S.G. Henderson and B.L. Nelson (Eds.), *Handbook in OR & MS, Vol. 13*

DOI: 10.1016/S0927-0507(06)13013-3

# Chapter 13
# Analysis for Design

*Ward Whitt*

*Department of Industrial Engineering and Operations Research, Columbia University, USA*
*E-mail: ww2040@columbia.edu*

**Abstract**

In this chapter we discuss analysis for the design of simulation experiments. By that we mean, not the traditional (important) methods to design statistical experiments, but rather techniques that can be used, before a simulation is conducted, to estimate the computational effort required to obtain desired statistical precision for contemplated simulation estimators. In doing so, we represent computational effort by simulation time, and that in turn by either the number of replications or the run length within a single simulation run. We assume that the quantities of interest will be estimated by sample means. In great generality, the required length of a single simulation run can be determined by computing the asymptotic variance and the asymptotic bias of the sample means. Existing theory supports this step for a sample mean of a function of a Markov process. We would prefer to do the calculations directly for the intended simulation model, but that usually is prevented by model complexity. Thus, as a first step, we usually approximate the original model by a related Markovian model that is easier to analyze. For example, relatively simple diffusion-process approximations to estimate required simulation run lengths for queueing models can often be obtained by heavy-traffic stochastic-process limits.

## 1 Introduction

Simulations are *controlled experiments*. Before we can run a simulation program and analyze the output, we need to choose a simulation model and decide what output to collect; i.e., we need to *design* the simulation experiment. Since (stochastic) simulations require statistical analysis of the output, it is often appropriate to consider the perspective of *experimental design*, e.g., as in Cochran and Cox (1992), Montgomery (2000) and Wu and Hamada (2000).

Simulations are also *explorations*. We usually conduct simulations because we want to learn more about a complex system we inadequately understand. To head in the right direction, we should have some well-defined goals and questions when we start, but we should expect to develop new goals and questions as

we go along. When we think about experimental design, we should observe that the time scale for computer simulation experiments tends to be much shorter than the time scale for the agricultural and medical experiments that led to the theory of experimental design. With the steadily increasing power of computers, computer simulation has become a relatively rapid process. After doing one simulation, we can quickly revise it and conduct others. Therefore, it is almost always best to think of simulation as an *iterative process*: We conduct a simulation experiment, look at the results and find as many new questions as answers to our original questions. Each simulation experiment suggests subsequent simulation experiments. Through a succession of these experiments, we gradually gain the better understanding we originally sought. To a large extent, it is fruitful to approach simulation in the spirit of *exploratory data analysis*, e.g., as in Tukey (1977), Velleman and Hoaglin (1981) and Chapter 1 of NIST/SEMATECH (2003).

Successful simulation studies usually involve an artful mix of both experimental design and exploration. We would emphasize the spirit of exploration, but we feel that some experimental design can be a big help. When we plan to hike in the mountains, in addition to knowing what peak we want to ascend, it is also good to have a rough idea how long it will take to get there: Should the hike take two hours, two days or two weeks?

That is just the kind of rough information we need for simulations. A major purpose of simulation experiments, often as a means to other ends, is to estimate unknown quantities of interest. When we plan to conduct a simulation experiment, in addition to knowing what quantities we want to estimate, it is also good to have a rough idea how long it will take to obtain a reliable estimate: Should the experiment take two seconds, two hours or two years?

As in Whitt (1989), in this chapter we discuss techniques that can be used, before a simulation is conducted, to estimate the computational effort required to obtain desired statistical precision for contemplated simulation estimators. Given information about the required computational effort, we can decide what cases to consider and how much computational effort to devote to each. We can even decide whether to conduct the experiment at all. We can also decide if we need to exploit *variance-reduction techniques* (or *efficiency-improvement techniques*), see Chapters 10–12 and 14–16.

The theoretical analysis we discuss should complement the experience we gain from conducting many simulation experiments. Through experience, we learn about the amount of computational effort required to obtain desired statistical precision for simulation estimators in various settings. The analysis and computational experience should reinforce each other, giving us better judgment. The methods in this chapter are intended to help develop more reliable expectations about statistical precision. We can use this knowledge, not only to design better simulation experiments, but also to evaluate simulation output analysis, done by others or ourselves.

At first glance, the experimental design problem may not seem very difficult. First, we might think, given the amazing growth in computer power, that the

computational effort rarely needs to be that great, but that is not the case: Many simulation estimation goals remain out of reach, just like many other computational goals; e.g., see Papadimitriou (1994).

Second, we might think that we can always get a rough idea about how long the runs should be by doing one *pilot run* to estimate the required simulation run lengths. However, there are serious difficulties with that approach. First, such a preliminary experiment requires that we set up the entire simulation before we decide whether or not to conduct the experiment. Nevertheless, if such a sampling procedure could be employed consistently with confidence, then the experimental design problem would indeed not be especially difficult. In typical simulation experiments, we want to estimate steady-state means for several different input parameters. Unfortunately, doing a pilot run for one set of parameters may be very misleading, because the required run length may change dramatically when the input parameters are changed.

To illustrate how misleading one pilot run can be, consider a simulation of a *queueing model*. Indeed, we shall use queueing models as the context examples throughout the chapter. Now consider the simulation of a single-server queue with unlimited waiting space (the $G/G/1/\infty$ model, e.g., see Cohen (1982) or Cooper (1982)), with the objective of estimating the mean steady-state (or long-run average) number of customers in the system, as a function of basic model data such as the arrival stochastic process and the service-time distribution. This queueing experimental design problem is interesting and important primarily because a uniform allocation of data over all cases (parameter values) is not nearly appropriate. Experience indicates that, for given statistical precision, the required amount of data increases dramatically as the traffic intensity $\rho$ (arrival rate divided by the service rate) increases toward the critical level for stability and as the arrival-and-service variability (appropriately quantified) increases. For example, the required simulation run length to obtain 5% relative error (width of confidence interval divided by the estimated mean) at a high traffic intensity such as 0.95 tends to be 100 times greater than at a lower traffic intensity such as 0.50. (The operative formula underlying this rough estimate is $f(\rho) \equiv (1-\rho)^{-2}$; note that $f(0.95)/f(0.50) = 400/4 = 100$. If we consider the more extreme case $\rho = 0.995$, then the factor is 10,000. If we used a criterion of absolute error instead of relative error, then the operative formula becomes even more impressive: then $f(\rho) \equiv (1-\rho)^{-4}$.)

In this queueing example, and throughout this paper, we use simulation time as our characterization of computational effort. (For a theoretical discussion of this issue, see Glynn and Whitt, 1992.) Some computational experience or additional experiments on the selected computer are needed to convert simulation time into computational effort. Since there is a degree of freedom in choosing the measuring units for time, it is important to normalize these time units. For example, in a queueing model we might measure time in terms of the number of arrivals that enter the system or we might stipulate that a representative service-time distribution has mean 1. On the positive side, focusing on required simulation time has the advantage that it yields characterizations

of computational effort that are independent of the specific computer used to conduct the simulation. It seems best to try to account for that important factor separately.

We assume that the quantities of interest will be estimated by *sample means*. (There are other estimation procedures; e.g., see Chapters 8 and 9.) With sample means, in great generality the required amount of simulation time can be determined by computing quantities called the *asymptotic variance* and the *asymptotic bias* of the sample means. Thus, we want to estimate these quantities before conducting the simulation. In general, that is not so easy to do, but existing theory supports this step for a sample mean of a function of a Markov process. However, the stochastic processes of interest in simulation models are rarely Markov processes. Thus, it is usually necessary to first approximate the given stochastic process by a Markov process in order to apply the techniques in this paper.

It is important to approach this approximation step with the right attitude. Remember that we usually only want to obtain a rough estimate of the required simulation run length. Thus, we may well obtain the desired insight with only a very rough approximation. We do not want this analysis step to take longer than it takes to conduct the simulation itself. So we want to obtain the approximation quickly and we want to be able to do the analysis quickly. Fortunately, it is often possible to meet these goals.

For example, we might be interested in simulating a non-Markovian open network of single-server queues. We might be interested in the queue-length distributions at the different queues. To obtain a rough estimate of the required simulation run length, we might first solve the traffic-rate equations to find the net arrival rate at each queue. That step is valid for non-Markovian queueing networks as well as Markovian queueing networks; e.g., see Chen and Yao (2001), Kelly (1979) or Walrand (1988). Given the net arrival rate at each queue, we can calculate the traffic intensity at each queue by multiplying the arrival rate times the mean service time. Then we might focus on the *bottleneck queue*, i.e., the queue with the highest traffic intensity. We do that because the overall required run length is usually determined by the bottleneck queue. Then we analyze the bottleneck queue separately (necessarily approximately).

We might approximate the bottleneck queue by the Markovian $M/M/1$ queue with the same traffic intensity, and apply the techniques described in this paper to the Markovian queue-length process in order to estimate the required simulation run length. Alternatively, to capture the impact of the arrival and service processes beyond their means, we might use heavy-traffic limit theorems to approximate the queue-length process of the bottleneck queue by a reflected Brownian motion (RBM); e.g., see Chen and Yao (2001) and Whitt (2002). We then apply the techniques described in this paper to the limiting RBM, which is also a Markov process. By the methods described in these last two paragraphs, we can treat quite general queueing-network models, albeit roughly.

*Here is how the rest of the chapter is organized*: We start in Section 2 by describing the *standard statistical framework*, allowing us to estimate the statistical precision of sample-mean estimators, both before and after the simulation experiment is conducted. In Section 2 we define the asymptotic variance and the asymptotic bias of a sample mean. We relate these asymptotic quantities to the ordinary variance and bias of a sample mean. We show the critical role played by the asymptotic variance in confidence intervals and thus for the required sample size to obtain desired statistical precision. We first discuss the classical statistical case of independent and identically distributed (i.i.d.) random variables, which arises naturally when the simulation estimate is based on *independent replications*. For i.i.d. random variables, the asymptotic variance coincides with the variance of a single random variable. Finally, we discuss the problem of initial transients and correlations that arise when we form the sample mean from a stochastic process observed over time within a single run.

In Section 3, following Whitt (1992), we indicate how to compute the asymptotic variance and the asymptotic bias of functions of continuous-time Markov chains. We describe a recursive algorithm for functions of birth-and-death processes. In Section 4 we consider several birth-and-death process examples, including the $M/M/1$ and $M/M/\infty$ queueing models. These examples show that model structure can make a big difference in the computational effort required for estimation by simulation.

In Section 5 we consider diffusion processes, which are continuous analogues of birth-and-death processes. We give integral representations of the asymptotic parameters for diffusion processes, which enable computation by numerical integration. In Section 6 we discuss applications of stochastic-process limits to the planning process. Following Whitt (1989) and Srikant and Whitt (1996), we show how heavy-traffic limits yield relatively simple diffusion approximations for the asymptotic variance and the asymptotic bias of sample-mean estimators for single-server and many-server queues. The time scaling in the heavy-traffic limits plays a critical role. In Section 7 we consider not collecting data for an initial portion of a simulation run to reduce the bias. Finally, in Section 8 we discuss directions for further research.

## 2 The standard statistical framework

### *2.1 Probability model of a simulation*

We base our discussion on a probability model of a (stochastic) simulation experiment: In the model, the simulation experiment generates an initial segment of a stochastic process, which may be a discrete-time stochastic process $\{X_n\colon\ n \geqslant 1\}$ or a continuous-time stochastic process $\{X(t)\colon\ t \geqslant 0\}$. We form the relevant sample mean

$$\overline{X}_n \equiv n^{-1}\sum_{i=1}^{n} X_i \quad \text{or} \quad \overline{X}_t \equiv t^{-1}\int_0^t X(s)\,\mathrm{d}s,$$

and use the sample mean to estimate the long-run average,

$$\mu = \lim_{n\to\infty} \overline{X}_n \quad \text{or} \quad \mu = \lim_{t\to\infty} \overline{X}_t,$$

which is assumed to exist as a proper limit with probability one (w.p.1). Under very general regularity conditions, the long-run average coincides with the expected value of the limiting steady-state distribution of the stochastic process. For example, supporting theoretical results are available for regenerative processes, Chapter VI of Asmussen (2003); stationary marked point processes, Section 2.5 of Sigman (1995); and generalized semi-Markov processes (GSMPs), Glynn (1989).

These stochastic processes arise in both observations from a single run and from independent replications. For example, in observations from a single run, a discrete-time stochastic process $\{X_n\colon\ n \geqslant 1\}$ arises if we consider the waiting times of successive arrivals to a queue. The random variable $X_n$ might be the waiting time of the $n$th arrival before beginning service; then $\mu$ is the long-run average waiting time of all arrivals, which usually coincides with the mean steady-state waiting time. On the other hand, $X_n$ might take the value 1 if the $n$th arrival waits less than or equal to $x$ minutes, and take the value 0 otherwise; then $\mu \equiv \mu(x)$ is the long-run proportion of customers that wait less than or equal to $x$ minutes, which usually corresponds to the probability that the steady-state waiting time is less than or equal to $x$ minutes.

Alternatively, in observations from a single run, a continuous-time stochastic process $\{X(t)\colon\ t \geqslant 0\}$ arises if we consider the queue length over time, beginning at time 0. The random variable $X(t)$ might be the queue length at time $t$ or $X(t)$ might take the value 1 if the queue length at time $t$ is less than or equal to $k$, and take the value 0 otherwise.

With independent replications (separate independent runs of the experiment), we obtain a discrete-time stochastic process $\{X_n\colon\ n \geqslant 1\}$. Then $X_n$ represents a random observation obtained from the $n$th run. For example, $X_n$ might be the queue length at time 7 in the $n$th replication or $X_n$ might be the average queue length over the time interval [0, 7] in the $n$th replication. Then the limit $\mu$ represents the long-run average over many independent replications, which equals the expected value of the random variable in any single run. Such expected values describe the expected *transient* (or time-dependent) behavior of the system.

### *2.2 Bias, mean-squared error and variance*

By assuming that the limits exist, we are assuming that we would obtain the exact answer if we devoted unlimited computational effort to the simulation experiment. In statistical language, e.g., see Lehmann and Casella (1998), we are assuming that the estimators $\overline{X}_n$ and $\overline{X}_t$ are *consistent estimators* of the quantity to be estimated, $\mu$. For finite sample size, we can describe the statistical precision by looking at the bias and the mean-squared error. The *bias*,

which we denote by $\bar{\beta}_n$ in the discrete-time case and $\bar{\beta}_t$ in the continuous-time case, indicates how much the expected value of the estimator differs from the quantity being estimated, and in what direction. For example, in the discrete-time case, the bias of $\overline{X}_n$ is

$$\bar{\beta}_n = \mathrm{E}[\overline{X}_n] - \mu.$$

The mean-squared error ($\mathrm{MSE}_n$ or $\mathrm{MSE}_t$) is the expected squared error, e.g.,

$$\mathrm{MSE}_n = \mathrm{E}\big[|\overline{X}_n - \mu|^2\big].$$

If there is no bias, then the MSE coincides with the variance of $\overline{X}_n$, which we denote by $\bar{\sigma}_n^2$, i.e.,

$$\bar{\sigma}_n^2 \equiv \mathrm{Var}(\overline{X}_n) \equiv \mathrm{E}\big[\big|\overline{X}_n - \mathrm{E}[\overline{X}_n]\big|^2\big].$$

Then we can write

$$\bar{\sigma}_n^2 \equiv \mathrm{Var}(\overline{X}_n) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j),$$

where $\mathrm{Cov}(X_i, X_j)$ is the *covariance*, i.e.,

$$\mathrm{Cov}(X_i, X_j) \equiv \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j].$$

Analogous formulas hold in continuous time. For example, then the variance of the sample mean $\overline{X}_t$ is

$$\bar{\sigma}_t^2 \equiv \mathrm{Var}(\overline{X}_t) = t^{-2} \int_0^t \int_0^t \mathrm{Cov}\big(X(u), X(v)\big)\,\mathrm{d}u\,\mathrm{d}v.$$

Unfortunately, these general formulas usually are too complicated to be of much help when doing preliminary planning. What we will be doing in the rest of this paper is showing how to develop simple approximations for these quantities.

### *2.3 The classical case: independent replications*

In statistics, the classical case arises when we have a discrete-time stochastic process $\{X_n\colon\ n \geqslant 1\}$, where the random variables $X_n$ are mutually independent and identically distributed (i.i.d.) with mean $\mu$ and finite variance $\sigma^2$, and we use the sample mean $\overline{X}_n$ to estimate the mean $\mu$. Clearly, the classical case arises whenever we use independent replications to do estimation, which of course is the great appeal of independent replications.

In the classical case, the sample mean $\overline{X}_n$ is a consistent estimator of the mean $\mu$ by the *law of large numbers* (LLN). Then there is no bias and the MSE coincides with the variance of the sample mean, $\bar{\sigma}_n^2$, which is a simple function

of the variance of a single observation $X_n$,

$$\bar{\sigma}_n^2 = \mathrm{MSE}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Moreover, by the *central limit theorem* (CLT), $\overline{X}_n$ is asymptotically normally distributed as the sample size $n$ increases, i.e.,

$$n^{1/2}[\overline{X}_n - \mu] \Rightarrow \mathrm{N}\big(0, \sigma^2\big) \quad \text{as } n \to \infty,$$

where $\mathrm{N}(a, b)$ is a normal random variable with mean $a$ and variance $b$, and "$\Rightarrow$" denotes convergence in distribution.

We thus use this *large-sample theory* to justify the approximation

$$\mathrm{P}(\overline{X}_n \leqslant x) \approx \mathrm{P}\left(\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \leqslant x\right) = \mathrm{P}\left(\mathrm{N}(0, 1) \leqslant \frac{x - \mu}{\sqrt{\sigma^2/n}}\right).$$

Based on this normal approximation, a $(1 - \alpha)100\%$ *confidence interval* for $\mu$ based on the sample mean $\overline{X}_n$ is

$$\left[\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

where

$$\mathrm{P}\big(-z_{\alpha/2} \leqslant \mathrm{N}(0, 1) \leqslant +z_{\alpha/2}\big) = 1 - \alpha.$$

A common choice is a 95% confidence interval, which corresponds to $\alpha = 0.05$; then $z_{\alpha/2} = 1.96 \approx 2$.

The *statistical precision* is typically described by either the *absolute width* or the *relative width* of the confidence interval, denoted by $w_a(\alpha)$ and $w_r(\alpha)$, respectively, which are

$$w_a(\alpha) = \frac{2z_{\alpha/2}\sigma}{\sqrt{n}} \quad \text{and} \quad w_r(\alpha) = \frac{2z_{\alpha/2}\sigma}{\mu\sqrt{n}}.$$

There are circumstances where each measure is preferred. For specified *absolute width* or *relative width* of the confidence interval, $\varepsilon$, and for specified level of precision $\alpha$, the *required sample size* $n_a(\varepsilon, \alpha)$ or $n_r(\varepsilon, \alpha)$ is then

$$n_a(\varepsilon, \alpha) = \frac{4\sigma^2 z_{\alpha/2}^2}{\varepsilon^2} \quad \text{or} \quad n_r(\varepsilon, \alpha) = \frac{4\sigma^2 z_{\alpha/2}^2}{\mu^2\varepsilon^2}. \tag{1}$$

From these formulas, we see that $n_a(\varepsilon, \alpha)$ and $n_r(\varepsilon, \alpha)$ are both inversely proportional to $\varepsilon^2$ and directly proportional to $\sigma^2$ and $z_{\alpha/2}^2$.

Standard statistical theory describes how observations (data) can be used to estimate the unknown quantities $\mu$ and $\sigma^2$. We might use a two-stage sampling procedure, exploiting the first stage to estimate the required sample size. However, here we are concerned with what we can do without any data at all.

We propose applying additional information about the model to obtain rough preliminary estimates for these parameters without data. Following the general approach of this paper, we suggest trying to estimate $\mu$ and $\sigma^2$ before conducting the simulation by analyzing the probability distribution of the outcome of a single replication, $X_n$ (using knowledge about the model). Admittedly, this preliminary estimation is often difficult to do; our approach is usually more useful in the context of one long run, which is discussed in the next section.

However, more can be done in this context than is often thought. Again, we must remember that we are only interested in making a rough estimate. Thus, we should be ready to make back-of-the-envelope calculations.

To illustrate what can be done, suppose that we focus on the relative-width criterion. With the relative-width criterion, it suffices to estimate the *squared coefficient of variation* (SCV, variance divided by the square of the mean)

$$c^2 \equiv \frac{\sigma^2}{\mu^2},$$

instead of both $\mu$ and $\sigma^2$. With the relative-width criterion, the required sample size is

$$n_r(\varepsilon, \alpha) = \frac{4c^2 z^2_{\alpha/2}}{\varepsilon^2}.$$

From the analysis above, we see that we only need to estimate a single parameter, the SCV $c^2$, in order to carry out this preliminary analysis. In many cases, we can make reasonable estimates based on "engineering judgment". For that step, it helps to have experience with variability as quantified by the SCV. First, note that the SCV measures the level of variability *independent of the mean*: The SCV of a random variable is unchanged if we multiply the random variable by a constant. We are thus focusing on the variability independent of the mean. Clearly, it is important to realize that the mean itself plays no role with the relative-width criterion.

Once we learn to focus on the SCV, we quickly gain experience about what to expect. A common reference case for the SCV is an exponential distribution, which has $c^2 = 1$. A unit point mass (deterministic distribution) has $c^2 = 0$. Distributions relatively more (less) variable than exponential have $c^2 >(<) 1$. In many instances we actually have a rough idea about the SCV. We might be able to judge in advance that the SCV is one of: (i) less than 1, but probably not less than 0.1, (ii) near 1, (iii) bigger than 1, but probably not bigger than 10, or (iv) likely to be large, even bigger than 10. In other words, it is not unreasonable to be able to estimate the SCV to within an order of magnitude (within a factor of 10). And that may be good enough for the desired rough estimates we want to make.

In lieu of information about the SCV, to obtain a rough estimate we can just let $c^2 = 1$. To proceed, we can also let $\alpha = 0.05$, so that $z_{\alpha/2} \approx 2$. Then, if we set $\varepsilon = 10^{-k}$, the required simulation run length is

$$n_r\big(10^{-k}, 0.05\big) = 16 \times (10)^{2k}.$$

Thus, when $c^2 = 1$, 10% relative precision requires about 1600 replications, while 1% relative precision requires 160,000 replications. If $c^2 \neq 1$, then we would multiply the number of required replications above by $c^2$. We thus can easily factor in the only unknown, the SCV $c^2$.

We have just reviewed the classical i.i.d. case, which is directly relevant when we use independent replications. However, in this chapter we concentrate on the more complicated case in which we consider an initial segment of a stochastic process from a single simulation run. It is good to have the classical i.i.d. case in mind, though, to understand the impact of bias and dependence upon the required computational effort.

### *2.4 An initial segment from a single run*

Now suppose that we intend to estimate a long-run average within a single run by the sample mean from an initial segment of a stochastic process, which could evolve in either discrete time or continuous time. The situation is now much more complicated, because the random observations need not be i.i.d. Indeed, they need not be either independent or identically distributed. Unlike the case of independent replications, we now face the problems of bias and dependence among the random variables.

Fortunately, there are generalizations of the classical i.i.d. framework that enable us to estimate the bias and the mean squared error as a function of the sample size in terms of only two fundamental parameters: the asymptotic bias and the asymptotic variance; see Whitt (1992) and references therein. That theory tells us that, under regularity conditions, both the bias and the MSE are of order $1/n$.

Within a single run, the stochastic processes tend to become stationary as time evolves. Indeed, now we assume that $X_n \Rightarrow X(\infty)$ as $n \to \infty$ (in the discrete-time case) and $X(t) \Rightarrow X(\infty)$ as $t \to \infty$ (in the continuous-time case). The stochastic processes fail to be stationary throughout all time primarily because it is necessary (or at least more convenient) to start the simulation in a special initial state. We thus can reduce the bias by choosing a good initial state or by deleting (not collecting statistics over) an initial portion of the simulation run. Choosing an appropriate initial state can be difficult if the stochastic process of interest is not nearly Markov. For example, even for the relatively simple $M/G/s/\infty$ queueing model, with $s$ servers and nonexponential service times, it is necessary to specify the remaining service time of all customers initially in service.

The asymptotic bias helps us to determine if it is necessary to choose a special initial state or delete an initial portion of the run. The asymptotic bias also helps us estimate the final bias, whether or not we choose a special initial state or delete an initial portion of the run. It also helps us determine what proportion of the full run should be deleted if we follow that procedure.

Under regularity conditions, there is a parameter $\bar{\beta}$ called the *asymptotic bias* such that

$$\bar{\beta} = \lim_{n\to\infty} n\bar{\beta}_n; \tag{2}$$

see Whitt (1992) and references therein. Given the definition of the bias $\bar{\beta}_n$, we see that the asymptotic bias must be

$$\bar{\beta} = \sum_{i=1}^{\infty} \big(\mathrm{E}[X_i] - \mu\big);$$

the regularity conditions ensure that the sum is absolutely convergent. We thus approximate the bias of $\overline{X}_n$ for any sufficiently large $n$ by

$$\bar{\beta}_n \approx \frac{\bar{\beta}}{n}.$$

This approximation reduces the unknowns to be estimated from the function $\{\bar{\beta}_n\colon\ n \geqslant 1\}$ to the single parameter $\bar{\beta}$. Corresponding formulas hold in continuous time.

Given that we can ignore the bias, either because it is negligible or because it has been largely removed by choosing a good initial state or by deleting an initial portion of the run, we can use the asymptotic variance to estimate the width of confidence intervals and thus the required run length to yield desired statistical precision. Under regularity conditions, there is a parameter $\bar{\sigma}^2$ called the *asymptotic variance* such that

$$\bar{\sigma}^2 = \lim_{n\to\infty} n\bar{\sigma}_n^2, \tag{3}$$

where (under the assumption that $\{X_n\colon\ n \geqslant 1\}$ is a stationary process)

$$\bar{\sigma}^2 = \mathrm{Var}(X_1) + 2\sum_{i=1}^{\infty} \mathrm{Cov}(X_1, X_{1+i}),$$

with $\sigma^2$ being the variance of $X_n$ and $\mathrm{Cov}(X_1, X_{1+i})$ being the lag-$i$ *autocovariance*. Because of the dependence, $\bar{\sigma}^2$ often is much bigger than $\sigma^2$. We thus approximate $\bar{\sigma}_n^2$, the variance of $\overline{X}_n$, for any sufficiently large $n$ by

$$\bar{\sigma}_n^2 \equiv \mathrm{Var}(\overline{X}_n) \approx \frac{\bar{\sigma}^2}{n}.$$

Again, this approximation reduces the unknowns to be estimated from the function $\{\bar{\sigma}_n^2\colon\ n \geqslant 1\}$ to the single parameter $\bar{\sigma}^2$.

In continuous time, we have the related asymptotic variance formula

$$\bar{\sigma}^2 = \lim_{t\to\infty} t\bar{\sigma}_t^2,$$

where (under the assumption that $\{X(t)\colon\ t \geqslant 0\}$ is a stationary process)

$$\bar{\sigma}^2 = 2\int_0^\infty \mathrm{Cov}\big(X(0), X(t)\big)\,\mathrm{d}t.$$

In continuous or discrete time, a critical assumption here is that the asymptotic variance $\bar{\sigma}^2$ is actually finite. The asymptotic variance could be infinite for two reasons: (i) heavy-tailed distributions and (ii) long-range dependence. In our context, we say that $X_n$ or $X(t)$ has a heavy-tailed distribution if its variance is infinite. In our context, we say that there is long-range dependence (without heavy-tailed distributions) if the variance $\mathrm{Var}(X_n)$ or $\mathrm{Var}(X(t))$ is finite, but nevertheless the asymptotic variance is infinite because the autocovariances $\mathrm{Cov}(X_j, X_{j+k})$ or $\mathrm{Cov}(X(t), X(t+k))$ do not decay quickly enough as $k \to \infty$; e.g., see Beran (1994), Samorodnitsky and Taqqu (1994) and Chapter 4 of Whitt (2002).

Assuming that $\bar{\sigma}^2 < \infty$, we can apply CLTs for weakly dependent random variables (involving other regularity conditions, e.g., see Section 4.4 of Whitt, 2002) to deduce that $\overline{X}_n$ (as well as $\overline{X}_t$) is again asymptotically normally distributed as the sample size $n$ increases, i.e.,

$$n^{1/2}[\overline{X}_n - \mu] \Rightarrow \mathrm{N}\big(0, \bar{\sigma}^2\big) \quad \text{as } n \to \infty,$$

so that the asymptotic variance $\bar{\sigma}^2$ plays the role of the ordinary variance $\sigma^2$ in the classical i.i.d. setting.

We thus again can use the large-sample theory to justify a normal approximation. The new $(1-\alpha)100\%$ confidence interval for $\mu$ based on the sample mean $\overline{X}_n$ is

$$\left[\overline{X}_n - z_{\alpha/2}\frac{\bar{\sigma}}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\bar{\sigma}}{\sqrt{n}}\right],$$

which is the same as for independent replications except that the asymptotic variance $\bar{\sigma}^2$ replaces the ordinary variance $\sigma^2$.

The formulas for the confidence interval relative width, $w_r(\alpha)$, and the required run length, $n_r(\varepsilon, \alpha)$, are thus also the same as for independent replications in (1) except that the asymptotic variance $\bar{\sigma}^2$ is substituted for the ordinary variance $\sigma^2$; e.g., the required simulation run length with a relative-width criterion is

$$n_r(\varepsilon, \alpha) = \frac{4\bar{\sigma}^2 z_{\alpha/2}^2}{\mu^2\varepsilon^2} \quad \text{and} \quad n_r\big(10^{-k}, 0.05\big) \approx \frac{\bar{\sigma}^2}{\mu^2} 16 \times (10)^{2k}. \tag{4}$$

From (1) and (4), we immediately see that the required run length is approximately $\bar{\sigma}^2/\sigma^2$ times greater when sampling from one run than with independent sampling (assuming that we could directly observe independent samples from the steady-state distribution, which of course is typically not possible).

As with independent replications, established simulation methodology and statistical theory tells how to estimate the unknown quantities $\mu$, $\bar{\beta}$ and $\bar{\sigma}^2$ from data; e.g., see Bratley et al. (1987) and Fishman (2001). Instead, we apply additional information about the model to obtain rough preliminary estimates for these parameters without data. For $\bar{\sigma}^2$, the representation of the asymptotic variance in terms of the autocovariances is usually too complicated to be of much help, but fortunately there is another approach, which we will describe in Section 3.

## 3 The asymptotic parameters for a function of a Markov chain

From the previous section, it should be apparent that we can do the intended preliminary planning if we can estimate the asymptotic bias and the asymptotic variance. We now start to describe how we can calculate these important parameters. We first consider functions of a Markov chain. That illustrates available general results. However, fast back-of-the-envelope calculations usually depend on diffusion approximations, based on stochastic-process limits, after doing appropriate scaling, which we discuss later in Sections 5 and 6. Indeed, the scaling is usually the key part, and that is so simple that back-of-the-envelope calculations are actually possible.

In this section, drawing on Whitt (1992), which itself is primarily a survey of known results (including Glynn (1984) and Grassman (1987a, 1987b) among others), we observe that (again under regularity conditions) we can calculate the asymptotic bias and the asymptotic variance whenever the stochastic process of interest is a function of a (positive-recurrent irreducible) Markov chain, i.e., when $X_n = f(Y_n)$ for $n \geqslant 1$, where $f$ is a real-valued function and $\{Y_n\colon\ n \geqslant 1\}$ is a Markov chain or when $X(t) = f(Y(t))$ for $t \geqslant 0$, where again $f$ is a real-valued function and $\{Y(t)\colon\ t \geqslant 0\}$ is a Markov chain. As noted before, we usually obtain the required Markov structure by approximating the given stochastic process by a related one with the Markov property.

In fact, as in Whitt (1992), we only discuss the case in which the underlying Markov chain has a finite state space (by which we mean countably finite, i.e., $\{0, 1, \ldots, m\}$, not $[c, d]$), but the theory extends to more general state spaces under regularity conditions. For illustrations, see Glynn (1994) and Glynn and Meyn (1996). But the finite-state-space condition is very useful: Under the finite-state-space condition, we can compute the asymptotic parameters numerically, without relying on special model structure. However, when we do have special structure, we can sometimes go further to obtain relatively simple closed-form formulas. We also obtain relatively simple closed-form formulas

when we establish diffusion-process approximations via stochastic-process limits.

### 3.1 Continuous-time Markov chains

We will discuss the case of a continuous-time Markov chain (CTMC); similar results hold for discrete-time Markov chains. Suppose that the CTMC $\{Y(t)\colon\ t \geqslant 0\}$ is irreducible with finite state space $\{0, 1, \ldots, m\}$ (which implies that it is positive recurrent). Our sample-mean estimator is

$$\overline{X}_t \equiv t^{-1} \int_0^t X(s)\,\mathrm{d}s, \quad t \geqslant 0,$$

where $X(t) = f(Y(t))$. (With the discrete state space, we write both $f(j)$ and $f_j$ for the value of $f$ at argument $j$.)

A finite-state CTMC is characterized by its *infinitesimal generator matrix* $Q \equiv (Q_{i,j})$; $Q_{i,j}$ is understood to be the derivative (from above) of the *probability transition function*

$$P_{i,j}(t) \equiv \mathrm{P}\big(Y(s+t) = j | Y(s) = i\big)$$

with respect to time $t$ evaluated at $t = 0$. However, in applications the model is specified by defining the infinitesimal generator $Q$. Then the probability transition function can be obtained subsequently as the solution of the ordinary differential equations

$$P'(t) = P(t)Q = QP(t),$$

which takes the form of the matrix exponential

$$P(t) = \mathrm{e}^{Qt} \equiv \sum_{k=0}^{\infty} \frac{Q^k t^k}{k!}.$$

Key asymptotic quantities associated with a CTMC are the stationary probability vector $\pi$ and the fundamental matrix $Z$. (By convention, we let vectors be row vectors.) The *stationary probability vector* $\pi$ can be found by solving the system of equations

$$\pi Q = 0 \quad \text{with } \sum_{i=0}^{m} \pi_i = 1.$$

The quantity we want to estimate is the expected value of the function $f$ (represented as a row vector) with respect to the stationary probability (row) vector $\pi$, i.e.,

$$\mu = \pi f^{\mathrm{T}} = \sum_{i=0}^{m} \pi_i f_i,$$

where "T" is the *transpose*.

We would not need to conduct a simulation to estimate $\mu$ if indeed we can calculate it directly as indicated above. As noted before, in intended applications of this planning approach, the actual model of interest tends to be more complicated than a CTMC, so that we cannot calculate the desired quantity $\mu$ directly. We introduce a CTMC that is similar to the more complicated model of interest, and use the CTMC analysis to get rough estimates of both $\mu$ and the required computational effort in order to estimate $\mu$ by simulation. We will illustrate for queueing models later.

To continue, the fundamental matrix $Z$ describes the time-dependent deviations from the stationary vector, in particular,

$$Z_{i,j} \equiv \int_0^\infty \left[P_{i,j}(t) - \pi_j\right] \mathrm{d}t.$$

Given the stationary probability vector, $\pi$, the fundamental matrix $Z$ can be found by first forming the square matrix $\Pi$, all of whose rows are the vector $\pi$, and then calculating

$$Z = (\Pi - Q)^{-1} - \Pi,$$

with the inverse always existing; again see Whitt (1992) and references therein. We now consider how the desired asymptotic parameters can be expressed in terms of the stationary probability vector $\pi$ and the fundamental matrix $Z$, including ways that are especially effective for computation.

### *3.2 Poisson's equation*

For that purpose, it is useful to introduce Poisson's equation. The stationary probability vector $\pi$ and the fundamental matrix $Z$ can be viewed as solutions $x$ to *Poisson's equation*

$$xQ = y,$$

for appropriate (row) vectors $y$. It can be shown that Poisson's equation has a solution $x$ if and only if $ye^{\mathrm{T}} = 0$, where $e$ is the row vector of 1's and $e^{\mathrm{T}}$ is its transpose. Then all solutions are of the form

$$x = -yZ + \left(xe^{\mathrm{T}}\right)\pi.$$

For example, $\pi$ is obtained by solving Poisson's equation when $y$ is the zero vector (and normalizing by requiring that $xe^{\mathrm{T}} = 1$). Then elements of $Z$ can be obtained by choosing other vectors $y$, requiring that $xe^{\mathrm{T}} = 0$.

In passing, we remark that there also is an alternate *column-vector form of Poisson's equation*, namely,

$$Qx^{\mathrm{T}} = y^{\mathrm{T}},$$

which has a solution if and only if $\pi y^{\mathrm{T}} = 0$. Then all solutions are of the form

$$x^{\mathrm{T}} = -Zy^{\mathrm{T}} + \bigl(\pi x^{\mathrm{T}}\bigr)e^{\mathrm{T}}.$$

It is significant that, for a CTMC, the asymptotic bias $\bar{\beta}$ defined in (2) and the asymptotic variance $\bar{\sigma}^2$ defined in (3) can be expressed directly in terms of $\pi$, $Z$, the function $f$ and (for $\beta$) the initial probability vector, say $\xi$, i.e.,

$$\bar{\beta}(\xi) = \xi Z f^{\mathrm{T}} \equiv \sum_{i=0}^{m}\sum_{j=0}^{m} \xi_i Z_{i,j} f_j$$

and

$$\bar{\sigma}^2 = 2\bigl(f\pi^{\mathrm{T}}\bigr)Zf^{\mathrm{T}} \equiv 2\sum_{i=0}^{m}\sum_{j=0}^{m} f_i \pi_i Z_{i,j} f_j.$$

Moreover, the asymptotic parameters $\bar{\beta}(\xi)$ and $\bar{\sigma}^2$ are themselves directly solutions to Poisson's equation. In particular,

$$\bar{\beta}(\xi) = xf^{\mathrm{T}},$$

where $x$ is the unique solution to Poisson's equation for $y = -\xi + \pi$ with $xe^{\mathrm{T}} = 0$. Similarly,

$$\bar{\sigma}^2 = 2xf^{\mathrm{T}},$$

where $x$ is the unique solution to Poisson's equation for $y_i = -(f_i - \mu)\pi_i$ with $xe^{\mathrm{T}} = 0$.

### *3.3 Birth-and-death processes*

*Birth-and-death (BD) processes* are special cases of CTMCs in which $Q_{i,j} = 0$ when $|i - j| > 1$; then we often use the notation $Q_{i,i+1} \equiv \lambda_i$ and $Q_{i,i-1} \equiv \mu_i$, and refer to $\lambda_i$ as the birth rates and $\mu_i$ as the death rates. For BD processes and skip-free CTMCs (which in one direction can go at most one step), Poisson's equation can be efficiently solved *recursively*.

To describe the recursive algorithm for BD processes, we start by observing that for a BD process Poisson's equation $xQ = y$ is equivalent to the system of equations

$$x_{j-1}\lambda_{j-1} - x_j(\lambda_j + \mu_j) + x_{j+1}\mu_{j+1} = y_j, \quad j \geqslant 0,$$

where $x_{-1} = x_{m+1} = 0$. Upon adding the first $j + 1$ equations, we obtain the desired recursive algorithm,

$$x_{j+1} = \frac{\lambda_j x_j + s_j}{\mu_{j+1}},$$

where

$$s_j = \sum_{i=0}^{j} y_i.$$

Hence, Poisson's equation for BD processes can indeed be solved recursively.

For BD processes and their continuous-time relatives – diffusion processes – the asymptotic parameters can be expressed directly as sums and integrals, respectively. For BD processes,

$$\bar{\beta}(\xi) = \sum_{j=0}^{m-1} \frac{1}{\lambda_j \pi_j} \sum_{i=0}^{j} (f_i - \mu)\pi_i \sum_{k=0}^{j} (\xi_k - \pi_k)$$

and

$$\bar{\sigma}^2 = 2 \sum_{j=0}^{m-1} \frac{1}{\lambda_j \pi_j} \left[ \sum_{i=0}^{j} (f_i - \mu)\pi_i \right]^2,$$

where, as for CTMCs, $\pi$ is the steady-state probability vector, while $\mu$ is the expected value of $f$ with respect to $\pi$. However, for BD processes, it is usually easier to use the recursive algorithm for computation. Indeed, the recursive algorithm for the asymptotic bias and the asymptotic variance parallels the well known recursive algorithm for the steady-state probability vector $\pi$.

## 4 Birth-and-death examples

In this section we consider examples of BD processes, primarily of queueing models. These examples show that the model structure can make a big difference in the computational effort required for estimation by simulation.

**Example 1** (The $M/M/1$ queue). Consider the queue-length (number in system, including the customer in service, if any) process $\{Q(t)\colon\ t \geqslant 0\}$ in the $M/M/1$ queue with unlimited waiting space. This model has a Poisson arrival process with constant rate and i.i.d. service times with an exponential distribution. The state space here is infinite, but the theory for the asymptotic parameters extends to this example. The queue-length process is a BD process with constant arrival (birth) rate and constant service (death) rate.

Let the service rate be 1 and let the arrival rate and traffic intensity be $\rho$. Fixing the service rate gives meaning to time in the simulation run length. Let $f(i) = i$ for all $i$, so that we are estimating the steady-state mean. The steady-state mean and variance are

$$\mu = \frac{\rho}{1-\rho} \quad \text{and} \quad \sigma^2 = \frac{\rho}{(1-\rho)^2};$$

e.g., see Cohen (1982).

To estimate the required simulation run length from a single long run, we use the asymptotic bias and the asymptotic variance. Let the system start out empty, so that the initial state is 0. As an argument of $\bar{\beta}(\xi)$, let 0 also denote the initial probability vector $\xi$ that puts mass 1 on the state 0. Then

$$\bar{\beta}(0) = \frac{-\rho}{(1-\rho)^3} \quad \text{and} \quad \bar{\sigma}^2 = \frac{2\rho(1+\rho)}{(1-\rho)^4}.$$

These formulas can be derived from the general BD formulas or directly; see Abate and Whitt (1987a, 1988a, 1988b).

Ignoring the initial transient (assuming that the queue-length process we observe is a stationary process), the required run length with a relative-width criterion, specified in general in (4), here is

$$t_r(\varepsilon, \alpha) = \frac{8(1+\rho)z^2_{\alpha/2}}{\rho(1-\rho)^2\varepsilon^2} \quad \text{and} \quad t_r\big(10^{-k}, 0.05\big) \approx \frac{32(1+\rho)(10)^{2k}}{\rho(1-\rho)^2}.$$

For 10% statistical precision ($\varepsilon = 0.1$) with 95% confidence intervals ($\alpha = 0.05$), when the traffic intensity is $\rho = 0.9$, the required run length is about 675,000 (mean service times, which corresponds to an expected number of arrivals equal to $0.9 \times 675{,}000 = 608{,}000$); when the traffic intensity is $\rho = 0.99$, the required run length is 64,300,000 (mean service times, which corresponds to an expected number of arrivals equal to $0.9 \times 64{,}300{,}000 = 57{,}900{,}000$). To summarize, for high traffic intensities, the required run length is of order $10^6$ or more mean service times. We can anticipate great computational difficulty as the traffic intensity $\rho$ increases toward the critical value for stability.

Compared to independent sampling of the steady-state queue length (which is typically not directly an option), the required run length is greater by a factor of

$$\frac{\bar{\sigma}^2}{\sigma^2} = \frac{2(1+\rho)}{\rho(1-\rho)^2},$$

which equals 422 when $\rho = 0.9$ and 40,200 when $\rho = 0.99$. Clearly, the dependence can make a great difference.

Now let us consider the bias. The relative bias is

$$\frac{\bar{\beta}_t(0)}{\mu} \approx \frac{\bar{\beta}(0)}{t\mu} = \frac{-1}{(1-\rho)^2 t},$$

so that, for $\rho = 0.9$ the relative bias starting empty is about $100/t$, where $t$ is the run length. For a run length of 675,000, the relative bias is $1.5 \times 10^{-4}$ or 0.015%, which is indeed negligible compared to the specified 10% relative width of the confidence interval. Hence the bias is in the noise; it can be ignored for high traffic intensities. The situation is even more extreme for higher traffic intensities such as $\rho = 0.99$.

**Example 2** (A small example with large asymptotic parameters). It is interesting to see that the asymptotic bias $\bar{\beta}(\xi)$ and the asymptotic variance $\bar{\sigma}^2$ can be arbitrarily large in a very small BD model with bounded rates. Suppose that $m = 2$, so that the BD process has only 3 states: 0, 1 and 2. Consider the symmetric model in which $\lambda_0 = \mu_2 = x$ and $\lambda_1 = \mu_1 = 1$, where $0 < x \leqslant 1$. Then the stationary distribution is

$$\pi_0 = \pi_2 = \frac{1}{2+x} \quad \text{and} \quad \pi_1 = \frac{x}{2+x}.$$

Let $f_i = i$ for all $i$, so that we are estimating the mean $\mu$. Then the mean is $\mu = 1$ and the asymptotic variance is

$$\bar{\sigma}^2 = \frac{4}{x(2+x)} \approx \frac{2}{x} \quad \text{for small } x.$$

This model has a high asymptotic variance $\bar{\sigma}^2$ for small $x$ because the model is *bistable*, tending to remain in the states 0 and 2 a long time before leaving. To see this, note that the mean first passage time from state 0 or state 2 to state 1 is $1/x$.

Note that the large asymptotic variance $\bar{\sigma}^2$ cannot be detected from the variance of the steady-state distribution, $\sigma^2$. As $x \downarrow 0$, $\sigma^2$, the variance of $\pi$, increases to 1. Thus, the ratio $\bar{\sigma}^2/\sigma^2$ is of order $\mathrm{O}(1/x)$. The steady-state distribution has moderate variance, but the process has quite strong dependence (but not so strong that the asymptotic variance becomes infinite).

The asymptotic bias starting in state 0 (or state 2) is also large for small $x$. The asymptotic bias starting in state 0 is

$$\bar{\beta}(0) = \frac{-(x+1)^2}{x(x+2)^2} \approx \frac{-1}{4x} \quad \text{for small } x.$$

As a function of the key model parameter $x$, the bias is much more important here than it was for the previous $M/M/1$ queue example. Here, *both* the asymptotic bias and the asymptotic variance are of order $\mathrm{O}(1/x)$, so that as a function of $x$, for very small $x$, the width of the confidence interval is $\mathrm{O}(1/\sqrt{x})$, while the bias is of order $\mathrm{O}(1/x)$. Thus the bias tends to be much more important in this example. In particular, the run length required to make the bias suitably small is of the *same order* as the run length required to make the width of a confidence interval suitably small. For this example, using simulation to estimate the mean $\mu$ when the parameter $x$ is very small would be difficult at best.

This model is clearly pathological: For very small $x$, a relatively long simulation run of this model starting in state 0 could yield a sample path that is identically zero. We might never experience even a single transition! This example demonstrates that it can be very helpful to know something about model structure when conducting a simulation.

**Example 3** (The $M/M/\infty$ queue). A queueing system with many servers tends to behave quite differently from a single-server queue. A queueing system with many servers can often be well approximated by an infinite-server queue. Thus we consider the number of busy servers at time $t$, also denoted by $Q(t)$, in an $M/M/\infty$ queue. As before, let the mean individual service time be 1, but now let the arrival rate be $\lambda$ (since the previous notion of traffic intensity is no longer appropriate). Now the arrival rate $\lambda$ can be arbitrarily large.

The first thing to notice is that as $\lambda$ increases, the required computational effort for given simulation run length in the simulation increases, simply because the expected number of arrivals in the interval $[0, t]$ is $\lambda t$. Thus, with many servers, we need to do a further adjustment to properly account for computational effort. To describe the computational effort, it is appropriate to multiply the time by $\lambda$. Thus, for the $M/M/\infty$ model with mean individual service rate 1, we let $c_r = \lambda t_r$ represent the required computational effort associated with the required run length $t_r$.

It is well known that the steady-state number of busy servers in the $M/M/\infty$ model, say $Q(\infty)$, has a Poisson distribution with mean $\lambda$; e.g., see Cooper (1982). Thus, the mean and variance of the steady-state distribution are

$$\mu \equiv \mathrm{E}\big[Q(\infty)\big] = \lambda \quad \text{and} \quad \sigma^2 \equiv \mathrm{Var}\big[Q(\infty)\big] = \lambda.$$

The asymptotic parameters also are relatively simple. As for the $M/M/1$ queue, we assume that we start with an empty system. Then the asymptotic bias and asymptotic variance are

$$\bar{\beta}(0) = -\lambda \quad \text{and} \quad \bar{\sigma}^2 = 2\lambda.$$

From the perspective of the asymptotic variance and relative error, we see that

$$\frac{\bar{\sigma}^2}{\mu^2} = \frac{2\lambda}{\lambda^2} = \frac{2}{\lambda},$$

so that simulation efficiency increases as $\lambda$ increases. However, the required computational effort to achieve relative $(1 - \alpha)\%$ confidence interval width of $\varepsilon$ is

$$c_r(\varepsilon, \alpha) \equiv \lambda t_r(\varepsilon, \alpha) = \frac{8z_\alpha^2}{\varepsilon^2}$$

which is *independent of* $\lambda$. Thus, from the perspective of the asymptotic variance, the required computational effort does not increase with the arrival rate, which is very different from the single-server queue.

Unfortunately, the situation is not so good for the relative bias. First, the key ratio is

$$\frac{\bar{\beta}(0)}{\mu} = \frac{-\lambda}{\lambda} = -1.$$

Thus the required run length to make the bias less than $\varepsilon$ is $1/\varepsilon$, and the required computational effort is $\lambda/\varepsilon$, which is *increasing in* $\lambda$. Unlike for the

$M/M/1$ queue, as the arrival rate $\lambda$ increases, the bias (starting empty) eventually becomes the dominant factor in the required computational effort.

For this $M/M/\infty$ model, it is natural to pay more attention to bias than we would with a single-server queue. A simple approach is to choose a different initial condition. The bias is substantially reduced if we start with a fixed number of busy servers not too different from the steady-state mean $\lambda$. Indeed, if we start with exactly $\lambda$ busy servers (assuming that $\lambda$ is an integer), then the bias is asymptotically negligible as $\lambda$ increases. Note, however, that this special initial condition does not directly put the system into steady state, because the steady-state distribution is Poisson, not deterministic.

If, instead, we were working with the $M/G/\infty$ model, then in addition we would need to specify the remaining service times of all the $\lambda$ customers initially in service at time 0. Fortunately, for the $M/G/\infty$ model, there is a natural way to do this: The steady-state distribution of the number of busy servers is again Poisson with mean $\lambda$, just as for the $M/M/\infty$ model. In addition, in steady-state, conditional upon the number of busy servers, the remaining service times of those customers in service are distributed as i.i.d. random variables with the *stationary-excess* (or equilibrium residual-life) cumulative distribution function (c.d.f.) $G_e$ associated with the service-time c.d.f. $G$, i.e.,

$$G_e(t) \equiv m^{-1} \int_0^t \left[1 - G(u)\right] \mathrm{d}u, \tag{5}$$

where $m$ is the mean of $G$ (here $m = 1$); e.g., see Takács (1962).

It is natural to apply this insight to more general many-server queueing models. Even in more general $G/G/s$ models, it is natural to initialize the simulation by putting $s$ customers in the system at time 0 and letting their remaining service times be distributed as $s$ i.i.d. random variables with c.d.f. $G_e$. For large $s$, that should be much better than starting the system empty.

For many-server queues, we may be interested in different congestion measures. By Little's law ($L = \lambda W$), we know that the expected steady-state number of busy servers in the $G/G/s/\infty$ model is exactly $\lambda$ (provided that $\lambda < s$). Thus, in simulation experiments, we usually are more interested in estimating quantities such as $\mathrm{E}[(Q(\infty) - s)^+]$, where $(x)^+ \equiv \max\{0, x\}$, or $\mathrm{P}(Q(\infty) > s + k)$. Note that we can calculate the asymptotic bias and the asymptotic variance for these quantities in the $M/M/s$ model by applying the BD recursion with appropriate functions $f$. With large $s$, it often helps to start the recursion at $s$ and move away in both directions. The initial value at $s$ can be taken to be 1; afterwards the correct value is obtained by choosing the appropriate normalizing constant.

## 5 Diffusion processes

*Diffusion processes* are continuous analogues of BD processes; e.g., see Karlin and Taylor (1981) and Browne and Whitt (1995). In this chapter we discuss diffusion processes because we are interested in them as approximations

of other processes that we might naturally simulate using discrete-event simulation. We want to use the diffusion processes to approximate the asymptotic bias and the asymptotic variance of sample means in the original process.

Diffusion processes tend to be complicated to simulate directly because they have continuous, continuously fluctuating, sample paths. Nevertheless, there also is great interest in simulating diffusion processes and stochastic differential equations, e.g., for finance applications, and special techniques have been developed; see Kloeden et al. (1994) and Kloeden and Platen (1995). Hence the analysis in this section may have broader application.

For diffusion processes, there are integral representations of the asymptotic parameters, paralleling the sums exhibited for BD processes. Corresponding to the finite-state-space assumption for the BD processes, assume that the state space of the diffusion is the finite interval $[s_1, s_2]$ and let the diffusion be reflecting at the boundary points $s_1$ and $s_2$, but under regularity conditions the integral representations will be valid over unbounded intervals. Let $\{Y(t)\colon\ t \geqslant 0\}$ be the diffusion process and let $X(t) = f(Y(t))$ for a real-valued function $f$. The diffusion process is characterized by its drift function $\mu(x)$ and its diffusion function $\sigma^2(x)$.

Let $\pi$ be the stationary probability density. The stationary probability density can be represented as

$$\pi(y) = \frac{m(y)}{M(s_2)}, \quad s_1 \leqslant y \leqslant s_2,$$

where

$$m(y) \equiv \frac{2}{\sigma^2(y)s(y)}$$

is the *speed density*,

$$s(y) \equiv \exp\left\{-\int_{s_1}^{y} \frac{2\mu(x)}{\sigma^2(x)}\,\mathrm{d}x\right\}$$

is the *scale density* and

$$M(y) = \int_{s_1}^{y} m(x)\,\mathrm{d}x, \quad s_1 \leqslant y \leqslant s_2,$$

provided that the integrals are finite.

Let $p(t, x, y)$ be the transition kernel. Then, paralleling the fundamental matrix of a CTMC, we can define the *fundamental function of a diffusion process*, $Z \equiv Z(x, y)$, by

$$Z(x, y) \equiv \int_0^{\infty} \big[p(t, x, y) - \pi(y)\big]\,\mathrm{d}t.$$

As before, let $\mu$ be the average of $f$ with respect to the stationary probability density $\pi$, i.e.,

$$\mu = \int_{s_1}^{s_2} \pi(x) f(x)\,\mathrm{d}x.$$

Then the integral representations for the asymptotic bias $\bar{\beta}(\xi)$ starting with initial probability density $\xi$ and the asymptotic variance $\bar{\sigma}^2$ are

$$\bar{\beta}(\xi) = 2\int_{s_1}^{s_2} \frac{1}{\sigma^2(y)\pi(y)} \times \left[\int_{s_1}^{y} \big(f(x)-\mu\big)\pi(x)\,\mathrm{d}x \int_{s_1}^{y} \big(\xi(z)-\pi(z)\big)\,\mathrm{d}z\right]\mathrm{d}y$$

and

$$\bar{\sigma}^2 = 4\int_{s_1}^{s_2} \frac{1}{\sigma^2(y)\pi(y)} \left[\int_{s_1}^{y} \big(f(x)-\mu\big)\pi(x)\,\mathrm{d}x\right]^2 \mathrm{d}y.$$

We now discuss two examples of diffusion processes, which are especially important because they arise as limit processes for queueing models, as we explain in Section 6.

**Example 4** (RBM). Suppose that the diffusion process is reflected Brownian motion (RBM) on the interval $[0,\infty)$ with drift function $\mu(x) = a$ and diffusion function $\sigma^2(x) = b$, where $a < 0 < b$, which we refer to by RBM$(a, b)$; see Harrison (1985), Whitt (2002) and references therein for more background. RBM is the continuous analog of the queue-length process for the $M/M/1$ queue (as we will explain in the next section). It is a relatively simple stochastic process with only the two parameters $a$ and $b$.

In fact, we can analyze the RBM$(a, b)$ processes by considering only the special case in which $a = -1$ and $b = 1$, which we call *canonical* RBM because there are *no free parameters*. We can analyze RBM$(a, b)$ in terms of RBM$(-1, 1)$ because we can relate the two RBMs by appropriately scaling time and space. For that purpose, let $\{R(t; a, b, X)\colon t \geqslant 0\}$ denote RBM$(a, b)$ with initial distribution according to the random variable $X$. The key relation between the general RBM and canonical RBM is

$$\big\{R(t; a, b, X)\colon t \geqslant 0\big\} \stackrel{\mathrm{d}}{=} \big\{c^{-1}R\big(d^{-1}t; -1, 1, cX\big)\colon t \geqslant 0\big\}$$

or, equivalently,

$$\big\{R(t; -1, 1, X)\colon t \geqslant 0\big\} \stackrel{\mathrm{d}}{=} \left\{cR\left(dt; a, b, \frac{X}{c}\right)\colon t \geqslant 0\right\},$$

where

$$c = \frac{|a|}{b}, \qquad d = \frac{b}{a^2}, \qquad a = \frac{-1}{cd} \quad \text{and} \quad b = \frac{1}{c^2 d},$$

where "$\stackrel{\mathrm{d}}{=}$" means equal in distribution (here as stochastic processes).

Hence it suffices to focus on canonical RBM. It has stationary density $\pi(x) = 2e^{-2x}$, $x \geqslant 0$. If we initialize RBM with its stationary distribution, then we obtain a stationary process. Let $R^* \equiv \{R^*(t; a, b)\colon\ t \geqslant 0\}$ denote *stationary* RBM, initialized by the stationary distribution.

If $f(x) = x$ for canonical RBM, then we would be estimating the steady-state mean $\mu = 1/2$. In this case, the asymptotic bias is $\bar{\beta}(0) = -1/4$ (Theorem 1.3 of Abate and Whitt, 1987a) and the asymptotic variance (for $R^*$) is $\bar{\sigma}^2 = 1/2$ (Abate and Whitt, 1988b).

To describe the general RBM with parameters $a$ and $b$, we apply the scaling relations in Section 6.1. As a consequence of those scaling properties, the mean and variance of the steady-state distribution of RBM$(a, b)$ are

$$\mu_{a,b} = \frac{b}{2|a|} \quad \text{and} \quad \sigma^2_{a,b} = \mu^2_{a,b} = \frac{b^2}{4a^2},$$

and the asymptotic parameters are

$$\bar{\beta}_{a,b}(0) = \frac{-b^2}{4|a|^3} \quad \text{and} \quad \bar{\sigma}^2_{a,b} = \frac{b^3}{2a^4}.$$

For the relative-width criterion, the key ratios are

$$\frac{\bar{\beta}_{a,b}(0)}{\mu_{a,b}} = \frac{-b}{2a^2} \quad \text{and} \quad \frac{\bar{\sigma}^2_{a,b}}{\mu^2_{a,b}} = \frac{2b}{a^2}.$$

Thus we see that the relative asymptotic bias is about the same as the relative asymptotic variance. Since the bias of the sample mean $\overline{X}_t$ is of order $\mathrm{O}(1/t)$, while the square root of the variance of the sample mean $\overline{X}_t$ is of order $\mathrm{O}(1/\sqrt{t}\,)$, the bias tends to be negligible for large $t$.

**Example 5** (OU). Suppose that the diffusion process is the Ornstein–Uhlenbeck (OU) diffusion process on the interval $(-\infty, \infty)$ with drift function $\mu(x) = ax$ and diffusion function $\sigma^2(x) = b$, where $a < 0 < b$, which we refer to as OU$(a, b)$. It is the continuous analog of the queue-length process in the $M/M/\infty$ queue when we center appropriately.

We also can analyze the OU$(a, b)$ processes by considering only the special case in which $a = -1$ and $b = 1$, which we call *canonical* OU. We can analyze OU$(a, b)$ in terms of OU$(-1, 1)$ because we can relate the two OUs by appropriately scaling time and space, just as we did for RBM. For that purpose, let $\{Z(t; a, b, X)\colon\ t \geqslant 0\}$ denote OU$(a, b)$ with initial distribution according to the random variable $X$. The key relation between the general OU$(a, b)$ and canonical OU$(-1, 1)$ is

$$\{Z(t; a, b, X)\colon\ t \geqslant 0\} \stackrel{\mathrm{d}}{=} \{c^{-1}Z(d^{-1}t; -1, 1, cX)\colon\ t \geqslant 0\}$$

or, equivalently,

$$\{Z(t; -1, 1, X): t \geqslant 0\} \stackrel{\mathrm{d}}{=} \left\{cZ\left(\mathrm{d}t; a, b, \frac{X}{c}\right): t \geqslant 0\right\},$$

where

$$c = \frac{|a|}{b}, \qquad d = \frac{b}{a^2}, \qquad a = \frac{-1}{cd} \quad \text{and} \quad b = \frac{1}{c^2 d}.$$

Then the stationary density of canonical OU is normal with mean 0 and variance 1/2. The mean of canonical OU starting at $x$ is

$$\mathrm{E}\big[Z(t; -1, 1, x)\big] = x\mathrm{e}^{-t}, \quad t \geqslant 0.$$

Paralleling our treatment of RBM, let $Z^* \equiv \{Z^*(t; a, b): t \geqslant 0\}$ be *stationary* OU, obtained by initializing the OU$(a, b)$ with the stationary normal distribution. For stationary canonical OU, the autocovariance function is

$$\mathrm{Cov}\big(Z^*(0), Z^*(t)\big) = \frac{1}{2}\mathrm{e}^{-t}, \quad t \geqslant 0.$$

Hence, the asymptotic parameters for canonical OU are

$$\bar{\beta}(\xi) = \xi \quad \text{and} \quad \bar{\sigma}^2 = \frac{1}{2}.$$

Just as with RBM, we can apply Section 6.1 to determine the effect of scaling. The mean and variance of the steady-state distribution of OU$(a, b)$ are

$$\mu_{a,b} = 0 \quad \text{and} \quad \sigma^2_{a,b} = \frac{b}{2|a|},$$

and the asymptotic parameters are

$$\bar{\beta}_{a,b}(x) = x\frac{b^2}{|a|^3} \quad \text{and} \quad \bar{\sigma}^2_{a,b} = \frac{b^3}{2a^4}.$$

The relative-width criterion makes less sense here because the random variables are not nonnegative.

## 6 Stochastic-process limits

In this section we discuss stochastic-process limits that make the RBM and OU diffusion processes serve as useful approximations for queueing models. We start by discussing the impact of scaling space and time. The scaling is often the key part.

### 6.1 Scaling of time and space

To obtain relatively simple approximate stochastic processes, we often consider stochastic-process limits, as in Whitt (2002). (We elaborate further.) To establish appropriate stochastic-process limits, we usually consider not just one stochastic process but a family of stochastic processes constructed by scaling time and space. It is thus important to know how the asymptotic parameters change under such scaling.

Suppose that we have a stochastic process $Z \equiv \{Z(t): t \geqslant 0\}$ and we want to consider the scaled stochastic process $Z_{u,v} \equiv \{Z_{u,v}(t): t \geqslant 0\}$, where

$$Z_{u,v}(t) \equiv uZ(vt), \quad t \geqslant 0,$$

for positive real numbers $u$ and $v$. Suppose that $Z(t) \Rightarrow Z(\infty)$ as $t \to \infty$. Then $Z_{u,v}(t) \Rightarrow Z_{u,v}(\infty)$ as $t \to \infty$, where

$$Z_{u,v}(\infty) = uZ(\infty).$$

Let $\mu$ be the mean and $\sigma^2$ the variance of $Z(\infty)$; let $\mu_{u,v}$ be the mean and $\sigma^2_{u,v}$ the variance of $Z_{u,v}(\infty)$. Then

$$\mu_{u,v} = u\mu \quad \text{and} \quad \sigma^2_{u,v} = u^2\sigma^2.$$

The relation is different for the asymptotic parameters: Observe that $\mathrm{E}Z_{u,v}(t) = u\,\mathrm{E}Z(vt)$ for $t \geqslant 0$ and, under the assumption that $Z$ is a stationary process,

$$\mathrm{Cov}\big(Z_{u,v}(0), Z_{u,v}(t)\big) = u^2\,\mathrm{Cov}\big(Z(0), Z(vt)\big), \quad t \geqslant 0.$$

As a consequence, the asymptotic bias and the asymptotic variance are

$$\bar{\beta}_{u,v} = \frac{u}{v}\bar{\beta} \quad \text{and} \quad \bar{\sigma}^2_{u,v} = \frac{u^2}{v}\bar{\sigma}^2.$$

Thus, once we have determined the asymptotic parameters of a stochastic process of interest, it is easy to obtain the asymptotic parameters of associated stochastic processes constructed by scaling time and space. If the scaling parameters $u$ and $v$ are either very large or very small, then the scaling can have a great impact on the required run length. Indeed, as we show below, in standard queueing examples the scaling is dominant.

### 6.2 RBM approximations

Consider the queue-length (number in system) stochastic process $\{Q_\rho(t): t \geqslant 0\}$ in the $G/G/s/\infty$ with traffic intensity (rate in divided by maximum rate out) $\rho$, with time units fixed by letting the mean service time be 1, without the usual independence assumptions. As reviewed in Whitt (1989, 2002), in remarkable generality (under independence assumptions and beyond), there is a *heavy-traffic stochastic-process limit* for the scaled queue-length processes,

obtained by dividing time $t$ by $(1-\rho)^2$ and multiplying space by $(1-\rho)$, i.e.,

$$\{(1-\rho)Q_\rho(t(1-\rho)^{-2}):\ t \geqslant 0\} \Rightarrow \{R(t;a,b):\ t \geqslant 0\} \quad \text{as } \rho \uparrow 1$$

for appropriate parameters $a$ and $b$, where $\{R(t;a,b):\ t \geqslant 0\}$ is RBM$(a,b)$ and again "$\Rightarrow$" denotes convergence in distribution, but here in the function space $D$ containing all sample paths.

The limit above is very helpful, because the number of relevant parameters has been greatly reduced. We see that the queue behavior for large $\rho$ should primarily depend upon only $\rho$ and the two parameters $a$ and $b$. Moreover, it turns out the parameters $a$ and $b$ above can be conveniently characterized (in terms of scaling constants in central limit theorems for the arrival and service processes). For example, in the standard $GI/GI/s/\infty$ model the heavy-traffic limit holds with

$$a = -s \quad \text{and} \quad b = s(c_a^2 + c_s^2),$$

where $c_a^2$ and $c_s^2$ are the SCVs of an interarrival time and a service time, respectively (provided that the second moments are finite). Similar limits hold for workload processes, recording the amount of remaining unfinished work in service time in the system.

We thus can apply the stochastic-process limit with the scaling properties in Section 6.1 and the properties of RBM to obtain approximations paralleling the exact results for the $M/M/1$ queue. We apply the stochastic-process limit to obtain the approximation

$$\{Q_\rho(t):\ t \geqslant 0\} \approx \{(1-\rho)^{-1}R(t(1-\rho)^2;a,b):\ t \geqslant 0\}.$$

The resulting approximations for the mean and variance of the steady-state distribution of the queue-length process are thus

$$\mathrm{E}[Q_\rho(\infty)] \approx \frac{b}{2|a|(1-\rho)} \quad \text{and} \quad \sigma_\rho^2 \equiv \mathrm{Var}(Q_\rho(\infty)) \approx \frac{b^2}{4a^2(1-\rho)^2};$$

the approximations for the asymptotic parameters are

$$\bar{\beta}_\rho(0) \approx \frac{-b^2}{4|a|^3(1-\rho)^3} \quad \text{and} \quad \bar{\sigma}_\rho^2 \approx \frac{b^3}{2a^4(1-\rho)^4}.$$

In the $GI/GI/s/\infty$ case, we just substitute the specific parameters $a$ and $b$ above. The resulting approximate asymptotic variance is

$$\bar{\sigma}_\rho^2 \equiv \bar{\sigma}^2_{(s,\rho,c_a^2,c_s^2)} = \frac{(c_a^2+c_s^2)^3}{2s(1-\rho)^4}.$$

Note that these formulas agree with the limits of the $M/M/1$ formulas as $\rho \uparrow 1$. Thus, we see that the $M/M/1$ formulas are remarkably descriptive more generally. But we also see the impact of $s$ servers and the $GI$ arrival and service processes: The asymptotic variance is directly proportional to $1/s$ and to the

*third power* of the overall "variability parameter" $(c_a^2 + c_s^2)$ as well as to the *fourth* power of $(1-\rho)^{-1}$.

More generally, we see how the parameters $s$, $\rho$, $a$ and $b$ in more general $G/G/s/\infty$ models (with nonrenewal arrival processes and non-i.i.d. service times) will affect the required simulation run length. Once we have established the corresponding heavy-traffic limit and identified the new values of $a$ and $b$ for these alternative models, we can apply the results above. For the relative-width criterion, the key ratios are

$$\frac{\bar{\beta}_{a,b}(0)}{\mu_{a,b}} \approx \frac{-b}{2a^2(1-\rho)^2} \quad \text{and} \quad \frac{\bar{\sigma}^2_{a,b}}{\mu^2_{a,b}} \approx \frac{2b}{a^2(1-\rho)^2}.$$

Values of the key parameters $a$ and $b$ in alternative models have been determined; e.g., see Sections 5.2–5.5 of Whitt (1989) and Fendick et al. (1989).

### *6.3 Many-server queues*

The analysis above applies to multiserver queues, but when the number of servers is large, the RBM approximation tends to be inappropriate. When the number of servers is large, it is often preferable to consider different limits in which the number $s$ of servers is allowed to increase as the arrival rate $\lambda$ increases; see Halfin and Whitt (1981), Chapter 10 of Whitt (2002), Whitt (2005) and references therein. Alternatively, when there is a large number of servers, a more elementary direct approach is to consider an infinite-server model as an approximation for the model with finitely many servers. We thus might approximate the queue-length (number in system) process in the $G/G/s$ model (with finite or infinite waiting room) by the stochastic process representing the number of busy servers in the associated $G/G/\infty$ model.

When we do consider the $G/G/\infty$ model, it is natural to develop approximations based on heavy-traffic stochastic-process limits, where now heavy-traffic is defined by having the arrival rate $\lambda$ increase without bound. For that purpose, let $Q_\lambda(t)$ denote the number of busy servers at time $t$ as a function of the arrival rate $\lambda$. Again, under regularity conditions, there is a heavy-traffic stochastic-process limit, but it takes a different form. Now

$$\left\{\frac{Q_\lambda(t)-\lambda}{\sqrt{\lambda}}\colon\ t \geqslant 0\right\} \Rightarrow \{L(t)\colon\ t \geqslant 0\} \quad \text{as } \lambda \to \infty,$$

where the limit process $L \equiv \{L(t)\colon\ t \geqslant 0\}$ is a zero-mean Gaussian process, i.e., for which $(L(t_1), \ldots, L(t_k))$ has a $k$-dimensional normal distribution for all positive integers $k$ and all time points $0 < t_1 < \cdots < t_k$; see Section 10.3 of Whitt (2002), Glynn and Whitt (1991) and references therein. As with the previous RBM limit, this limit generates a natural approximation; here it is

$$\{Q(t)\colon\ t \geqslant 0\} \approx \{\lambda + \sqrt{\lambda}L(t)\colon\ t \geqslant 0\}.$$

For the $G/M/\infty$ special case, when the service times are i.i.d. and exponentially distributed (still with mean 1), the Gaussian limit process $L$ is $\mathrm{OU}(-1, 1 + c_a^2)$, where $c_a^2$ is the normalization constant in the central limit theorem for the arrival process, corresponding to the SCV of an interarrival time in the renewal ($GI$) case. Thus the approximate asymptotic variance is

$$\bar{\sigma}^2 \equiv \bar{\sigma}^2_{Q_\lambda} \approx \lambda \frac{(1 + c_a^2)^3}{2}.$$

For the more general $G/GI/\infty$ model, when the service times are *not* exponentially distributed, the limiting Gaussian process is *not* Markov (except if $G$ is a mixture of an exponential and a point mass at 0; see Glynn (1982)). If $G$ denotes the c.d.f. of the service time and $G^{\mathrm{c}}(t) \equiv 1 - G(t)$ is the associated complementary c.d.f. (c.c.d.f.), then the autocovariance function of the stationary version $L^*$ of the limit process $L$ is

$$\begin{aligned}&\mathrm{Cov}\big(L^*(0), L^*(t)\big)\\ &\quad = \int_0^\infty G(u)G^{\mathrm{c}}(t+u)\,\mathrm{d}u + c_a^2 \int_0^\infty G^{\mathrm{c}}(t+u)G^{\mathrm{c}}(u)\,\mathrm{d}u.\end{aligned}$$

In the special case of the $M/GI/\infty$ model, $c_a^2 = 1$ and the autocovariance function simplifies, becoming

$$\mathrm{Cov}\big(L^*(0), L^*(t)\big) = \int_0^\infty G^{\mathrm{c}}(t+u)\,\mathrm{d}u = G_{\mathrm{e}}^{\mathrm{c}}(t),$$

where $G_{\mathrm{e}}^{\mathrm{c}}$ is c.c.d.f. associated with the stationary-excess c.d.f. $G_{\mathrm{e}}$ in (5). Since $G_{\mathrm{e}}$ has mean $(c_s^2 + 1)/2$, the asymptotic variance of $L^*$ is $(c_s^2 + 1)/2$. Thus, for the $M/GI/\infty$ model the approximate asymptotic variance of $Q_\lambda$ is

$$\bar{\sigma}^2 \equiv \bar{\sigma}^2_{Q_\lambda} \approx \frac{\lambda(c_s^2 + 1)}{2}.$$

With many-server queues, we are often less interested in the queue-length process than other stochastic processes. For example in the $M/M/s/0$ (Erlang loss or B) model, which has $s$ servers, no extra waiting room and arrivals blocked or lost without affecting future arrivals when arrivals find all servers busy, instead of the number of busy servers, we are often interested in the steady-state blocking probability. In the corresponding $M/M/s/\infty$ (Erlang delay or C) model, which has $s$ servers and unlimited extra waiting room, we are often interested in the steady-state delay probability, i.e., the probability that an arrival must wait before beginning service, or the probability that an arrival must have to wait more than some designated time, such as 20 seconds (a common target in telephone call centers).

To consider the simulation run length required to estimate these alternative characteristics, we may nevertheless use the infinite-server model and the analysis above as a rough guide. To get better estimates we can consider the multi-server model with $s$ servers (instead of letting $s = \infty$). It has been found

useful to consider limits in which $s$ increases along with $\lambda$. It turns out to be appropriate to let $s$ and $\lambda$ increase so that

$$\frac{\lambda - s}{\sqrt{\lambda}} \to \gamma \quad \text{as } \lambda \to \infty.$$

Then, just as in the infinite-server case, under regularity conditions there is again a heavy-traffic limit for $[Q_\lambda(t) - \lambda]/\sqrt{\lambda}$ but now with a different limit process $L$; see Halfin and Whitt (1981), Srikant and Whitt (1996, 1999), Puhalskii and Reiman (2000) and Whitt (2005). That in turn allows us to approximate the asymptotic variance and estimate the required simulation run length. The issue of required simulation run lengths for many-server loss systems is the main focus of Srikant and Whitt (1996, 1999).

## 7 Deleting an initial portion of the run to reduce bias

We have seen that for various Markov processes we can estimate the bias of a sample mean associated with any contemplated initial distribution and simulation run length. If the estimated bias is too large, then we can try to reduce the bias by choosing alternative initial conditions. We can estimate the bias reduction gained by choosing alternative initial distributions, because the asymptotic bias $\bar{\beta}(\xi)$ is a function of the initial probability distribution $\xi$.

If the estimated bias is too large, and it is difficult to change the initial conditions, then we might instead consider not collecting data for an initial portion of the simulation run, given the natural initial conditions. However, it is more difficult to estimate the bias reduction from not collecting data from an initial portion of the run. For that purpose, we need to know the time-dependent mean $\mathrm{E}[X(t)]$, where $\{X(t)\colon\ t \geqslant 0\}$ is the stochastic process being observed. The asymptotic bias when we do not collect data over an initial interval $[0, c]$ is

$$\bar{\beta}(\xi, c) = \int_c^\infty \big(\mathrm{E}X(t) - \mu\big)\,\mathrm{d}t.$$

A rough approximation for the asymptotic bias $\bar{\beta}(\xi, c)$ can be based on the exponential approximation

$$\mathrm{E}\big[X(t)\big] - \mu \approx \mathrm{e}^{-t/\bar{\beta}}, \quad t \geqslant 0,$$

where the parameter $\bar{\beta}$ is chosen to yield the correct asymptotic bias $\bar{\beta} = \bar{\beta}(\xi, 0)$. Then we obtain the approximation

$$\bar{\beta}(\xi, c) \approx \int_c^\infty \mathrm{e}^{-t/\bar{\beta}}\,\mathrm{d}t = \bar{\beta}\mathrm{e}^{-c/\bar{\beta}}.$$

Unfortunately, however, the exponential approximation is not very reliable, because the time-dependent mean rarely has such a simple exponential

form. For better estimates of the reduced bias, we need to estimate the time-dependent mean $\mathrm{E}X(t)$. Fortunately, for some commonly occurring stochastic processes, expressions for the time-dependent mean are available. For example, exact and approximate expressions for the time-dependent mean for $M/M/1$ and RBM are contained in Abate and Whitt (1987a, 1987b, 1988b).

For the $M/GI/\infty$ model with arrival rate $\lambda$ and service-time c.d.f. $G$, starting empty,

$$\mathrm{E}\big[Q_\lambda(t)\big] = \mathrm{E}\big[Q_\lambda(\infty)\big]G_\mathrm{e}(t) = \lambda G_\mathrm{e}(t), \quad t \geqslant 0,$$

where $G_\mathrm{e}$ is the stationary-excess c.d.f., just as in the covariance function; see Section 4 of Eick et al. (1993). So the asymptotic bias is $-\lambda(c_s^2+1)/2$, just like the asymptotic variance.

## 8 Directions for further research

We described two classes of models that have been analyzed rather thoroughly to understand the required simulation run lengths: single-server queues and many-server queues (here approximated by infinite-server queues). Other important classes of stochastic models should be analyzed in the same way.

The analysis here is based on the normal approximation for the sample mean reviewed in Section 2. The conditions in the central limit theorem yielding the normal approximation are not satisfied when there are heavy-tailed distributions or long-range dependence. Since these features tend to arise in practice, it is natural to include them in simulations. As can be seen from Chapter 4 of Whitt (2002), there is alternative asymptotic theory for heavy-tailed distributions or long-range dependence, and there is a body of statistical techniques, as in Beran (1994), but more needs to be done to plan simulations in that context. In general, simulations in face of such stochastic complexity are difficult. Work to cope with such complexity is described in Chapter 11.

## Acknowledgement

The author was supported by National Science Foundation Grant DMS-02-2340.

## References

Abate, J., Whitt, W. (1987a). Transient behavior of regulated Brownian motion, I and II. *Advances in Applied Probability* 19, 560–631.

Abate, J., Whitt, W. (1987b). Transient behavior of the $M/M/1$ queue: Starting at the origin. *Queueing Systems* 2, 41–65.

Abate, J., Whitt, W. (1988a). Transient behavior of the $M/M/1$ queue via Laplace transforms. *Advances in Applied Probability* 20, 145–178.

Abate, J., Whitt, W. (1988b). The correlation functions of $RBM$ and $M/M/1$. *Stochastic Models* 4, 315–359.
Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edition. Springer-Verlag, New York.
Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman and Hall, London.
Bratley, P., Fox, B.L., Schrage, L.E. (1987). *A Guide to Simulation*, 2nd edition. Springer-Verlag, New York.
Browne, S., Whitt, W. (1995). Piecewise-linear diffusion processes. In: Dshalalow, J. (Ed.), *Advances in Queueing*. CRC Press, Boca Raton, FL, pp. 463–480.
Chen, H., Yao, D.D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*. Springer-Verlag, New York.
Cochran, W.G., Cox, G.M. (1992). *Experimental Designs*, 2nd edition. Wiley, New York.
Cohen, J.W. (1982). *The Single Server Queue*, 2nd edition. North-Holland, Amsterdam.
Cooper, R.B. (1982). *Introduction to Queueing Theory*, 2nd edition. North-Holland, New York.
Eick, S.G., Massey, W.A., Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research* 41, 731–742.
Fendick, K.W., Saksena, V.R., Whitt, W. (1989). Dependence in packet queues. *IEEE Transactions on Communications* 37, 1173–1183.
Fishman, G.S. (2001). *Discrete Event Simulation*. Springer-Verlag, New York.
Glynn, P.W. (1982). On the Markov property of the $GI/G/\infty$ Gaussian limit. *Advances in Applied Probability* 14, 191–194.
Glynn, P.W. (1984). Some asymptotic formulas for Markov chains with application to simulation. *Journal of Statistical Computation and Simulation* 19, 97–112.
Glynn, P.W. (1989). A GSMP formalism for discrete-event systems. *Proceedings of the IEEE* 77, 14–23.
Glynn, P.W. (1994). Poisson's equation for the recurrent $M/G/1$ queue. *Advances in Applied Probability* 26, 1044–1062.
Glynn, P.W., Meyn, S. (1996). A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability* 24, 916–931.
Glynn, P.W., Whitt, W. (1991). A new view of the heavy-traffic limit for infinite-server queues. *Advances in Applied Probability* 23, 188–209.
Glynn, P.W., Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research* 40, 505–520.
Grassman, W.K. (1987a). The asymptotic variance of a time average for a birth–death process. *Annals of Operations Research* 8, 165–174.
Grassman, W.K. (1987b). Means and variances of time averages in Markovian environments. *European Journal of Operations Research* 31, 132–139.
Halfin, S., Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29, 567–588.
Harrison, J.M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.
Karlin, S., Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, San Diego, CA.
Kelly, F.P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.
Kloeden, P.E., Platen, E. (1995). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin.
Kloeden, P.E., Platen, E., Schurz, H. (1994). *Numerical Solution of SDE Through Computer Experiments*. Springer-Verlag, Berlin.
Lehmann, E.L., Casella, G. (1998). *Theory of Point Estimation*, 2nd edition. Wiley, New York.
Montgomery, D.C. (2000). *Design and Analysis of Experiments*, 5th edition. Wiley, New York.
NIST/SEMATECH (2003). Exploratory data analysis. In: *e-Handbook of Statistical Methods*, Chapter 1. Available at *http://www.itl.nist.gov/div898/handbook/*.
Papadimitriou, C.H. (1994). *Computational Complexity*. Addison–Wesley, Reading, MA.
Puhalskii, A.A., Reiman, M.I. (2000). The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. *Advances in Applied Probability* 32, 564–595.
Samorodnitsky, G., Taqqu, M.S. (1994). *Stable Non-Gaussian Random Processes*. Chapman and Hall, New York.

Sigman, K. (1995). *Stationary Marked Point Processes, An Intuitive Approach*. Chapman and Hall, New York.

Srikant, R., Whitt, W. (1996). Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation* 6, 7–52.

Srikant, R., Whitt, W. (1999). Variance reduction in simulations of loss models. *Operations Research* 47, 509–523.

Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press, New York.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison–Wesley, Reading, MA.

Velleman, P., Hoaglin, D. (1981). *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury, Belmont, CA.

Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ.

Whitt, W. (1989). Planning queueing simulations. *Management Science* 35, 1341–1366.

Whitt, W. (1992). Asymptotic formulas for Markov processes. *Operations Research* 40, 279–291.

Whitt, W. (2002). *Stochastic-Process Limits*. Springer-Verlag, New York.

Whitt, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research* 30 (1), 1–27.

Wu, C.F.J., Hamada, M.S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.