \,+\, \rho(1 - \mathrm{e}^{-\gamma x})\,,$$

where $\gamma \stackrel{\text{def}}{=} \mu - \lambda$; the probabilistic meaning of this formula is that the probability $\mathbb{P}(W_\infty = 0)$ that a customer gets served immediately equals $1 - \rho$, whereas a customer experiences delay w.p. ρ, and conditionally upon this the delay has an exponential distribution with rate parameter γ. Further, this is also the distribution of the steady-state workload $V(\infty)$, and the steady-state queue length $Q(\infty)$ is geometric with success parameter $1 - \rho$ (cf. A1), i.e., $\mathbb{P}\big(Q(\infty) = n\big) \;=\; (1 - \rho)\rho^n$, $n \in \mathbb{N}$. There are also explicit formulas for a number of transient characteristics such as the busy period density and the transition probabilities $p_{ij}^t \;=\; \mathbb{P}\big(Q(s + t) = j \,\big|\, Q(s) = i\big)$ of the Markov process $\{Q(t)\}_{t\geq 0}$, but the expressions are complicated and involve Bessel functions, even infinite sums of such, cf. [16, Section III.9].

Beyond the M/M/1 queue, the easiest special case of the GI/G/1 queue is GI/M/1 (exponential services), where steady-state quantitites have an

[2] M stands for Markovian or Memoryless. Note that the arrival process is just a Poisson process with rate λ.

explicit distribution given that one has solved the transcendental equation $\mathbb{E}e^{\gamma(V-T)} = 1$. Also, M/G/1 (Poisson arrivals) simplifies considerably, as discussed in Example 5.15 below.

2 The Monte Carlo Method

Given the stochastic origin of the integration problem (1.3), it is natural to consider computing $\mathbb{P}(W_n > x)$ by appealing to a sampling-based method. In particular, suppose that we could implement algorithms (on one's computer) capable of generating two independent sequences of i.i.d. r.v.'s $\{V_n\}_{n\geq 0}$ and $\{T_n\}_{n\geq 0}$ with the appropriate service-time and interarrival-time distributions. Then, by recursively computing the W_k according to the Lindley recursion (1.1), we would thereby obtain the r.v. W_n. By repeatedly drawing additional V_k and T_k, one could then obtain R i.i.d. copies $W_{1n}, \ldots, W_{Rn}$ of W_n. The probability $z \stackrel{\text{def}}{=} \mathbb{P}(W_n > x)$ could then be computed via the sample proportion of the W_{rn} that are greater than x, namely via the estimator

$$\widehat{z} \stackrel{\text{def}}{=} \widehat{z}_R \stackrel{\text{def}}{=} \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{W_{rn} > x\}$$

($\mathbb{1}$ = indicator function). The LLN (1.3) guarantees, of course, that the algorithm converges to $z = \mathbb{P}(W_n > x)$ as the number R of independent replications tends to ∞.

This example exposes a basic idea in the area of stochastic simulation, namely to simulate independent realizations of the stochastic phenomenon under consideration and then to compute an estimate for the probability or expectation of interest via an appropriate estimator obtained from independent samples.

To be more precise, suppose that we want to compute $z = \mathbb{E}Z$. The idea is to develop an algorithm that will generate i.i.d. copies $Z_1, \ldots, Z_R$ of the r.v. Z and then to estimate z via the sample-mean estimator

$$\widehat{z} \stackrel{\text{def}}{=} \widehat{z}_R \stackrel{\text{def}}{=} \frac{1}{R}\sum_{r=1}^{R} Z_r\,. \tag{2.1}$$

In other words, one runs R independent computer experiments replicating the r.v. Z, and then computes z from the sample. Use of random sampling or a method for computing a probability or expectation is often called the *Monte Carlo method*. When the estimator $\widehat{z}$ of $z = \mathbb{E}Z$ is an average of i.i.d. copies of Z as in (2.1), then we refer to $\widehat{z}$ as a *crude Monte Carlo* (CMC) estimator.

Note that an LLN also holds for many dependent and asymptotically stationary sequences, see, for example, the discussion surrounding (1.3) for