Stochastic-Process Limits

An Introduction to Stochastic-Process Limits And their Application to Queues

Ward Whitt

AT&T Labs - Research
The Shannon Laboratory
Florham Park, New Jersey

Draft
June 13, 2001

Preface

0.1. What Is This Book About?

This book is about *stochastic-process limits*, i.e., limits in which a sequence of stochastic processes converges to another stochastic process. Since the converging stochastic processes are constructed from initial stochastic processes by appropriately scaling time and space, the stochastic-process limits provide a macroscopic view of uncertainty. The stochastic-process limits are interesting and important because they generate simple approximations for complicated stochastic processes and because they help explain the statistical regularity associated with a macroscopic view of uncertainty.

This book emphasizes the continuous-mapping approach to obtain new stochastic-process limits from previously established stochastic-process limits. The continuous-mapping approach is applied to obtain stochastic-process limits for *queues*, i.e., probability models of service systems or waiting lines. These limits for queues are called *heavy-traffic limits*, because they involve a sequence of models in which the offered loads are allowed to increase towards the critical value for stability. These heavy-traffic limits generate simple approximations for complicated queueing processes under normal loading and reveal the impact of variability upon queueing performance. By focusing on the application of stochastic-process limits to queues, this book also provides an introduction to heavy-traffic stochastic-process limits for queues.

0.2. In More Detail

More generally, this is a book about *probability theory* – a subject which has applications to every branch of science and engineering. Probability theory can help manage a portfolio and it can help engineer a communication network. As it should, probability theory tells us how to compute probabilities, but probability theory also has a more majestic goal: *Probability theory aims to explain the statistical regularity associated with a macroscopic view of uncertainty.*

In probability theory, there are many important ideas. But one idea might fairly lay claim to being the central idea: That idea is conveyed by the *central limit theorem*, which explains the ubiquitous bell-shaped curve: Following the giants – De Moivre, Laplace and Gauss – we have come to realize that, under regularity conditions, a sum of random variables will be approximately normally distributed if the number of terms is sufficiently large.

In the last half century, through the work of Erdös and Kac (1946, 1947), Doob (1949), Donsker (1951, 1952), Prohorov (1956), Skorohod (1956) and others, a broader view of the central limit theorem has emerged. We have discovered that there is not only statistical regularity in the n^{th} sum as n gets large, but there also is statistical regularity in the first n sums. That statistical regularity is expressed via a stochastic-process limit, i.e., a limit in which a sequence of stochastic processes converges to another stochastic process: A sequence of continuous-time stochastic processes generated from the first n sums converges in distribution to Brownian motion as n increases. That generalization of the basic central limit theorem (CLT) is known as *Donsker's theorem.* It is also called a functional central limit theorem (FCLT), because it implies convergence in distribution for many functionals of interest, such as the maximum of the first n sums. The ordinary CLT becomes a simple consequence of Donsker's FCLT, obtained by applying a projection onto one coordinate, making the ordinary CLT look like a view from Abbott's (1952) *Flatland.*

As an extension of the CLT, Donsker's FCLT is important because it has many significant applications, beyond what we would imagine knowing the CLT alone. For example, there are many applications in Statistics: Donsker's FCLT enables us to determine asymptotically-exact approximate distributions for many test statistics. The classic example is the *Kolmogorov-Smirnov statistic*, which is used to test whether data from an unknown source can be regarded as an independent sample from a candidate distribution. The stochastic-process limit identifies a relatively simple approximate distribution for the test statistic, for any continuous candidate cumulative distribution function, that can be used when the sample size is large. Indeed, early work on the Kolmogorov-Smirnov statistic by Doob (1949) and Donsker (1952) provided a major impetus for the development of the general theory of stochastic-process limits. The evolution of that story can be seen in the books by Billingsley (1968, 1999), Csörgő and Horváth (1993), Pollard (1984), Shorack and Wellner (1986) and van der Waart and Wellner (1996).

Donsker's FCLT also has applications in other very different directions. The application that motivated this book is the application to queues: *Donsker's FCLT can be applied to establish heavy-traffic stochastic-process limits for queues.* A heavy-traffic limit for an open queueing model (with input from outside that eventually departs) is obtained by considering a sequence of queueing models, where the input load is allowed to increase toward the critical level for stability (where the input rate equals the maximum potential output rate). In such a heavy-traffic limit, the steady-state performance descriptions, such as the steady-state queue length, typically grow without bound. Nevertheless, with appropriate scaling of both time and space, there may be a nondegenerate stochastic-process limit for the entire queue-length process, which can yield useful approximations and can provide insight into system performance. The approximations can be useful even if the actual queueing systems do not experience heavy traffic. *The stochastic-process limits strip away unessential details and reveal key features determining performance.*

We are especially interested in the scaling of time and space that occurs in these heavy-traffic stochastic-process limits. It is customary to focus attention on the limit process, which serves as the approximation, but the scaling of time and space also provides important insights. For example, the scaling may reveal a *separation of time scales*, with different phenomena occurring at different time scales. In heavy-traffic limits for queues, the separation of time scales leads to unifying ideas, such as the *heavy-traffic averaging principle* (Section 2.4.2) and the *heavy-traffic snapshot principle* (Remark 5.9.1).

We obtain these many consequences of Donsker's FCLT by applying the continuous-mapping approach: Various continuous-mapping theorems imply that convergence in distribution is preserved under appropriate functions, with the simple case being a single function that is continuous. The continuous-mapping approach is much more effective with the FCLT than the CLT because many more random quantities of interest can be represented as functions of the first n partial sums than can be represented as functions of only the n^{th} partial sum. Since many heavy-traffic stochastic-process limits for queues follow from Donsker's FCLT and the continuous-mapping approach, we see that the statistical regularity revealed by the heavy-traffic limits for queues can be regarded as a consequence of the central limit theorem.

In this book we tell the story about the expanded view of the central limit theorem in more detail. We focus on stochastic-process limits, Donsker's theorem and the continuous-mapping approach. We also put life into the general theory by providing a detailed discussion of one application — queues. We give an introductory account that should be widely accessible. To help visualize the statistical regularity associated with stochastic-process limits, we perform simulations and plot stochastic-process sample paths.

However, we hasten to point out that there already is a substantial literature on stochastic-process limits, Donsker's FCLT and the continuous-mapping approach, including two editions of the masterful book by Billingsley (1968, 1999). What distinguishes the present book from previous books on this topic is our focus on *stochastic-process limits with nonstandard scaling and nonstandard limit processes.*

An important source of motivation for establishing such stochastic-process limits for queueing stochastic processes comes from evolving communication networks: Beginning with the seminal work of Leland, Taqqu, Willinger and Wilson (1994), extensive *traffic measurements* have shown that the network traffic is remarkably bursty, exhibiting complex features such as *heavy-tailed probability distributions, strong (or long-range) dependence and self-similarity.* These features present difficult engineering challenges for network design and control; e.g., see Park and Willinger (2000) and Krishnamurthy and Rexford (2001). Accordingly, a goal in our work is to gain a better understanding of these complex features and the way they affect the performance of queueing models.

To a large extent, the complex features – the heavy-tailed probability distributions, strong dependence and self-similarity – can be *defined* through their impact on stochastic-process limits. Thus, a study of stochastic-process limits, in a sufficiently broad context, is directly a study of the complex features observed in network traffic. From that perspective, it should be clear that this book is intended as a response (but not nearly a solution) to the engineering challenge posed by the traffic measurements.

We are interested in the way complex traffic affects network performance. Since a major component of network performance is congestion (queueing effects), we abstract network performance and focus on the way the complex traffic affects the performance of queues. The heavy-traffic limits show that the complex traffic can have a dramatic impact on queueing performance! We show that there are again heavy-traffic limits with these complex features, but both the scaling and the limit process may change. As in the standard case, the stochastic-process limits reveal key features determining performance.

The heavy-tailed distributions and strong dependence can lead to stochastic-process limits with *jumps in the limit process*, i.e., stochastic-process limits in which the limit process has discontinuous sample paths. The jumps have engineering significance, because they reveal sudden big changes, when viewed in a long time scale.

Much of the more technical material in the book is devoted to establishing stochastic-process limits with jumps in the limit process, but there already are books discussing stochastic-process limits with jumps in the limit process. Indeed, Jacod and Shiryaev (1987) establish many such stochastic-process limits. To be more precise, from the technical standpoint, what distinguishes this book from previous books on this topic is our focus on stochastic-process limits with *unmatched jumps* in the limit process; i.e., stochastic process limits in which the limit process has jumps unmatched in the converging processes.

For example, we may have a sequence of stochastic processes with continuous sample paths converging to a stochastic process with discontinuous sample paths. Alternatively, before scaling, we may have stochastic processes, such as queue-length stochastic processes, that move up and down by unit steps. Then, after introducing space scaling, the discontinuities are asymptotically negligible. Nevertheless, the sequence of scaled stochastic processes can converge in distribution to a limiting stochastic process with discontinuous sample paths.

Jumps are not part of Donsker's FCLT, because Brownian motion has continuous sample paths. But the classical CLT and Donsker's FCLT do not capture all possible forms of statistical regularity that can prevail. Other forms of statistical regularity emerge when the assumptions of the classical CLT no longer hold. For example, if the random variables being summed have heavy-tailed probability distributions (which here means having infinite variance), then the classical CLT for partial sums breaks down. Nevertheless, there still may be statistical regularity, but it assumes a new form. Then there is a different FCLT in which the limit process has jumps!

But the jumps in this new FCLT are *matched jumps*; each jump corresponds to an exceptionally large summand in the sums. At first glance, it is not so obvious that unmatched jumps can arise. Thus, we might regard stochastic-process limits with unmatched jumps in the limit process as pathological, and thus not worth serious attention. Part of the interest here lies in the fact that such limits, not only can occur, but routinely do occur in interesting applications. In particular, unmatched jumps in the limit process frequently occur in heavy-traffic limits for queues in the presence of heavy-tailed probability distributions. For example, in a single-server queue, the queue-length process usually moves up and down by unit steps. Hence, when space scaling is introduced, the jumps in the scaled queue-length process are asymptotically negligible. Nevertheless, occasional exceptionally long service times can cause a rapid buildup of customers, causing the sequence of scaled queue-length processes to converge to a limit process with discontinuous sample paths. We give several examples of stochastic-process limits with unmatched jumps in the limit process in Chapter 6.

Stochastic-process limits with unmatched jumps in the limit process present technical challenges: Stochastic-process limits are customarily established by exploiting the function space D of all right-continuous $\mathbb{R}^k$-valued functions with left limits, endowed with the Skorohod (1956) J_1 topology (notion of convergence), which is often called "the Skorohod topology." However, that topology does not permit stochastic-process limits with unmatched jumps in the limit process.

As a consequence, to establish stochastic-process limits with unmatched jumps in the limit process, we need to use a nonstandard topology on the underlying space D of stochastic-process sample paths. Instead of the standard J_1 topology on D, we use the M_1 topology on D, which also was introduced by Skorohod (1956). Even though the M_1 topology was introduced a long time ago, it has not received much attention. Thus, a major goal here is to provide a systematic development of the function space D with the M_1 topology and associated stochastic-process limits.

It turns out the standard J_1 topology is stronger (or finer) than the M_1 topology, so that previous stochastic-process limits established using the J_1 topology also hold with the M_1 topology. Thus, while the J_1 topology sometimes cannot be used, the M_1 topology can almost always be used. Moreover, the extra strength of the J_1 topology is rarely exploited. Thus, we would be so bold as to suggest that, *if only one topology on the function space D is to be considered, then it should be the M_1 topology.*

Even though our motivation comes from queueing models and their application to describe the performance of evolving communication networks, there are many other possible applications of stochastic-process limits with jumps in the limit process. Indeed, stochastic-process limits with jumps in the limit process can arise whenever there are abrupt changes. There are natural applications to insurance, because insurance claim distributions often have heavy tails. There also are natural applications to finance, especially in the area of risk management; e.g., related to electricity derivatives. See Embrechts, Klüppelberg and Mikosch (1997), Adler, Feldman and Taqqu (1998) and Asmussen (2000).

In some cases, the fluctuations in a stochastic process are so strong that no stochastic-process limit is possible with a limiting stochastic process having sample paths in the function space D. In order to establish stochastic-process limits involving such dramatic fluctuations, we introduce larger function spaces than D, which we call E and F. The names are chosen to suggest a natural progression starting from the space C of continuous functions and going beyond D. We define topologies on the spaces E and F analogous to the M_2 and M_1 topologies on D. Thus we exploit our study of the M topologies on D in this later work.

Even though the special focus here is on heavy-traffic stochastic-process limits for queues allowing unmatched jumps in the limit process, many heavy-traffic stochastic-process limits for queues have no jumps in the limit process. That is the case whenever we can directly apply the continuous-mapping approach with Donsker's FCLT. Then we deduce that reflected Brownian motion can serve as an asymptotically-exact approximation for several queueing processes in a heavy-traffic limit. In the queueing chapters we show how those classic heavy-traffic limits can be established and applied. Indeed, the book is also intended to serve as a general introduction to heavy-traffic stochastic-process limits for queues.

0.3. Organization of the Book

The book has fifteen chapters, which can be roughly grouped into four parts, ordered according to increasing difficulty. The level of difficulty is far from uniform: The first part is intended to be accessible with less background. It would be helpful (necessary?) to know something about probability and queues.

The *first part*, containing the first five chapters, provides an informal introduction to stochastic-process limits and their application to queues. The first part provides a broad overview, mostly without proofs, intending to complement and supplement other books, such as Billingsley (1968, 1999).

Chapter 1 uses simulation to help the reader directly experience the statistical regularity associated with stochastic-process limits. Chapter 2 discusses applications of the random walks simulated in Chapter 1. Chapter 3 introduces the mathematical framework for stochastic-process limits. Chapter 4 provides an overview of stochastic-process limits, presenting Donsker's theorem and some of its generalizations. Chapter 5 provides an introduction to heavy-traffic stochastic-process limits for queues.

The *second part*, containing Chapters 6 – 10, shows how the unmatched jumps can arise and expands the treatment of queueing models. The first chapter, Chapter 6 uses simulation to demonstrate that there should indeed be unmatched jumps in the limit process in several examples. Chapter 7 continues the overview of stochastic-process limits begun in Chapter 4. The remaining chapters in the second part apply the stochastic-process limits, with the continuous-mapping approach, to obtain more heavy-traffic limits for queues.

The *third part*, containing Chapters 11 – 14, is devoted to the technical foundations needed to establish stochastic-process limits with unmatched jumps in the limit process. The earlier queueing chapters draw on the third part to a large extent. The queueing chapters are presented first to provide motivation for the technical foundations.

The third part begins with Chapter 11, which provides more details on the mathematical framework for stochastic-process limits, expanding upon the brief introduction in Chapter 3. Chapter 12 focuses on the function space D of right-continuous $\mathbb{R}^k$-valued functions with left limits, endowed with one of the nonstandard Skorohod (1956) M topologies (M_1 or M_2). As a basis for applying the continuous-mapping approach to establish new stochastic-process limits in this context, Chapter 13 shows that commonly used functions from D or $D \times D$ to D preserve convergence with the M topologies. The third part concludes with Chapter 14, which establishes heavy-traffic limits for networks of queues.

The *fourth part*, containing Chapter 15, is more exploratory. It initiates new directions for research. Chapter 15 introduces the new spaces larger than D that can be used to express stochastic-process limits for scaled stochastic processes with even greater fluctuations.

The organization of the book is described in more detail at the end of Chapter 3, in Section 3.6.

Additional material is contained in an *Internet Supplement*. The Internet Supplement has three purposes: First, it is intended to maintain a list of corrections for errors found after the book has been published. Second, it is intended to provide supporting details, such as omitted proofs, for material in the book. Third, it is intended to provide supplementary material related to the subject of the book. Pointers to the Internet Supplement will be provided throughout the book. The initial contents of the Internet Supplement appear at the end of the book in Appendix B. The Internet Supplement is available online:

http://www.research.att.com/~wow/supplement.html

0.4. What is Missing?

Even though this book is long, it only provides introductions to stochastic-process limits and heavy-traffic stochastic-process limits for queues.

There are several different kinds of limits that can be considered for probability distributions and stochastic processes. Here we only consider central limit theorems and natural generalizations to the functions space D. We omit other kinds of limits such as large deviation principles. For large deviation principles, the continuous-mapping approach can be applied using contraction principles. Large deviations principles can be very useful for queues; see Shwartz and Weiss (1995). For a sample of other interesting probability limits (related to the Poisson clumping heuristic), see Aldous (1989).

Even though much of the book is devoted to queues, we only discuss heavy-traffic stochastic-process limits for queues. There is a large literature on queues. Nice *general introductions to queues*, at varying mathematical levels, are contained in the books by Asmussen (1987), Cooper (1982), Hall (1991), Kleinrock (1975, 1976) and Wolff (1989).

Queueing theory is intended to aid in the *performance analysis* of complex systems, such as computer, communication and manufacturing systems. We discuss performance implications of the heavy-traffic limits, but we do not discuss performance analysis in detail. Jain (1991) and Gunther (1998) discuss the performance analysis of computer systems; Bertsekas and Gallager (1987) discuss the performance analysis of communication networks; and Buzacott and Shanthikumar (1993) and Hopp and Spearman (1996) discuss the performance analysis of manufacturing systems.

Since we are motivated by evolving communication networks, we discuss queueing models that arise in that context, but we do not discuss the context itself. For background on evolving communication networks, see Keshav (1997), Kurose and Ross (2000) and Krishnamurthy and Rexford (2001). For research on communication network performance, see Park and Willinger (2000) and recent proceedings of *IEEE INFOCOM* and *ACM SIGCOMM*:

http://www.ieee-infocom.org/2000/

http://www.acm.org/pubs/contents/proceedings/series/comm/

Even within the relatively narrow domain of *heavy-traffic stochastic-process limits for queues*, we only provide an introduction. Harrison (1985) provided a previous introduction, focusing on Brownian motion and Brownian queues, the heavy-traffic limit processes rather than the heavy-traffic limits themselves. Harrison shows how martingales and the Ito stochastic calculus can be applied to calculate quantities of interest and solve control problems. Newell (1982) provides useful perspective as well with his focus on deterministic and diffusion approximations. Harrison and Newell show that the limit processes can be used directly as approximations without considering stochastic-process limits. In contrast, we emphasize insights that can be gained from the stochastic-process limits, e.g., from the scaling.

The subject of heavy-traffic stochastic-process limits remains a very active research topic. Most of the recent interest focuses on *networks of queues with multiple classes of customers*. A principal goal is to determine good polices for scheduling and routing. That focus places heavy-traffic stochastic-process limits in the mainstream of operations research.

Multi-class queueing networks are challenging because the obvious stability criterion – having the traffic intensity be less than one at each queue – can in fact fail to be sufficient for stability; see Bramson (1994a, b). Thus, for general multi-class queueing networks, the very definition of heavy traffic is in question. For some of the recent heavy-traffic stochastic-process limits, new methods beyond the continuous-mapping approach have been required; see Bramson (1998) and Williams (1998a,b).

Discussion of the heavy-traffic approach to multi-class queueing networks, including optimization issues, can be found in the recent books by Chen and Yao (2001) and Kushner (2001), in the collections of papers edited by Yao (1994), Kelly and Williams (1995), Kelly, Zachary and Ziedins (1996), Dai (1998), McDonald and Turner (2000) and Park and Willinger (2000), and in recent papers such as Bell and Williams (2001), Harrison (2000, 2001a,b) and Kumar (2000). Hopefully, this book will help prepare readers to appreciate that important work and extend it in new directions.

Contents