STOCHASTIC MODELING AND THE THEORY OF QUEUES

Ronald W. Wolff

University of California, Berkeley

Prentice Hall, Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging-in-Publication Data
Wolff, Ronald W.,
Stochastic modeling and the theory of queues / Ronald W. Wolff.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-13-846692-0
1. Stochastic processes—Mathematical models. 2. Queueing theory.
I. Title.
QA274.W65 1988 88-19624
519.2—dc19 CIP

Editorial/production supervision and
interior design: Patrice Fraccio
Manufacturing buyer: Mary Noonan

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN 0-13-846692-0

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

PREFACE

Many years ago, I started to write a first-year graduate-level book on stochastic processes and the theory of queues, for two main reasons:

(1) Expositions of renewal theory began with "hard stuff," e.g., Blackwell's Theorem. Time-average behavior (the "easy stuff"), where most of the applications are, was hardly mentioned. Even when the easy stuff was mentioned, e.g., in a problem at the end of a chapter, it was not pointed out that the easy stuff does not depend on the hard stuff.

(2) Expositions of queueing theory were in even worse shape. The subject seemed to be a catalog of models that were solved by unrelated and often highly technical mathematical methods. Messy mathematical expressions were derigueur. Unifying ideas were missing.

My early treatment of the easy stuff, which was based on Tauberian theorems, wasn't so easy, and while I had most of the right unifying ideas for queueing theory, my understanding of them wasn't what it is today. Partially for these reasons, this first effort fizzled.

Some years later, Stuart Dreyfus and I decided to write an introductory operations research text, where he would handle optimization, and I would handle stochastic models. The treatment of stochastic models in most introductory texts is little more than formula plugging. I was confident that unifying ideas can be explained at the undergraduate level.

For this purpose, the time-average approach was simplified. A basic idea was to envisage processes evolving over time, either by drawing pictures, or observing real processes that we were modeling. A time average then became an average over sample paths of the process. Because all the models we would encounter were regenerative, and had renewal processes embedded in them, we needed only the strong law of large numbers to derive time-average properties. With this approach, the inspection paradox and related effects became easy to understand. Tauberian theorems were banished!

This fit beautifully with fundamental queueing results such as "$L = \lambda w$," a relation between continuous- and discrete-time averages. The concept of work was introduced, and what is now called the PASTA property of the Poisson process was explained and applied. Elementary queueing models were compared in important and understandable ways. At that time, I wrote most of what now are the more elementary parts of Chapter 5.

A publisher who was very interested in our project sent us several reviews. The reviewers all held the view that we were writing two incompatible books. The problem seemed to be that I was after deep conceptual understanding, through a unified approach, while Dreyfus, feeling that there is no unified approach to optimization in general, was writing more of a guided tour of diverse optimization methods. Although the publisher was still enthusiastic, we agreed with the reviewers, and abandoned the joint project.

Having overcome problems that stalled my first effort, and with nearly 100 pages of new notes, I returned to my original objective. This is the result.

In the intervening years, many other books have come out, and of course we know more now. In fact, several of the topics covered in latter chapters have exploded into subfields. This growth has made the need for a unified approach even more important today than it was when my first effort began.

While written as a graduate text, this book contains a substantial amount of material not found in other general-purpose queueing books. In addition, sections at the end of seven chapters discuss important points in detail, and relate them to the literature. These features should make this book a useful general reference both for research and applications.

Most chapters have extensive sets of problems. Some of them extend the theory, while others supply it, sometimes in artful ways. Many involve problem formulation, i.e., the ability to translate into appropriate mathematics a problem stated in phsyical terms. This is the reason for the word *modeling* rather than either *models* or *processes* in the title.

While this has turned out to be a larger book than planned, it is my hope that the unified approach and the stress on intuitive ideas have not been diluted by its size, and that they will improve our understanding of stochastic phenomena.

We strive for rigor as well as intuitive understanding, but we also have placed a limit on assumed mathematical background that sometimes makes this difficult to achieve, e.g., in the definition of what are called stopping times. These instances occur rarely, and the benefit of this limit is greater accessibility. Furthermore, it is my firm belief that a thorough understanding of the stochastic phenomena in this book does not require a higher level of mathematical abstraction.

It is a pleasure to acknowledge the corrections and suggestions received from many friends, colleagues, students, and former students. Betsy Greenberg, George Shanthikumar, Karl Sigman, and Ward Whitt looked over the entire manuscript, and Gordon Newell recently used portions of the manuscript in a course. Their suggestions have led to numerous improvements. I also thank Philip Bitar and Yat-Wah Wan for numerous corrections and other suggestions.

Preparation of this manuscript would have been impossible without the generous technical support provided by the Department of Industrial Engineering and Operations Research at Berkeley. Finally, I thank Genji Yamazaki and the Tokyo Metropolitan Institute of Technology for their support during the editorial phase of the manuscript.

Ronald W. Wolff

ON USING THIS BOOK

As a text, this book can be used in several ways. There is a natural division: Chapters 1 thorugh 4 for a renewal-theory (regenerative process) based course on stochastic processes, and the rest of the book for a course on queueing theory. In the latter case, there is more material than can be covered in one course. Many of the details of Chapter 11 may be omitted, as may whole topics such as Chapter 7, but it is my preference to at least explain what every major topic is about. For example, Chapter 7 helps put Chapter 6 in perspective.

At the cost of some redundancy, the introduction to queueing theory in Chapter 5 is written so that fundamental ideas can be understood without detailed knowledge of earlier chapters. Subsequent chapters rely more heavily on the earlier ones. Chapters 6 through 8 can be read independently of each other, and, aside from some knowledge of renewal theory, Chapter 9 is nearly self-contained.

The minimum prerequisites for this book are a year of calculus and a calculus-based course in probability theory. For Markov chains, some knowledge of matrix algebra is also helpful.

From time to time, we use mathematical concepts and results that go somewhat beyond these prerequisites. These include different definitions of an integral, and the validity of interchanging operations such as integral and sum. These issues arise primarily in proofs, rather than in examples or problems, and are discussed in the Appendix.

Shortcomings in probability background are more difficult to pin down. Inadequate practice at problem solving generally, and in conditional arguments in particular, is common. Thus one often hears "I thought I understood the lecture, but I can't do the homework." Certain elementary but useful matters of technique, e.g., indicator functions, may be unfamiliar. In stochastic processes, it is often very useful to fix a point in the sample space, so that a sequence of random variables becomes a sequence of real numbers. This may also seem new, but it is important for a proper understanding of the strong law of large numbers.

Chapter 1 is intended to help remedy for these shortcomings. It is intended as a review. Topics of particular importance for the book that may also be unfamiliar are treated in some detail toward the end of this chapter. For first-year graduate students, a heterogenous group, I find that at least a week of review of this material is worthwhile.

We occasionally footnote sections, portions of sections, or even portions of proofs with an asterisk (*). The reason, often given in the text, is one or more of these: The mathematical level is higher, this is a special topic that may be omitted, or what follows is intricate and does not contribute to understanding the results.

INTRODUCTION

In this book, we strive for a balance between the theory of stochastic processes and applications of the theory to model building and the analysis of specific models. The term *stochastic modeling* is intended to convey the latter emphasis.

Much of the theory is based on the notion of *starting over* of what we call *regeneration cycles*. This is the main concept underlying renewal theory; it provides the basis for our analysis of regenerative processes in general, and discrete- and continuous-time Markov chains in particular. Nearly all the queueing models in this book are regenerative, and this approach provides a theoretical foundation for these models as well.

Because this concept is so basic, it is explicitly introduced and used early in Chapter 2, much earlier than other books of comparable level.

Renewal theory is primarily concerned with *asymptotic* or *limiting* behavior. We place great emphasis on *time averages*, which have the technical name of *Cesàro limits*. For example, we ask ''what fraction of time is a process in state *B*,'' rather than the corresponding limiting probability of the process being in state *B* as time approaches infinity. We call limits of the latter type *pointwise*.

There are two reasons for this emphasis: First, a fraction of time is generally of greater practical significance than the corresponding ''limit at infinity.'' It has a more meaningful connection with performance measures for the models we will build. Secondly, the theory underlying the existence and representation of time averages is more elementary, and the averages hold under more general conditions than in the case for the corresponding pointwise limits.

Related to time averages is the notion of observing an ongoing process ''at random,'' where more formally, we consider stationary versions of processes.

The time-average point of view helps us to think in a clear and consistent manner, and to avoid mistakes. For example, the *inspection paradox* and related batch effects, which at first seem counterintuitive, are introduced early in Chapter 2 and carefully explained from the time-average point of view. The objective is to improve intuition so that when these effects occur in more complex models later (or, for that matter, in everyday life!), intuition leads in the right direction.

Realizations (sample paths) of elementary processes are shown in several figures. It is important to be able to think in these terms and to visualize processes evolving over time. This, after all, is what we would see observing the real thing—realized values of quantities that we model as random variables.

Theorems about the existence of pointwise limits, notably Blackwell's Theorem and the Key Renewal Theorem, are important for a full theoretical treatment of renewal theory. Sometimes they are relevant in applications. This topic is also treated in Chapter 2, but separately from and after time averages. Except for periodic chains in Chapter 3, this more advanced material is not needed in the rest of this book, and may be omitted. (Graduate students should be aware of these results and understand through elementary examples that pointwise limits may not exist in the lattice case.)

Another reason for the clear separation of time averages and pointwise limits is this: By mixing the two, or what is worse, hardly mentioning time averages, the reader is likely to come away with the mistaken impression that distributions must be *nonlattice* or *spread out* (see Chapter 2) for certain results to hold. This impression is reinforced by far too many papers in the literature that make these assumptions when they are not necessary. Thirty lashes!

Our point of view is the same for discrete-time processes, notably discrete-time Markov chains (Chapter 3). The distinction between periodic and aperiodic states is the analogue of lattice versus nonlattice distributions. A time average is now an average over discrete time points, usually the integers, and in Markov chains, a fraction of time becomes a *fraction of transitions*. By formulating in terms of time averages, an important theorem about the existence and uniqueness of a stationary probability vector is proven without mention of the distinction between periodic and aperiodic states.

Theoretically more difficult results about pointwise limits for discrete-time chains are covered toward the end of Chapter 3, and may be omitted.

Continuous-time Markov chains are treated in Chapter 4, along with (briefly) semi-Markov processes. New theoretical difficulties can occur for these chains, notably the possibility of an infinite number of transitions in finite time, called an *explosion*. This possibility will not occur (with positive probability) for applications of these chains in the rest of the book. Consequently, our approach is to introduce sufficient conditions that both eliminate this possibility and cover our subsequent needs. Both here and in Chapter 3, we briefly treat what are called *stationary measures*, because they are useful in Chapter 6.

Many examples in Chapters 2 through 4 and problems at the end of each chapter emphasize modeling for a wide range of applications, including queues, inventories, transportation, replacement and maintenance policies, and reliability. The organization of these chapters, however, is based on theoretical considerations: classes of stochastic processes and their properties.

Queueing theory, which is the subject of the rest of the book, is about modeling and analysis of congestion and delays at service facilities. Nearly all of the models we discuss are stochastic processes that fall under one or more headings in earlier chapters. However, queueing theory cannot be systematically treated by a series of examples in chapters on stochastic processes. Nor should it be viewed as a "handbook" of models that, as we proceed from chapter to chapter, are treated by what may appear to be unrelated mathematical methods.

We present a unified treatment of queueing theory, based on the time-average point of view introduced in Chapter 2.

Chapter 5 is an overview, a queueing short course. Measures of performance such as L and w are defined as (continuous or discrete) time averages. Important concepts and results are introduced and applied as early as possible, including "$L = \lambda w$," the concept of *work-in-system* and its relation to other quantities, and the result *Poisson arrivals see time averages (PASTA)*. These results are treated in a more theoretical manner toward the end of the chapter. Queues that can be modeled as continuous-time Markov chains are included for two purposes: to show how to model complicated situations, and to facilitate the comparision of single and multichannel queues, a theme we return to several times in later chapers.

The remaining chapters are more specialized, and each has a substantial literature. We will not attempt to summarize them here; the introduction and the last section of each chapter should meet this need.

Except in Chapter 5, where we briefly consider rush-hour behavior, our queueing models are regenerative processes. See Sections 2-21 and 11-6 for discussion of recent results on the regenerative nature of queues once thought not to have this property. For these results, it is necessary to generalize how we define a regenerative process; see Section 2-21. While beyond our scope, properties of these more general regenerative processes and their application to queueing networks is currently an active research area.

CONTENTS