INVITED PAPER

SIMULATION METHODS FOR QUEUES: AN OVERVIEW

Peter W. GLYNN [1] and Donald L. IGLEHART [2]

Department of Operations Research, Stanford University, Stanford, CA 94305, U.S.A.

Received 21 April 1988

Abstract

This paper gives an overview of those aspects of simulation methodology that are (to some extent) peculiar to the simulation of queueing systems. A generalized semi-Markov process framework for describing queueing systems is used through much of the paper. The main topics covered are: output analysis for simulation of transient and steady-state quantities, variance reduction methods that exploit queueing structure, and gradient estimation methods for performance parameters associated with queueing networks.

Keywords: Simulation, queueing theory, output analysis, variance reduction, generalized semi-Markov processes, gradient estimation.

1. Introduction

This paper is intended to give the reader an overview of those aspects of simulation methodology that exploit (to some degree) the stochastic structure of queueing systems. As a consequence, certain more broadly based methodologies are not discussed here; see chapters 2, 3, and 8 of Bratley, Fox and Schrage [6] for a more complete picture of the current state of simulation methodology.

In section 2, we argue that simulation is an important numerical tool for the study of complex queueing systems. Section 3 develops the generalized semi-Markov process view of queueing systems, which runs as a unifying thread through much of this paper. In section 4, we describe the basic methodology for the analysis of simulation output associated with estimation of transient quantities. Section 5 describes the analogues of these methods for the steady-state estimation problem. The key role of general state space Markov chain theory in the analysis of steady-state output is also described there.

Section 6 describes some recent work on exploiting diffusion approximations for queues in heavy traffic to develop insight into the question of how the

[1] The research of this author was supported by the U.S. Army Research Office under Contract DAAG29-84-K-0030.

[2] The research of this author was supported by the U.S. Army Research Office under Contract DAAG-84-K-0030 and National Science Foundation Grant DCR-85-09668.

simulation run-length required to estimate steady-state quantities increases as the traffic intensity converges to unity. In section 7, the ideas of variance reduction and efficiency improvement are introduced. These methods permit the simulator to improve upon conventional estimation methods by incorporating knowledge of the stochastic system under consideration into the estimator. Sections 8 through 14 describe seven such techniques that exploit queueing-related stochastic structure. Finally, in section 15 two powerful new methods (based on perturbation analysis and likelihood ratios) for estimating gradients of queueing performance parameters are discussed.

2. Simulation as a numerical tool

Simulation is a powerful tool for studying complex queueing systems. Its popularity derives from the following factors:

1) The basic idea underlying simulation is conceptually simple to understand. As a result, it is a tool that is accessible to a wide range of users, including those without a strong background in the theory of stochastic processes.
2) By using simulation in conjunction with sophisticated graphics, one can observe the evolution of sample trajectories of the system. This can provide an insight into the system that is unavailable from a purely numerical study, in which the focus would be on computation of expectations of various performance measures.
3) Conventional numerical approaches often rely on solving specifically structured systems of equations. The appropriate system of equations differs from one type of expectation to another. For example, the method used to solve for the steady-state of a Markov chain is quite different from that used to calculate the transient probabilities. On the other hand, simulation offers a methodology which can calculate expectations of arbitrary functionals of the system, without any major change in the basic approach.
4) Conventional numerical techniques are generally oriented to the solution of stochastic systems having rather special probabilistic structure (e.g. discrete-time and continuous-time Markov chains). By contrast, simulation is a general-purpose tool for which the study of a general discrete-event system requires a programming effort comparable to its Markovian counterpart (in which, for example, all inter-arrival and service time distributions are assumed exponential).

As the above points indicate, a principal advantage of Monte Carlo simulation is that implementation of such algorithms by practitioners is relatively straightforward, thereby saving time and (possibly) money. In addition, there are often sound computational reasons for choosing a simulation algorithm over competing numerical approaches. We will illustrate this point with an example.

Suppose that one is interested in the transient behavior of an irreducible closed Jackson network, containing n customers, with s_i servers at the i'th station, $1 \leqslant i \leqslant d$. The state space S of the corresponding continuous-time Markov chain $X = (X(t)\colon t \geqslant 0)$ is the set of d-tuples $\{(k_1, \dots, k_d)\colon k_i \in \mathbb{Z}^+, k_1 + \dots + k_d = n\}$, where $\mathbb{Z}^+$ is the set of non-negative integers. Note that

$$|S| = \binom{n+d-1}{n}$$

so that the size of the state space S grows rapidly with n.

If one is interested in calculating $\alpha = E_\eta f(X(t))$ for $f\colon \mathbb{R}^d \to \mathbb{R}$ ($E_\eta(\cdot)$ denotes the expectation operator conditional on $X(0)$ having distribution η), a conventional numerical approach would typically rely on the fact that

$$\alpha = \eta^t \exp(At) f \tag{2.1}$$

where A is the generator of X. (We assume throughout this paper that all vectors are encoded as column vectors.) It is clear that the complexity of a conventional numerical algorithm for calculating (2.1) will increase (rapidly) with n, since such algorithms are highly sensitive to $|S|$.

For example, it is well known that the right-hand side of (2.1) can be expressed as

$$\eta^t \exp(At) f = \eta^t \sum_{m=0}^{\infty} \frac{A^m}{m!} t^m f.$$

One can recursively compute the products $g_m = A^m t^m f/m!$ via $g_m = A g_{m-1} \cdot t/m$; each of these matrix-vector multiplications takes on the order of $|S|^2$ operations. This analysis suggests that the complexity of this matrix-vector multiplication algorithm increases rapidly with n.

To be more precise, assume $\|f\| \leqslant 1$ where $\|f\| \equiv \max\{|f(x)|\colon x \in S\}$. If $\mu_1, \dots, \mu_d$ are the service rates at the d stations, note that

$$\begin{aligned}\|A\| &\equiv \max\left\{\sum_{y \in S} |A_{xy}|\colon x \in S\right\}\\ &\leqslant 2 \sum_{j=1}^{d} \mu_j s_j\end{aligned}$$

is independent of n. Thus, for a given error tolerance ϵ, one can select $M(\epsilon)$ (independently of n) such that

$$\left|\sum_{m > M(\epsilon)} \eta^t A^m t^m f/m!\right| < \epsilon.$$

Hence, the complexity to calculate $E_\eta f(X(t))$ to precision ϵ, via the above algorithm, for f's satisfying $\|f\| \leqslant 1$, is the number of operations required to calculate

$$\eta^t \sum_{m \leqslant M(\epsilon)} A^m t^m f/m!,$$

which is, of course, of order n^{2d-2}. (Recall that $|S|$ is of order n^{d-1}.) Thus, for a system having a large number of customers n (or stations d), this approach becomes infeasible.

On the other hand, Monte Carlo simulation is relatively insensitive to the size of the state space. Let $X_1, X_2, \ldots$ be i.i.d. copies of the process X and consider the estimator

$$\alpha_m = \frac{1}{m} \sum_{j=1}^{m} f(X_j(t)).$$

If $\| f \| \leqslant 1$, Chebyshev's inequality implies that

$$P\{ | \alpha_m - \alpha | > \epsilon \} \leqslant \frac{1}{\epsilon^2 m}.$$

Thus, the sample size m required to obtain ϵ-precision with probability $1 - \delta$ is independent of $|S|$. Furthermore, we claim that the number of observations needed to generate α_m is insensitive to $|S|$. To see this, note that X can be uniformized via a Poisson process $N(\cdot)$ having rate $\Sigma \mu_j s_j$ ($|A_{xx}| \leqslant \Sigma \mu_j s_j$ for all $x \in S$). Since the transition epochs of X form a subsequence of the jump times of N, it follows that the number of jumps of X up to t is dominated by $N(t)$. But $EN(t) = \Sigma_{j=1}^{d} \mu_j s_j \cdot t$, which is independent of n. We conclude that Monte Carlo simulation's efficiency is relatively insensitive to the magnitude of n.

The above analysis suggests the following rule of thumb. As the state space of a queueing system gets larger and larger, simulation becomes increasingly more competitive as a computational tool.

An additional attractive feature of Monte Carlo simulation, which is important computationally, is that the error analysis for Monte Carlo algorithms is generally straightforward. For example, the typical Monte Carlo approach to calculating a parameter α which can be expressed as the mean of a r.v. X (i.e., $\alpha = EX$) is to generate i.i.d. replicates $X_1, X_2, \ldots$ of the r.v. X and to use $\overline{X}(n) = n^{-1} \Sigma_{i=1}^{n} X_i$ as an estimator of α. The error in $\overline{X}(n)$ is usually assessed via a confidence interval. In this setting, if $EX^2 < \infty$, a simulator can simply use the interval

$$\left[\overline{X}(n) - z(\delta) \frac{s(n)}{\sqrt{n}},\ \overline{X}(n) + z(\delta) \frac{s(n)}{\sqrt{n}} \right] \tag{2.2}$$

as a $100(1-\delta)\%$ confidence interval for α. (In (2.2), $z(\delta)$ is the root of the equation $P\{ N(0, 1) \leqslant z(\delta) \} = 1 - \delta/2$ and $s(n)$ is the sample standard deviation.) Since the central limit theorem justifies the asymptotic validity of (2.2) as a $100(1-\delta)\%$ confidence interval for α, we see that (2.2) provides an easily calculated and asymptotically precise error assessment for the estimator $\overline{X}(n)$. Error assessments (even in the form of upper bounds on the error) are typically quite hard to calculate for conventional numerical schemes.

3. The role of GSMP's in queueing simulations

In order to describe the simulation of queueing systems, we shall find it convenient to use the formalism of generalized semi-Markov processes (GSMP's). GSMP's form a class of stochastic processes that succinctly describe the essential probabilistic features of queueing systems. A GSMP is characterized by:

S: a "physical" state space which is finite or countable (typically, S will be the set of all possible queue-length vectors)

$E(s)$: the set of events which can occur in $s \in S$ ($E(s)$ will usually correspond to the different arrival and departure events possible when the system is in state s)

$p(s'; s, e)$: the probability of jumping from s to s', given that event e triggers the transition from s (e.g. e might correspond to station i completing service, in which case $p(s'; s, e)$ might represent the probability of sending a customer from station i to station j; here $s' = w - e_i + e_j$ where e_i is the i'th unit vector

r_{se}: the rate at which the clock corresponding to event e runs down to zero in state s (e.g. in a queueing network, r_{se} might be unity except for events e which are "interrupted" in state s, in which case $r_{se} = 0$; such "interrupted" events occur in the modeling of pre-emptive resume queueing priorities)

$F(\cdot; s', e', s, e)$ the probability distribution which schedules a new event e' in state s', given that the previous state was s and the transition was triggered by e (e.g. these would typically be service and interarrival time distributions in a queueing network). (3.1)

We will now illustrate the GSMP modeling formalism by describing a general service time/inter-arrival time version of an open Jackson network. We assume that the network has d stations, each station containing one server. Specifically, each server uses a first come/first serve queueing discipline; the service times for the consecutive customers served at the i'th station are i.i.d. with common continuous distribution G_i. The external arrival stream to the i'th station is assumed to be a renewal process with continuous inter-arrival distribution F_i. Furthermore, customers are routed between stations according to a Markovian routing scheme with associated substochastic routing matrix $P = (P_{ij}\colon 1 \leqslant i, j \leqslant d)$. Finally, the routing, inter-arrival, and service time sequences described above are all assumed independent of one another.

Given such a network, we obtain a GSMP by letting $S = \mathbb{Z}^+ \times \ldots \times \mathbb{Z}^+$ (d times); the vector $s = (s_1, \ldots, s_d) \in S$ will then represent the queue-lengths (including the customer at the server) at each of the d stations. A state transition occurs via either of the following possibilities: an external arrival event or a departure event. Thus $E(s) = \{(i, 1)\colon 1 \leqslant i \leqslant d\} \cup \{(i, 2)\colon 1 \leqslant i \leqslant d, s_i \geqslant 1\}$, where $(i, 1)$ corresponds to an external arrival to station i and $(j, 2)$ denotes a departure from station j. Note that the continuity of the F_i's and G_j's implies that simultaneous events can not occur, in the sense that simultaneous external arrivals and departures are impossible. As for the routing probabilities $p(s';\ s,\ e)$, observe that:

$$p(s';\ s,\ (i,\ 1)) = \begin{cases} 1 & \text{if} \quad s' = s + e_i \\ 0 & \text{if} \quad s' \neq s + e_i \end{cases}$$

$$p(s';\ s,\ (i,\ 2)) = \begin{cases} P_{ij} & \text{if} \quad s' = s + e_j - e_i,\ s_i \geqslant 1 \\ 1 - \sum_{j=1}^{d} P_{ij} & \text{if} \quad s' = s - e_i,\ s_i \geqslant 1 \end{cases}$$

All "speeds" $r_{se} = 1$ and $F(\cdot;\ s',\ (i,\ 1),\ s,\ e) = F_i(\cdot)$, whereas $F(\cdot;\ s',\ (i,\ 2),\ s,\ e) = G_i(\cdot)$.

To precisely define the dynamics of a GSMP, we use an associated GSSMC (general state space Markov chain). To simplify the remainder of our discussion of GSMP's, we henceforth assume that the distributions $F(\cdot;\ s',\ e',\ s,\ e)$ are continuous with $F(0;\ s',\ e',\ s,\ e) = 0$; this permits us to ignore the possibility of simultaneous arrivals and departures. We also require that $r_{se} > 0$ for some $e \in E(s)$; otherwise, the system gets "stuck" in that state s. The key to the dynamics of a GSMP is to identify a clock with each of the events $e \in E(s)$ that are "active" in state $s \in S$. In a queueing context, the clock readings will typically correspond to the amounts of time remaining until the next arrival and departure events occur, for each of the arrival and service time processes active in s. A GSSMC is obtained by applying the method of supplementary variables to the GSMP; the idea is that the GSSMC describes not only the physical state of the system, but also the states of the clocks for each of the currently active events. Thus, the GSSMC $X = (X_n\colon\ n \geqslant 0)$ will take the form $X_n = (S_n,\ C_n)$, where S_n is the physical state at the n'th transition of the GSMP, and C_n is the associated vector of clock readings.

To describe the state space of X, let $\mathbb{R}_+ = [0,\ \infty)$, $E = \bigcup_{s \in S} E(s)$, and let

$$\begin{aligned} C_s = \{ c \in \mathbb{R}_+^E\colon\ & c_e > 0 \text{ iff } e \in E(s), \\ & c_e r_{se} \neq c_{e'} r_{se'} \text{ for } e \neq e' \text{ whenever} \\ & c_e r_{se} c_{e'} r_{se'} > 0 \}. \end{aligned}$$

Then, $\Sigma = \bigcup_{s \in S}(\{s\} \times C_s)$ is the state space for the chain X. For $(s,\ c) \in \Sigma$, let

$$\begin{aligned} t^* &\equiv t^*(s,\ c) = \min\{c_e/r_{se}\colon\ e \in E(s)\}, \\ c_e^* &\equiv c_e^*(s,\ c) = c_e - t^* r_{se},\ e \in E(s), \\ e^* &\equiv e^*(s,\ c) = e \text{ iff } c_e^* = 0 \text{ and } e \in E(s). \end{aligned}$$

Note that e^* is the (unique) event triggering a transition from state, s, while t^* is the interval between transitions of the GSMP, beginning in state s with clock vector c. At a transition from s to s' triggered by event e, new clock values are independently generated for each $e' \in N(s', s, e) \equiv E(s') - (E(s) - \{e\})$; the clock values are generated from the distributions $F(\cdot; s', e', s, e)$. For $e' \in O(s', s, e) \equiv E(s') \cap (E(s) - \{e\})$, the old clock reading is kept after the transition so that $c_{e'} = c_{e'}^*(s, c)$. Finally, for $e' \in (E(s) - \{e\}) - E(s')$, event e' ceases to be scheduled after the transition so that $c_{e'} = 0$; this occurs (for a departure event clock) whenever a customer departs to leave the server idle, for example.

We are now ready to define the transition function of X. For $(s, c) \in \Sigma$, $A = \{s'\} \times \{c' \in C_{s'}: c'_e \leqslant a_e, e \in E(s')\}$, set

$$P((s, c), A) = p(s'; s, e) \prod_{e' \in N(s', s, e^*)} F(a_{e'}; s', e', s, e^*) \times \prod_{e \in O(s', s, e^*)} I_{[0, a_e]}(c_e^*) \tag{3.2}$$

Let $X = (X_n = (S_n, C_n): n \geqslant 0)$ be the co-ordinate process having transition function (3.2). Then, for $n \geqslant 1$,

$$A(n) = \sum_{k=0}^{n-1} t^*(S_k, C_k) \tag{3.3}$$

is the time of the n'th transition of the GSMP. Thus, the GSMP $Q = (Q(t): t \geqslant 0)$ may be formally defined as

$$Q(t) = \sum_{n=0}^{\infty} S_n I(A(n) \leqslant t < A(n+1)\}, \tag{3.4}$$

where $A(0) = 0$. If we assume that $A(n) \to \infty$ a.s. (i.e. that the process is non-explosive), it follows that (3.2) through (3.4) yield a well-defined process Q. The process Q is said to be the canonical GSMP associated with "problem data" as specified by (3.1). (For more detail, see Whitt [28]).

A key point of the above analysis is that a large class of queueing network simulations have Markov structure. The GSMP framework is a context which allows us to precisely identify the nature of the associated Markov structure. This, in turn, has profound implications both for the development of improved simulation algorithms and for the analysis of existing methodologies.

4. Output analysis for transient simulations

Let $X = (X_n: n \geqslant 0)$ be the GSSMC corresponding to a GSMP. Let T be a finite-valued (possibly randomized) stopping time for X and let $Y = f(X_n: 0 \leqslant n \leqslant T)$ be an $\mathbb{R}^d$-valued r.v. The *transient simulation output analysis problem* concerns the estimation of quantities of the form

$$\alpha = g(EY) \tag{4.1}$$

where $g: \mathbb{R}^d \to \mathbb{R}$.

EXAMPLE 1

Let h be a real-valued function defined on the state space S of the GSMP Q and consider the problem of estimating the parameter $\alpha = Eh(Q(t))$. This is a special case of (4.1), in which $g(y) = y$, $Y = f(S_n, C_n\colon 0 \leqslant n \leqslant N(t)+1) = h(S_{N(t)})$, and $T - 1 = N(t) \equiv \max\{n \geqslant 0\colon \Lambda(n) \leqslant t\}$.

EXAMPLE 2

Given a subset A of the state space of Q, define $S(A) = \inf\{t \geqslant 0\colon Q(t) \in A\}$. If we set $Y = \Lambda(T(A))$, where $T(A) = \inf\{n \geqslant 0\colon S_n \in A\}$, and $g(y) = y$, then (4.1) incorporates the problem of calculating $\alpha = ES(A)$.

EXAMPLE 3

For two subsets A and B, consider the problem of estimating $\alpha = E\{S(A) \mid S(A) \leqslant S(B)\}$. The estimation of such a *conditional expectation* is a special case of (4.1), in which $d = 2$, $Y = (\Lambda(T(A))I(T(A) \leqslant T(B)),\ I(T(A) \leqslant T(B)))$, and $g(y_1, y_2) = y_1/y_2$. (To define g on all of $\mathbb{R}^d$, set $g(y, 0) = 0$.)

EXAMPLE 4

Given a real-valued function h, suppose that we wish to estimate the *variance* $\alpha = \operatorname{var} h(Q(t))$. Here, $Y = (h^2(S_{N(t)}),\ h(S_{N(t)}))$, $d = 2$, and $g(y_1, y_2) = y_1 - y_2^2$.

EXAMPLE 5

Let T be a stopping time for X, $Y = (f_1(X_n\colon 0 \leqslant n \leqslant T),\ f_2(X_n\colon 0 \leqslant n \leqslant T))$ (f_1, f_2 are real-valued), and $d = 2$. The *ratio estimation* problem $\alpha = EY(1)/EY(2)(Y = (Y(1), Y(2))$ is the instance of (4.1) in which $g(y_1, y_2) = y_1/y_2$. (Define $g(y, 0) = 0$.)

A methodology for solving (4.1) is straight-forward and readily obtained. Let $X_1, X_2, \ldots$ be i.i.d. copies of the process X, and set $\overline{Y}(n) = n^{-1}\Sigma_{i=1}^{n} Y_i$ with $\overline{Y}(0) = 0$ where $Y_i = f(X_{in}\colon 0 \leqslant n \leqslant T_i)$. An estimator $\alpha(n)$ for α is then given by

$$\alpha(n) = g(\overline{Y}(n)).$$

The following theorem summarizes the behavior of $\alpha(n)$ (see Serfling [26], p. 118, for a proof which can be easily modified to cover the current circumstances.)

THEOREM 1

(i) If $E\|Y\| < \infty$ and if g is continuous in a neighbourhood of EY, then $\alpha(n) \to \alpha$ a.s. as $n \to \infty$.

(ii) If $E\|Y\|^2 < \infty$ and if g is differentiable at EY, then $n^{1/2}(\alpha(n) - \alpha) \Rightarrow \sigma N(0, 1)$ as $n \to \infty$, where $\sigma^2 = \nabla g(EY)^t C \nabla g(EY)$, $C = EYY^t - EY \cdot EY^t$, and $\nabla g(EY)$ is the gradient of g evaluated at EY.

Thus, under suitable regularity hypotheses, $\alpha(n)$ is strongly consistent for α (i.e. converges to α a.s.) and satisfies a central limit theorem (CLT). The CLT asserts that the rate of convergence of $\alpha(n)$ to α is roughly of order $n^{-1/2}$. Thus, in order to add an additional significant figure of accuracy (i.e. increase accuracy by a factor of 10), one has to increase the number of observations by a factor of 100. This suggests that simulation is a relatively inefficient tool for obtaining high-accuracy solutions. Of course, in many applications, two or three significant figures of accuracy is quite acceptable and, as indicated in section 2, simulation can then be a competitive technique.

Because of the slow convergence rate of Monte Carlo simulation, error analysis is particularly important. Typically simulators assess error by producing confidence intervals for the parameter to be estimated. The CLT provided by theorem 1 is the key to obtaining such confidence intervals in the transient setting. Set

$$C(n) = \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^t - \overline{Y}(n)\overline{Y}^t(n)$$

$$s(n) = \left(\nabla g(\overline{Y}(n))^t C(n) \nabla g(\overline{Y}(n))\right)^{1/2},$$

and let $I(n) = [\alpha(n) - z(\delta)s(n)/n^{1/2},\ \alpha_n + z(\delta)s(n)/n^{1/2}]$ where $z(\delta)$ solves $P\{N(0, 1) \leqslant z(\delta)\} = 1 - \delta/2$. The following result states that $I(n)$ is an asymptotic $100(1-\delta)\%$ confidence interval for α. (The proof is easy.)

PROPOSITION 1

If $E\|Y\|^2 < \infty$, g is continuously differentiable in a neighborhood of EY, and $\sigma^2 = \nabla g(EY)^t C \nabla g(EY) > 0$, then $P\{\alpha \in I(n)\} \to 1 - \delta$ as $n \to \infty$.

Proposition 1 describes the basic output analysis algorithm for transient simulations. It indicates what data needs to be collected from the simulation, and how the data must be manipulated in order to produce confidence intervals that are asymptotically valid.

Several important refinements of proposition 1 are worth mentioning, however. First, it may be inconvenient to set the run length in terms of the number of observations generated, since the amount of computer time required to generate a fixed number of observations will in general be random (e.g. time to simulate a Jackson network to time t corresponds to a random number of state transitions). It is often more natural to specify run length in terms of a computer budget t, and to produce confidence intervals based on the data generated within that budget constraint.

Let $\beta_1, \beta_2, \ldots$ be the amounts of computer time required to generate $Y_1, Y_2, \ldots$ respectively; we make the natural assumption that the random vectors $(Y_1, \beta_1), (Y_2, \beta_2), \ldots$ are i.i.d. Note that the number of observations $\kappa(t) = \max\{n \geqslant 0: \sum_{k=1}^{n} \beta_k \leqslant t\}$ generated by time t is a renewal process. The natural point estimate for α is $\alpha_t = \alpha(\kappa(t))$ and the natural confidence interval to use is $I_t = I(\kappa(t))$.

THEOREM 2

(i) If $E\|Y\| < \infty$, $\beta_1 < \infty$ a.s., and if g is continuous in a neighborhood of EY, then $\alpha_t \to \alpha$ a.s.

(ii) If $E\|Y\|^2 < \infty$, $E\beta_1 < \infty$, and if g is continuously differentiable in a neighborhood of EY, then $t^{1/2}(\alpha_t - \alpha) \Rightarrow \sigma N(0, 1)$ as $t \to \infty$ where $\sigma^2 = E\beta_1 \cdot \nabla g(EY)^t C \nabla g(EY)$, and $C = EYY^t - EY \cdot EY^t$. Furthermore, if $\sigma^2 > 0$, then $P\{\alpha \in I_t\} \to 1 - \delta$ as $t \to \infty$.

We refer to Glynn and Whitt [11] for a proof of theorem 2.

A second refinement concerns running the simulation according to a *sequential stopping rule*, in which the run length is not set in advance. Instead, the simulation is allowed to run until either an absolute or relative error precision ϵ is met. Note that the confidence interval $I(n)$ has half-with $z(\delta)s(n)/n^{1/2}$. Thus,

$$T(\epsilon) = \min\{n \geqslant 1:\ z(\delta)s(n)n^{-1/2} + n^{-1} < \epsilon\} \tag{4.2}$$

is a sequential stopping rule designed to stop the simulation when the interval half-width drops below ϵ (the additional n^{-1} term is included to prevent the simulation from stopping too early); this is an absolute precision stopping rule. (In a relative precision rule, ϵ would be scaled to the magnitude of α). Set $\tilde{I}(\epsilon) = I(T(\epsilon))$.

THEOREM 3

If $E\|Y\|^2 < \infty$, g is continuously differentiable in a neighborhood of EY, and $\sigma^2 = \nabla g(EY)^t C \nabla g(EY) > 0$, then $P\{\alpha \in \tilde{I}(\epsilon)\} \to 1 - \delta$ as $\epsilon \downarrow 0$.
See Glynn and Whitt [12] for a proof.

A major concern in all of the confidence interval methodologies described above is that they are based on asymptotic limit theory. For small to moderate run lengths, the limit theory may be misleading. From a practical standpoint, this manifests itself by producing confidence intervals with incorrect coverage levels (e.g. a confidence interval with a stated coverage level of 90% may only cover α 80% of the time). Several factors typically lead to confidence interval degradation. Firstly, the r.v.'s Y_1, $Y_2, \ldots$ may be highly skewed and asymmetric so that the multivariate CLT for $\overline{Y}(n)$ exhibits a slow rate of convergence. Secondly, non-linearities in g may exacerbate the situation. For example, $\overline{Y}(n)$ is unbiased for EY, but $g(\overline{Y}(n))$ is generally biased for $g(EY)$ when g is non-linear. These factors lead to slow convergence in the CLT for $\alpha(n)$ which, in turn, leads to poor coverage in $I(n)$. Development of small-sample corrections to the confidence intervals described above is an active area of current research.

5. Output analysis for steady-state simulations

In this section, we describe the output analysis problem for estimation of steady-state parameters. For f: $\Sigma \to \mathbb{R}^d$, let $Y_n = f(X_n)$ where $X = (X_n: n \geqslant 0)$ is the GSSMC associated with the queueing system. Suppose the sequence $(Y_n: n \geqslant 0)$ is ergodic in the sense that there exists a finite-valued (deterministic) vector μ such that

$$\overline{Y}(n) \Rightarrow \mu \tag{5.1}$$

as $n \to \infty$. The *steady-state simulation output analysis problem* concerns estimating $\alpha = g(\mu)$, where g: $\mathbb{R}^d \to \mathbb{R}$ is given.

In many applications, the simulator's interest centers on steady-state characteristics of the process Q, as opposed to X. Specifically, for h: $S \to \mathbb{R}^d$, interest frequently is concentrated on the long-run behavior of

$$\frac{1}{t}\int_0^t h(Q(s))\,\mathrm{d}s$$

and, more generally,

$$\alpha = k\left(\lim_{t\to\infty} \frac{1}{t}\int_0^t h(Q(s))\,\mathrm{d}s\right) \tag{5.2}$$

(if the limit exists) for k: $\mathbb{R}^d \to \mathbb{R}$.

EXAMPLE 6

If Q is the queue-length process for an open queueing network, $h(q) = q$, and $k(x) = x_1 + \ldots + x_d$, then α (as defined by (5.2)) corresponds to the steady-state expected total number of customers in the system.

EXAMPLE 7

To calculate the steady-state variance of the total number of customers in the system, let $h(q) = ((\Sigma_1^d q_i)^2, \Sigma_1^d q_i)$ and set $k(y_1, y_2) = y_1 - y_2^2$.

EXAMPLE 8

To measure the steady-state ratio of customers at two stations i and j in a queueing network, let Q be the corresponding vector queue-length process, $h(q) = (q_i, q_j)$ and set $k(y_1, y_2) = y_1/y_2$.

The next result shows that the estimation problem (5.2) is but a special case of the steady-state simulation problem for the chain X. (For the proof, see the appendix.)

PROPOSITION 2

For h: $S \to \mathbb{R}^d$, suppose that

$$\frac{1}{n}\sum_{k=0}^{n-1} h(S_k)t^*(S_k, C_k) \xrightarrow{\text{a.s.}} \mu_1$$

$$\frac{1}{n}\sum_{k=0}^{n-1} t^*(S_k, C_k) \xrightarrow{\text{a.s.}} \mu_2$$

$$\frac{1}{n}\sum_{k=0}^{n-1} \| h(S_k) \| t^*(S_k, C_k) \xrightarrow{\text{a.s.}} \mu_3$$

where μ_1, μ_2, μ_3 are finite (deterministic) constants and $\mu_2 > 0$. Then,

$$\frac{1}{t}\int_0^t h(Q(s))\, \mathrm{d}s \xrightarrow{\text{a.s.}} \mu_1/\mu_2$$

as $t \to \infty$ and (5.2) is a special case of the steady-state simulation problem, in which $f(s, c) = (h(s)t^*(s, c), t^*(s, c))$ and g: $\mathbb{R}^{d+1} \to \mathbb{R}$ is defined by $g(x_1, \ldots, x_d, x_{d+1}) = k(x_1/x_{d+1}, \ldots, x_d/x_{d+1})$.

Before proceeding to the discussion of simulation methodology for the steady-state simulation problem, we will detour (for a moment) to discussing conditions guaranteeing the validity of (5.1). Such conditions show that the steady-state simulation problem is well-posed. A key to the analysis is the Markov chain X. An extensive body of theory has been developed to study the recurrence structure of GSSMC's. One particularly powerful approach involves proving that the chain X is recurrent in the sense of Harris. Specifically, the chain X is said to be *Harris recurrent* if there exists a set $A \subseteq \Sigma$, a positive number λ, an integer $n \geqslant 1$, and a probability distribution ϕ on Σ such that

(i) $P\{X_n \in A \text{ infinitely often} \mid X_0 = (s, c)\} = 1$ for all $(s, c) \in \Sigma$

(ii) $P\{X_n \in \cdot \mid X_0 = (s, c)\} \geqslant \lambda\phi(\cdot)$ for all $(s, c) \in A$. (5.3)

In most applications, A is typically a compact subset of Σ and condition (ii) translates into a requirement that the measures $P\{X_n \in \cdot \mid X_0 = (s, c)\}$ have a common density component which is uniformly bounded (in $(s, c) \in A$) away from zero; if the density component is continuous in (s, c), the compactness of A often suffices to obtain a uniform lower bound on the densities. This discussion suggests that we can reasonably expect many GSMP's to be Harris recurrent, particularly when considering those GSMP's for which the distributions $F(\cdot;\, s', e', s, e)$ have continuous positive (Lebesgue) density components.

An important observation in the study of Harris chains is that condition (5.3) permits one to "split" the transition function $P\{X_n \in \cdot \mid X_0 = (s, c)\}$ over the set A:

$$P\{X_n \in \cdot \mid X_0 = (s, c)\} = \lambda\phi(\cdot) + (1 - \lambda)Q((s, c), \cdot) \qquad (5.4)$$

for $(s, c) \in A$, where $Q((s, c), \cdot)$ is a probability distribution on Σ. Condition (5.4) states that if $X_{\tau'} \in A$, the distribution of $X_{\tau'+n}$ is determined by a "coin flip" in which the probability of success is λ. If the coin flip is successful, $X_{\tau'+n}$ is distributed according to $\phi(\cdot)$ (independently of the position of X at time τ'), whereas, if the coin flip is unsuccessful, $X_{\tau'+n}$ is distributed according to $Q(X_{\tau'}, \cdot)$.

Note that condition (5.3) (i) guarantees that A is hit infinitely often by X. Since the probability of a successful coin flip is λ at each visit to A, a "geometric trials" argument shows that eventually a successful coin flip must occur; let τ be the associated time at which X is distributed according to ϕ. It is evident that X_τ has a distribution independent of $X_0, \ldots, X_{\tau-n}$. This regenerative-type characteristic of the random time τ can be shown to imply that Harris chains possess a non-trivial σ-finite measure π (unique up to a multiplicative constant) which is stationary in the sense that

$$\pi(\cdot) = \int_\Sigma P\{X_1 \in \cdot \mid X_0 = x\}\, \pi(\mathrm{d}x).$$

Furthermore, if $\pi(\cdot)$ is finite, it can be expressed as

$$\pi(\cdot) = E_\phi\left\{\sum_{k=0}^{\tau-1} I(X_k \epsilon \cdot)\right\} / E_\phi \tau$$

where $E_\phi(\cdot)$ denotes the expectation conditional on the initial distribution of X equalling ϕ. (See Athreya and Ney [3] for details.) Finally, if $E_\phi \tau < \infty$ and if $E_\phi(\Sigma_{k=0}^{\tau-1} \| f(X_k) \|) < \infty$, the strong law (5.1) holds:

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \to \mu \text{ a.s.}$$

as $n \to \infty$ (regardless of the initial distribution of X); see Revuz [24], p. 139. Thus, the Markov chain theory described here suggests that the limit theorem (5.1) will frequently be in force when dealing with steady-state queueing simulations. Developing precise conditions on the basic building blocks of the GSMP (i.e. S, $E(s)$, r_{se}, $p(s'; s, e)$, $F(\cdot; s', e', s, e)$) that guarantee Harris recurrence of X is an active area of current research.

An additional benefit of this approach is that if $E_\phi \tau < \infty$ and if $E_\phi \Sigma_{k=0}^{\tau-1} \| f(X_k) \| < \infty$, the regenerative-type structure of X permits us to re-express the steady-state limit μ in terms of the ratio

$$\mu = E_\phi \sum_{k=0}^{\tau-1} f(X_k) / E_\phi \tau. \tag{5.5}$$

(see Glynn [9]). As we shall see, (5.5) will permit us to reduce the steady-state simulation problem to a special case of the transient simulation problem studied in section 4. This idea lies at the core of the *regenerative method* of steady-state simulation output analysis.

Specifically, note that if $E_{\phi}\Sigma_{k=0}^{\tau-1} \parallel h(S_k) \parallel t^*(S_k, C_k) < \infty$ and $E_{\phi}\Sigma_{k=0}^{\tau-1} t^*(S_k, C_k) < \infty$, then proposition 2, (5.1) and (5.5) may be combined to prove that

$$\frac{1}{t}\int_0^t h(Q(s))\,\mathrm{d}s \xrightarrow{\text{a.s.}} \frac{E_{\phi}\sum\limits_{k=0}^{\tau-1} h(S_k)t^*(S_k, C_k)}{E_{\phi}\sum\limits_{k=0}^{\tau-1} t^*(S_k, C_k)}$$

as $t \to \infty$. Then, estimating α as defined by (5.2) is equivalent to estimating the parameter $\alpha = g(E_{\phi}Y)$ where

$$Y = \left(\sum_{k=0}^{\tau-1} h(S_k)t^*(S_k, C_k),\ \sum_{k=0}^{\tau-1} t^*(S_k, C_k)\right) \tag{5.6}$$

and $g(y_1, \ldots, y_d, y_{d+1}) = k(y_1/y_{d+1}, \ldots, y_d/y_{d+1})$. Observing that Y depends on X only up to the randomized stopping time τ, we see that α is expressed precisely in the form (4.1), which is the transient simulation problem. Thus, all the transient simulation methodology described in section 4 can be readily applied to the estimation of the steady-state parameter α.

The regenerative method is particularly straightforward to carry out in the case that the process $Q = (Q(t)\colon t \geqslant 0)$ is an S-valued continuous-time Markov chain (CTMC). This typically occurs when all the distributions $F(x;\, s', e', s, e)$ are exponential (i.e. of the form $1 - \exp(-\lambda(s', e', s, e)x)$ for $x \geqslant 0$). If Q is an irreducible positive recurrent CTMC, it is well known that the process Q "regenerates" at those instants at which Q enters a fixed state, say $y \in S$.

Thus, the regenerative structure is available to the simulator without having to explicitly "split" the transition function of X as in (5.4). In particular, if $\tau(y) = \inf\{n \geqslant 1\colon S_n = y\}$ and $E_y\Sigma_{k=0}^{\tau(y)-1}(\parallel h(S_k) \parallel + 1)t^*(S_k, C_k) < \infty$ ($E_y(\cdot)$ is the expectation operator for the CTMC conditional on $Q(0) = y$), then α (as defined by (5.2)) takes the form $\alpha = g(E_yY(y))$ where $g(y_1, \ldots, y_d, y_{d+1}) = k(y_1/y_{d+1}, \ldots, y_d/y_{d+1})$ and $Y(y)$ is given as in (5.6), with $\tau(y)$ playing the role of τ. In applying the transient methodology of section 4 to the regenerative steady-state analysis of CTMC's, we recall that the algorithm involves simulating i.i.d. copies of the chain Q from time 0 to time $\Lambda(\tau(y))$. An interesting feature of the algorithm is that because of the fact that the evolution of Q is independent over "cycles" formed by entrance times to y, the above approach is equivalent to simulating one copy of Q until time $\Lambda(\tau_n(y))$, where $\tau_n(y)$ in the time of the n'th visit of $(S_n\colon n \geqslant 0)$ to y. Thus, forming the transient estimate $\alpha(n) = g(\overline{Y}(n))$ is equivalent to simulating Q for $\Lambda(\tau_n(y))$ time units and computing $k((\overline{h \circ Q})(\Lambda(\tau_n(y))))$, where $(\overline{h \circ Q})(t) = t^{-1}\int_0^t h(Q(s))\,\mathrm{d}s$.

Given that any state $y \in S$ may be chosen as the "regeneration" state, it is natural to ask whether the efficiency of the regenerative method for estimating $\alpha = k(\lim_{t\to\infty}(\overline{h \circ Q})(t))$ is affected by the choice of the regeneration state. By efficiency, we refer to the quality of the estimator available after t units of computer time have been expended. Theorem 2 (together with the regenerative method, which permits us to estimate α via $g(\overline{Y}_n(y)))$ shows that the regeneration state x is more efficient than y if $\sigma^2(x) \leqslant \sigma^2(y)$, where $\sigma^2(\cdot) = E\beta_1(\cdot)\nabla g(EY(\cdot))^t C(\cdot)\nabla g(EY(\cdot))$ (clearly, β_1, $\nabla g(EY)$, and C all depend on the regeneration state). It seems reasonable to assume that the computational effort β_j required to generate Q over the j'th cycle takes the form

$$\beta_j = \sum_k \chi(S_k)$$

for some function $\chi(\cdot)$, where the sum is over the states visited on the j'th cycle.

PROPOSITION 3

Suppose Q is a positive recurrent irreducible CTMC, and that $k(\cdot)$ is differentiable. If $\chi(y) \geqslant 0$ for all y and if $E\beta_1(x) < \infty$, $E\|Y(x)\|^2 < \infty$ for some x, then $\sigma^2(y) = \sigma^2(x)$ for all $y \in S$.

For the proof, see the appendix. Thus, the efficiency of the regenerative method for CTMC's is independent of the choice of the regeneration state. (This result does not extend in general to the regenerative method as based on the "splitting" approach of (5.4).)

A limitation of the regenerative technique described above for queueing systems (based on the "splitting" method) is that it takes significant effort (both analytically and in programming time) on the part of the simulator to adapt software so as to identify the regeneration times. An alternative approach to developing methodologies for the steady-state simulation problem requires strengthening (5.1) to a functional central limit theorem (FCLT) hypothesis on $(Y_n\colon n \geqslant 0)$, where $Y_n = f(X_n)$: There exists a matrix A such that

$$n^{1/2}\left(\frac{1}{n}\sum_{k=0}^{\lfloor nt \rfloor} Y_k - t\mu\right) \Rightarrow AB(t) \tag{5.7}$$

as $n \to \infty$ in $D[0, \infty)$ (the Skorohod space of right continuous functions with left limits having domain $[0, \infty)$ and range $\mathbb{R}^d$), where $B(\cdot)$ is a standard Brownian motion on $\mathbb{R}^d$ (so that $EB(s)B(s)^t = s \cdot I$). Maigret [19] proved FCLT's of the form (5.7) for Harris chains, by exploiting the martingale central limit theorem. (Although her proof was stated for $d = 1$, the result is easily extended to our current setting.) Thus, (5.7) is a hypothesis which one can reasonably expect to hold for a great many queueing systems.

Let $\alpha(n) = g(\overline{Y}(n))$ be the point estimator for the steady-state limit α. The following result is easily established, and summarizes the behavior of $\alpha(n)$.

PROPOSITION 4

Suppose (5.7) holds and let $C = A^t A$. If g is continuous at μ, then $\alpha(n) \Rightarrow \alpha = g(\mu)$ as $n \to \infty$. If, in addition, g is differentiable at μ, then $n^{1/2}(\alpha(n) - \alpha) \Rightarrow \sigma N(0, 1)$ as $n \to \infty$, where $\sigma^2 = \nabla g(\mu)^t C \nabla g(\mu)$.

Proposition 4 shows that if one can construct an estimator $s(n) \Rightarrow |\sigma|$ as $n \to \infty$, then $P\{\alpha \in [\alpha(n) - z(\delta)s(n)/n^{1/2}, \alpha(n) + z(\delta)s(n)/n^{1/2}]\} \to 1 - \delta$ as $n \to \infty$ whenever $\sigma^2 > 0$. Thus, a confidence interval methodology is easily obtained, once one is given a consistent estimator for σ^2. As a consequence, the development of consistent estimators for the parameter σ^2 appearing in proposition 4 is the central problem of steady-state simulation output analysis.

To construct a consistent estimator for σ^2, observe that if g is continuously differentiable at μ, then (5.7) implies that $\nabla g(\overline{Y}(n)) \Rightarrow \nabla g(\mu)$ as $n \to \infty$. Thus, the hard part of the estimation problem deals with estimating C. To estimate C, observe that if $\{n \| \overline{Y}(n) - \mu \|^2 : n \geqslant 1\}$ is uniformly integrable, then

$$C = \lim_{n \to \infty} n \cdot E(\overline{Y}(n) - \mu)(\overline{Y}(n) - \mu)^t. \tag{5.8}$$

Suppose $(Y_n: n \geqslant 0)$ is strictly stationary with $\Sigma_{n=0}^{\infty} \| E(Y_0 - \mu)(Y_n - \mu)^t \| < \infty$. By computing the right-hand side of (5.8), one easily shows that

$$C = E(Y_0 - \mu)(Y_0 - \mu)^t + \sum_{k=1}^{\infty} E(Y_0 - \mu)(Y_k - \mu)^t + E(Y_k - \mu)(Y_0 - \mu)^t. \tag{5.9}$$

Several different methods (e.g. spectral techniques, autoregressive methods) for estimating C rely on the representation (5.9); see chapter 3 of Bratley, Fox and Schrage [6]. Recall that (5.9) only holds exactly when the sequence $(Y_n: n \geqslant 0)$ is strictly stationary. Typically, a simulation of the GSSMC X gives rise to a non-stationary process, since X_0 will usually have a non-stationary distribution. In order to justify the use of stationary process ideas in a queueing simulation context, it is of some comfort to recognize that if X is an aperiodic Harris chain with a finite invariant measure, then

$$(X_{n+k}: k \geqslant 0) \Rightarrow (X_k^*: k \geqslant 0)$$

(weak convergence in the product topology) as $n \to \infty$, where $X^* = (X_k^*: k \geqslant 0)$ is strictly stationary (see Revuz [24], p. 198). Thus, one expects that many queueing systems are at least asymptotically stationary. This suggests that, at least in a large-sample context, algorithms based on a stationarity assumption should behave reasonably well.

As in the transient simulation setting, it is often more convenient to define sample size in terms of a budget constraint t on the computer time. As in the transient case, let $\beta_1, \beta_2, \ldots$ be the amounts of computer time required to generate $Y_1, Y_2, \ldots$ respectively. Then, $\kappa(t) = \max\{n \geqslant 0: \Sigma_{k=1}^{n} \beta_k \leqslant t\}$ is the number of observations generated in t time units. The natural point estimate for α is $\alpha_t = \alpha(\kappa(t))$.

THEOREM 4

Suppose that (5.7) holds and that g is differentiable at μ. If $n^{-1}\Sigma_{k=1}^{n} \beta_k \to \gamma$ a.s. $n \to \infty$ (where γ is a finite positive deterministic constant), then $t^{1/2}(\alpha_t - \alpha)$

$\Rightarrow \sigma N(0, 1)$ as $t \to \infty$, where $\sigma^2 = \gamma \nabla g(\mu)' C \nabla g(\mu)$. If, in addition, $\sigma^2 > 0$ and $s_t \Rightarrow \sigma$ as $t \to \infty$, then $P\{\alpha \in [\alpha_t - z(\delta)s_t/t^{1/2}, \alpha_t + z(\delta)s_t/t^{1/2}]\} \to 1 - \delta$ *as* $t \to \infty$.

For a proof, see Glynn and Whitt [11].

An analogue of the sequential stopping rule methodology for transient simulations also extends to the steady-state setting. Let $s(n)$ be an estimator for $(\nabla g(\mu)' C \nabla g(\mu))^{1/2}$ based on the observations $Y_1, \ldots, Y_n$ and let $T(\epsilon)$ be defined as in (4.2).

THEOREM 5

Suppose that (5.7) holds and that g is differentiable at μ. If $s(n) \to (\nabla g(\mu)' C \nabla g(\mu))^{1/2} > 0$ a.s. as $n \to \infty$, then $P\{\alpha \in [\alpha(T(\epsilon)) - \epsilon, \alpha(T(\epsilon)) + \epsilon]\} \to 1 - \delta$ as $\epsilon \downarrow 0$.

For a proof, see Glynn and Whitt [12].

6. Run-length determination for queueing simulations

As indicated in sections 4 and 5, a major concern of simulators centers on the issue of the amount of computer time t required to calculate a queueing parameter α to within a given precision ϵ. One approach to the problem involves using a sequential stopping rule of the form (4.2). Several difficulties with the method are apparent. Firstly, prior to initiating such a simulation, the simulator will typically have no idea as to how much computer time the simulation will use in computing the $T(\epsilon)$ observations determined by the stopping rule. Thus, the simulation may significantly exceed the desired computer time budget (or, alternatively, the simulator may need to terminate the simulation early) when such a sequential algorithm is used. Secondly, the nature of the random sample size $T(\epsilon)$ can introduce certain contaminating effects into the simulation output analysis procedure that generally would not be present (or, at the least, would be present in smaller quantities) in a deterministic run-length simulation. For example, such sequential methods are prone, particularly for largish ϵ, to early termination because of an overly optimistic variance estimate caused by the observations $Y_1, Y_2, \ldots, Y_{T(\epsilon)}$ clustering closely together (due to random circumstance). However, it turns out (fortunately) that such contaminating effects decrease to zero as $\epsilon \downarrow 0$. As a consequence, the sequential stopping rules of sections 4 and 5 continue to form an important class of simulation algorithms, particularly for ϵ small.

Nevertheless, the above discussion indicates that approximation of the simulation run-length required to compute α to within ϵ can be a valuable tool to the simulator. The principal application is to determine whether a given (planned) simulation is feasible. For example, if the approximate run-length is prohibitively large for the precision ϵ, the simulation would either need to be abandoned, be made more efficient, or the permissible error level ϵ would need to be increased. Even if feasibility is guaranteed by the approximation, the approximation has further applications. The approximate run-length could be used to assign a (deterministic) computer budget t to the simulation, thereby avoiding the contamination effects described earlier.

As a first step to developing an approximation, note that if α_t is an estimator for α satisfying $t^{1/2}(\alpha_t - \alpha) \Rightarrow \sigma N(0, 1)$, the $100(1-\delta)\%$ confidence interval has length approximately $2|\sigma|z(\delta)/t^{1/2}$. Assuming $\alpha \neq 0$, this suggests that the run-length required to obtain a $100(1-\delta)\%$ confidence interval for α of length $\epsilon|\alpha|$ is approximately $4\sigma^2 z^2(\delta)/\epsilon^2\alpha^2$. Assuming that $\{t(\alpha_t - \alpha)^2\colon t \geqslant 0\}$ is uniformly integrable; σ^2 can be evaluated as

$$\sigma^2 = \lim_{t\to\infty} t \cdot MSE(\alpha_t).$$

where $MSE(\alpha_t) \equiv E(\alpha_t - \alpha)^2$. Thus, the key to developing appropriate run-length approximations is an estimate of $MSE(\alpha_t)$ for large t.

Asmussen [2] and Whitt [29] have studied the run-length approximation problem for queues in heavy traffic. Specifically, consider the problem of estimating the steady-state waiting time EW of a customer in the GI/G/1/∞ queue for a traffic intensity just below unity. Assuming that EW is estimated by $n^{-1}\sum_{k=0}^{n-1} W_k$ (W_0, $W_1, \ldots$ are the waiting times of successive customers) and that the computer time required to generate $W_0, \ldots, W_{n-1}$ is n, it is argued there that

$$\begin{aligned} t \cdot MSE(\alpha_t) &= tE\left(\frac{1}{t}\sum_{k=0}^{t-1} W_k - EW\right)^2 \\ &\approx c \cdot (1-\rho)^{-2} \end{aligned}$$

for some constant c. Thus, in order to obtain estimates for EW of high relative precision, it is necessary for $t \gg (1-\rho)^{-2}$. As pointed out by Asmussen and Whitt, a similar analysis can be formally carried out for more general networks of queues in which diffusion approximations hold.

An important feature of this theory is that it forcefully points out the substantial amounts of computer time required to obtain (relatively) accurate solutions to queues in heavy traffic.

7. Variance reduction techniques (VRT's)

The idea underlying variance reduction is to exploit the stochastic structure of the queueing model under consideration, in order to obtain improved computational efficiency. The stochastic structure can often be used to construct a variety of alternative estimators for the quantity α to be estimated. The remainder of this section will be devoted to a discussion of how to choose the most "efficient" estimator in the class available to the simulator.

Consider two competing estimators $\alpha_1(n)$ and $\alpha_2(n)$. Suppose that $\alpha_i(n)$ is obtained by generating a sequence $(Y_{ij}\colon 1 \leqslant j \leqslant n)$, and that the time required to generate Y_{ij} is given by β_{ij}. Thus, the time required to generate $\alpha_i(n)$ is given by $\beta_{i1} + \ldots + \beta_{in}$. Conversely, given a computer budget t, the estimator available at time t is $\alpha_{it} \equiv \alpha_i(N_i(t))$, where $N_i(t) = \max\{n \geqslant 0\colon \Sigma_{k=1}^{n} \beta_{ik} \leqslant t\}$.

Assume that the estimators $\alpha_i(n)$ $(i = 1, 2)$ satisfy FCLT's: There exists finite σ_i and $\epsilon < 1$ such that

$$n^{1/2}(t\alpha_i(\lfloor nt \rfloor) - t\alpha) \Rightarrow \sigma_i B(t) \tag{7.1}$$

as $n \to \infty$, in $D[\epsilon, \infty)$, where $B(\cdot)$ is a standard real-valued Brownian motion. If we set $t = 1$ in (7.1) and assume that $\{n(\alpha_i(n) - \alpha)^2\colon \mathrm{n} \geqslant 1\}$ is uniformly integrable, we find that

$$\operatorname{var} \alpha_i(n) = \sigma_i^2/n + o(1/n).$$

We shall therefore say that $\alpha_2(n)$ has lower variance than the estimator $\alpha_1(n)$ if $\sigma_2^2 \leqslant \sigma_1^2$. Techniques which give rise to a new estimator $\alpha_2(n)$ having lower variance than the original estimator $\alpha_1(n)$ are called *variance reduction techniques* (VRT's).

To understand (7.1) better, suppose that $\alpha_i(n) = g_i(\overline{Y}_i(n))$, where $\overline{Y}_i(n) = n^{-1}\Sigma_{k=1}^{n} Y_{ik}$. This comprises a class of estimators extensively studied in both sections 4 and 5. If the FCLT (5.7) holds for $\overline{Y}_i(n)$, and g_i is continuously differentiable in a neighborhood of μ_i (μ_i is the centering constant appearing in (5.7)), then (7.1) is valid with $\sigma_i^2 = \nabla g_i(\mu_i)^t A_i^t A_i \nabla g_i(\mu_i)$ (A_i is the scaling matrix appearing in (5.7)).

It might, at first, seem reasonable to choose the estimator with the smallest possible variance. However, this choice neglects the fact that the simulation of the sequence $(Y_{ij}\colon j \geqslant 0)$ may be considerably cheaper for one estimator than the other. Thus, the computer time needed to generate observations should be factored into the decision as to which estimator to use. This can be considered mathematically by analyzing the estimators α_{it} $(i = 1, 2)$ rather than the $\alpha_i(n)$'s. The following result generalizes theorems 2 and 4 (see Glynn and Whitt [11]).

THEOREM 6

Suppose (7.1) holds and that $n^{-1}\Sigma_{k=1}^{n} \beta_{ik} \to \gamma_i$ a.s. as $n \to \infty$ with γ_i finite, deterministic, and positive. Then, $t^{1/2}(\alpha_{it} - \alpha) \Rightarrow \eta_i N(0, 1)$ as $t \to \infty$ where $\eta_i^2 = \gamma_i \sigma_i^2$.

Thus, the appropriate figure of merit in comparing the computational efficiency of two competing estimators $\alpha_1(n)$, $\alpha_2(n)$ is the quantity $e_i \equiv 1/\gamma_i \sigma_i^2$. The parameter e_i is called the *efficiency* of the estimator α_i. Thus, the estimator $\alpha_2(n)$ is said to be more efficient than $\alpha_1(n)$ if $e_2 \geqslant e_1$. Techniques which give rise to a new estimator $\alpha_2(n)$ having greater efficiency than the original estimator $\alpha_1(n)$ are called *efficiency improvement techniques* (EIT's).

In much of the simulation literature, the parameters γ_1, γ_2 are ignored in the analysis of computational efficiency. Thus, the class of EIT's is then identical to the class of VRT's. There are several reasons for ignoring the effect of γ_1, γ_2. First, in many VRT's, the additional computation required to form $\alpha_2(n)$ from $\alpha_1(n)$ is believed to be small, so that $\gamma_1 \approx \gamma_2$. Secondly, the exact size of the γ_i's is hard to quantify from a mathematical standpoint, since the γ_i's typically reflect the quality of the programming implementation, as well as various machine dependent considerations.

8. VRT 1: control variates

Suppose $\alpha(n)$ is a real-valued estimator for α and that $(\delta(n)\colon n \geqslant 1)$ is an $\mathbb{R}^d$-valued sequence with mean zero. A sequence $(\delta(n)\colon n \geqslant 1)$ which is known to converge to zero asymptotically is called a *control variate* sequence. The idea behind control variates is that the correlation structure between $\alpha(n)$ and $\delta(n)$ can be fruitfully used to obtain a variance reduction.

Suppose that $\alpha(n)$ and $\delta(n)$ satisfy a joint $d+1$ dimensional multivariate CLT: There exists a finite-valued matrix A such that

$$n^{1/2}(\alpha(n)-\alpha,\ \delta(n)) \Rightarrow AN(0,\ I) \tag{8.1}$$

as $n \to \infty$. Note that (8.1) implies that $\delta(n) \Rightarrow 0$ as $n \to \infty$. To exploit the correlation between $\alpha(n)$ and $\delta(n)$, set $\alpha(n,\ \lambda) = \alpha(n) - \lambda^t\delta(n)$ for $\lambda \in \mathbb{R}^d$. Observe that the continuous mapping principle, as applied to (8.1), yields

$$n^{1/2}(\alpha(n,\ \lambda)-\alpha) \Rightarrow \sigma(\lambda)N(0,\ 1) \tag{8.2}$$

as $n \to \infty$. The CLT (8.2) shows that the new estimator $\alpha(n,\ \lambda)$ is also consistent for α, in the sense that $\alpha(n,\ \lambda) \Rightarrow \alpha$ as $n \to \infty$. Since λ is at our disposal, we may choose it so as to minimize $\sigma^2(\lambda)$. To analyze $\sigma^2(\lambda)$, write $C = A^tA$ in the form

$$C = \left(\begin{array}{c|c} \sigma^2 & b^t \\ \hline b & D \end{array}\right)$$

where D is $d \times d$ and b is $d \times 1$. If the covariance matrix D is positive definite, it is easily shown that the quadratic form $\sigma^2(\lambda)$ is minimized at

$$\lambda^* = D^{-1}b,$$

and that $\sigma^2(\lambda^*) = \sigma^2 - b^tD^{-1}b$. Of course, in a typical simulation context, the value of λ^* will need to be estimated from data (although in certain applications, D may be computable prior to initiating the simulation). Assuming $(\lambda_n\colon n \geqslant 0)$ is consistent for λ^* in the sense that $\lambda_n \Rightarrow \lambda^*$ as $n \to \infty$, a converging-together argument proves that

$$n^{1/2}(\alpha(n,\ \lambda_n)-\alpha) \Rightarrow \sigma(\lambda^*)N(0,\ 1)$$

as $n \to \infty$. Thus, the "controlled" estimator $\alpha(n,\ \lambda_n)$ achieves a variance reduction as compared with the original estimator $\alpha(n)$. To determine whether $\alpha(n, \lambda_n)$ achieves an efficiency improvement, we of course need to balance the additional cost of computing $\delta(n)$ and λ_n from the simulation against the variance reduction obtained. This trade-off is determined both by the estimation problem at hand and the choice of control variables used.

A particularly convenient set of control variates is typically available for the steady-state simulation of queueing networks. For example, consider a queueing network with d stations and l customer types. Suppose that $V_{ij} = (V_{ij}(k)\colon k \geqslant 1)$ is the sequence of customer service times for customers of type j at station i. Assume that the sequence V_{ij} is made up of i.i.d. r.v.'s with common mean μ_{ij} and finite variance σ_{ij}^2. Let $N_{ij}(n)$ be the number of service completions for type j customers at station i in the first n transitions of the GSSMC $X = (X_n = (S_n, C_n)\colon n \geqslant 0)$. Choose a set $B \subseteq \{(i, j)\colon 1 \leqslant i \leqslant d, 1 \leqslant j \leqslant l\}$; this will correspond to the set of control variates used. Specifically, let

$$\delta(n) = \left[\left(\frac{1}{N_{ij}(n)} \sum_{k=1}^{N_{ij}(n)} V_{ij}(k) - \mu_{ij}\right) \colon (i, j) \in B\right]. \tag{8.4}$$

The following result is straightforward to prove, and shows that $\delta(n)$ is a control variate.

THEOREM 7

Suppose that $(V_{ij}\colon (i, j) \in B)$ is a collection of independent sequences. Then, if $N_{ij}(n)/n \Rightarrow \nu_{ij}$ as $n \to \infty$ where ν_{ij} is finite, positive, and deterministic (for all $(i, j) \in B$), it follows that

$$n^{1/2}\delta(n) \Rightarrow A_c N(0, I)$$

where $D = A_c^t A_c$ satisfies $D = \operatorname{diag}(\sigma_{ij}^2\colon (i, j) \in B)$.

The matrix D appearing in theorem 7 is precisely the D of (8.3). Hence, the control variates defined by (8.4) have known variance structure; see Bauer, Venkataraman and Wilson [5] for further details on how to exploit this fact.

Lavenberg, Moeller and Welch [17] examined control variate sequences based on (8.4) as well as several extensions which included information on the routing structure of the network. A family of controls, called work variables, was found to be particularly effective. The ratio of variance reduction $\sigma^2(\lambda^*)/\sigma^2$ ranged form 0.2 to 0.95 for the models simulated.

9. VRT 2: indirect estimation

As the discussion of the previous six sections has suggested, discrete-event simulations for queueing systems are most naturally formulated in terms of the queue-length process $Q = (Q(t)\colon t \geqslant 0)$. In many applications, the simulator may be more interested in estimating characteristics of certain waiting times in the system. An important identity which can be used here is Little's Law.

Let $Q_T = (Q_T(t)\colon t \geqslant 0)$ be the total number of customers in some subnetwork of the system. Let $((A_n, D_n)\colon n \geqslant 1)$ be the sequence in which A_n, D_n correspond to the arrival and departure times of the n'th customer to the subnetwork. Then, $W_n = D_n - A_n$ is the amount of time a customer "waits" in the subnetwork. The following result is Little's Law.

THEOREM 8

If $A_n/n \to \lambda^{-1}$ a.s. and $n^{-1}\Sigma_{k=1}^n W_k \to w$ a.s. $(0 < \lambda,\ w < \infty)$, then

$$\frac{1}{t}\int_0^t Q_T(s)\ \mathrm{d}s \to q \text{ a.s.}$$

as $t \to \infty$ and $q = \lambda w$.

Thus, the steady-state mean waiting time w can be estimated as $w = \lambda^{-1}q = g(\lambda^{-1}, q)$ where $g(x, y) = xy$. So, provided that we record both the queue-length process Q_T and the arrival times $(A_n\colon n \geqslant 1)$ we can estimate w by techniques of the type described in section 4 (without having to observe the waiting times directly).

In most queueing simulations, however, both the waiting time sequence $W = (W_n\colon n \geqslant 1)$ and queue-length process Q_T are directly observable. The question then arises as to whether to estimate the steady-state mean waiting time w directly or to base the estimation on the identity $w = \lambda^{-1}q$. As discussed in section 7, the criterion for selection is efficiency. Glynn and Whitt [13] analyze this situation by assuming a joint FCLT for the A_n's and W_n's: There exists a finite-valued matrix A_L such that

$$n^{1/2}\left(\frac{1}{n}A_{\lfloor nt\rfloor} - t\lambda^{-1},\ \frac{1}{n}\sum_{k=1}^{\lfloor nt\rfloor} W_k - tw\right) \Rightarrow A_L B(t) \tag{9.1}$$

as $n \to \infty$ in $D[0, \infty)$, where $B(\cdot)$ is a standard two-dimensional Brownian motion. To compute the direct estimator $n^{-1}\Sigma_{k=1}^n W_k$ requires observing the queueing simulation up to the time $D_n' = \max\{D_k\colon 1 \leqslant k \leqslant n\}$. Thus, the direct estimator for w can be compared with the product estimator $(n^{-1}A_n)\cdot(\int_0^{D_n'} Q_T(s)\ \mathrm{d}s/D_n')$. The following result can be found in Glynn and Whitt [13].

THEOREM 9

If (9.1) holds with $0 < \lambda,\ w < \infty$, then

$$n^{1/2}\left(\frac{1}{n}\sum_{k=1}^{n} W_k - \left(\frac{1}{n}A_n\right)\cdot\left(\frac{1}{D_n'}\int_0^{D_n'} Q_T(s)\ \mathrm{d}s\right)\right) \Rightarrow 0$$

as $n \to \infty$.

From (9.1) and theorem 9, it follows that both estimators have the same asymptotic variance parameter (namely, σ_w^2, the (2, 2) element of $A_L^t A_L$). Thus, there is no difference, in terms of (asymptotic) efficiency, between the direct estimator for w and the product estimator based on $w = \lambda^{-1}q$.

However, in many queueing simulations, there is additional stochastic structure that can be exploited. In particular, one can often calculate λ prior to running the simulation. This, in fact, is the typical situation where Q_T corresponds to the total number of customers in an open queue; λ is then just the external arrival rate to the network. Thus, the direct estimator for w can be compared with the indirect estimator $\lambda^{-1}(\int_0^{D_n'} Q_T(s)\,\mathrm{d}s/D_n')$. To compare efficiencies, suppose:

(i) $\left\{ n^{-1}\left[\left(A_n - n\lambda^{-1}\right)^2 + \left(\sum_{k=1}^{n} W_k - nw \right)^2 \right] : n \geqslant 1 \right\}$ is uniformly integrable

(ii) $E\{W_j \mid U_i = u\}$ is non-increasing in u for all i, $j \geqslant 1$, where $U_i = A_i - A_{i-1}$.

(9.2)

The second condition states that the waiting times are (roughly speaking) non-increasing in the inter-arrival times. This is a condition which we would expect to hold in many queueing systems.

THEOREM 10

If (9.1) and (9.2) hold, then

$$n^{1/2}\left(\frac{\lambda^{-1}}{D_n'} \int_0^{D_n'} Q_T(d)\,\mathrm{d}s - w \right) \Rightarrow \sigma_q N(0, 1)$$

where $\sigma_q^2 \geqslant \sigma_w^2$.

See Glynn and Whitt [13] for the proof. We conclude that the direct estimator for w is usually more efficient than the indirect estimator.

A similar analysis can be carried out for the problem of estimating the steady-state queue-length parameter q. Note that the identity $q = \lambda w$ suggests that in the presence of known λ, the estimation of q and w are equivalent problems. As a consequence, we can conclude from theorem 10 that the estimator $\lambda n^{-1}\Sigma_{k=1}^n W_k$ is usually more efficient for calculating q than the more obvious estimator $\int_0^{D_n'} Q_T(d)\,\mathrm{d}s/D_n'$. For numerical illustrations, see Law [18].

10. VRT 3: discrete-time conversion

Suppose that $Q = (Q(t)\colon t \geqslant 0)$ is an S-valued non-explosive CTMC and consider the problem of estimating $\alpha = g(\nu)$ via simulation, where

$$\nu = E\int_0^{A(\tau(y))} r(Q(s))\,\mathrm{d}s$$

and $r: S \to \mathbb{R}^d$, $g: \mathbb{R}^d \to \mathbb{R}$. From the discussion in section 5, it is evident that the problem of estimating the steady-state parameter $\alpha = k(\lim_{t\to\infty} t^{-1} \int_0^t h(Q(s))\,\mathrm{d}s)$ is a special case of the above. The key idea is to let $r(q) = (h(q), 1)$ and let $g(x_1, \ldots, x_d, x_{d+1}) = k(x_1/x_{d+1}, \ldots, x_d/x_{d+1})$.

The obvious approach to estimating $\alpha = g(\nu)$ is to recognize that it is a special case of the transient simulation problem (4.1) and to employ the methodology described there. Let $\alpha(n)$ be the estimator obtained by generating n i.i.d. copies $(Q_i: 1 \leqslant i \leqslant n)$ of the CTMC up to the stopping time $\Lambda(\tau(y))$. It turns out that a more efficient estimator can be found by simulating only the embedded discrete-time Markov chain (DTMC). This is the basis of the principle of *discrete-time conversion*.

Let $Z = (Z_n: n \geqslant 0)$ be the embedded DTMC associated with the CTMC Q i.e. Z_n is the n'th distinct state visited by Q. (If $p(s'; s, e) = 0$ for all $e \in E(s)$, $s \in S$, then $Z_n = S_n$.) Note that if $E\int_0^{\Lambda(\tau(y))} \| r(Q(s)) \| \, ds < \infty$, then

$$E\left\{\int_0^{\Lambda(\tau(y))} r(Q(s))\,\mathrm{d}s \mid Z\right\} = \sum_{n=0}^{\tau(y)-1} r(Z_n)/\lambda(Z_n) \tag{10.1}$$

where $\lambda(z)$ is the exponential holding time parameter in $z \in S$. As a consequence of (10.1), ν can be re-expressed as

$$\nu = E \sum_{k=0}^{\tau(y)-1} r(Z_k)/\lambda(Z_k). \tag{10.2}$$

Let $\alpha_d(n)$ be the transient estimator for $\alpha = g(\nu)$ which exploits the discrete-time representation (10.2). To be more precise, let $Z(1), \ldots, Z(n)$ be i.i.d. copies of the sequence $Z(Z(i) = (Z_{in}: n \geqslant 0))$ generated up to time $\tau(y)$. Then

$$\alpha_d(n) = g(\bar{Y}'(n))$$

where $Y_k' = \sum_{n=0}^{\tau_k(y)-1} r(Z_{kn})/\lambda(Z_{kn})$. The following theorem compares the variances of $\alpha(n)$ and $\alpha_d(n)$.

THEOREM 11

Suppose that $E(\int_0^{\Lambda(\tau(y))} \| r(Q(s)) \| \,\mathrm{d}s)^2 < \infty$ and g is differentiable at ν. Then

$$n^{1/2}(\alpha(n) - \alpha) \Rightarrow \sigma N(0, 1)$$

$$n^{1/2}(\alpha_d(n) - \alpha) \Rightarrow \sigma_d N(0, 1)$$

where $\sigma^2 = \operatorname{var}[\nabla g(\mu)^t \int_0^{\Lambda(\tau(y))} r(Q(s))\,\mathrm{d}s]$ and $\sigma_d^2 = \operatorname{var}[E\{\nabla g(\mu)^t \int_0^{\tau(y)} r(Q(s))\,\mathrm{d}s \mid Z\}]$.

The proof is essentially that of theorem 1. It is a well known fact that if $E\xi^2 < \infty$, then var $E\{\xi \mathscr{F}\} \leqslant$ var ξ for any sub σ-field $\mathscr{F}$ defined on the same probability space as ξ. Thus, $\sigma_d^2 \leqslant \sigma^2$, and so $\alpha_d(n)$ always delivers a guaranteed variance reduction as compared with $\alpha(n)$.

To compare efficiencies, note that simulating Z to time $\tau(y)$ takes less effort than generating Q to $\Lambda(\tau(y))$, because we can dispense with the need to generate exponential variates. Thus, $\alpha_d(n)$ is a double winner, in the sense that it improves upon $\alpha(n)$ both from a variance and computation standpoint. For more on the method of steady-state discrete-time conversion, see Hordijk, Iglehart and Schassberger [16] and Fox and Glynn [8].

11. VRT 4: extended conditional Monte Carlo

The basic principle used in discrete-time conversion is that we can reduce the variance of an estimator via conditioning. This general idea is known, in the simulation literature, as the method of *conditional Monte Carlo*. In this section, we consider an extension of conditional Monte Carlo in which the different observations over which the estimator averages are conditioned on different r.v.'s. Following Bratley, Fox and Shrage [6], p. 71, we call this method *extended conditional Monte Carlo*. In general we can not always expect to obtain a variance reduction by using this method. Nevertheless, there is a general class of queueing simulations for which the technique does work.

Given a real-valued Markov chain (MC) $X=(X_n\colon n\geqslant 0)$ and $f\colon \mathbb{R}\to\mathbb{R}$, let $Y_n=f(X_n)$. Suppose that $(Y_n\colon n\geqslant 0)$ is ergodic in the sense that there exists finite μ such that

$$\overline{Y}(n)\Rightarrow\mu$$

as $n\to\infty$; our goal here is to efficiently estimate the steady-state mean μ. An important queueing example of such a real-valued MC is the waiting time sequence $(W_n\colon n\geqslant 0)$ for the GI/G/1/∞ queue, which obeys the recursion

$$W_{n+1}=[W_n+\eta_{n+1}]^+, \tag{11.1}$$

where the sequence $(\eta_n\colon n\geqslant 1)$ is i.i.d. and independent of W_0.

Suppose that X is a stochastically increasing MC (SIMC); i.e., X can be represented in the form $X_{n+1}=h(X_n,\eta_{n+1})$, where h is non-decreasing in both arguments and $(\eta_n\colon n\geqslant 1)$ is an i.i.d. sequence independent of X_0. Note that (11.1) is a special case of a SIMC. Our next result shows that it is often sufficient to assume only monotonicity in the first co-ordinate of h (for the proof, see the appendix).

PROPOSITION 5

Suppose that X satisfies a recursion of the form $X_{n+1}=h(X_n,\eta_{n+1})$ where h is non-decreasing in the first co-ordinate and $(\eta_n\colon n\geqslant 1)$ is an i.i.d. sequence independent of X_0. If η_n has a continuous distribution, then $X\overset{\mathscr{D}}{=}X'$ ($\overset{\mathscr{D}}{=}$ denotes equality in distribution), where $X'_{n+1} = h'(X'_n, \eta'_{n+1})$, h' is non-decreasing in both co-ordinates, and $(\eta'_n\colon n \geqslant 1)$ is an i.i.d. sequence independent of X'_0 with $\eta'_n \overset{\mathscr{D}}{=} \eta_n$.

The idea behind extended conditional Monte Carlo is to replace the estimator $\overline{Y}(n)$ by

$$\overline{Y}_e(n) = \frac{1}{n} \sum_{k=1}^{n} E\{f(X_k) \mid X_{k-1}\}.$$

Note that $\overline{Y}_e(n)$ is a sample mean of the r.v.'s $f_e(X_0), \ldots, f_e(X_{n-1})$ where

$$f_e(x) = E\{f(X_1) \mid X_0 = x\}.$$

We assume that both $\overline{Y}(n)$ and $\overline{Y}_e(n)$ satisfy CLT's: There exists σ, σ_e such that

$$n^{1/2}(\overline{Y}(n) - \mu) \Rightarrow \sigma N(0, 1)$$

$$n^{1/2}(\overline{Y}_e(n) - \mu) \Rightarrow \sigma_e N(0, 1) \tag{11.2}$$

as $n \to \infty$. Furthermore, assume that

$$\left\{n\left[(\overline{Y}(n) - \mu)^2 + (\overline{Y}_e(n) - \mu)^2\right]\colon n \geqslant 1\right\} \tag{11.3}$$

is uniformly integrable. The following theorem shows that extended conditional Monte Carlo reduces variance if f is monotone and X is a SIMC.

THEOREM 12

Assume (11.2) and (11.3). If X is a SIMC and f is monotone, then $\sigma_e^2 \leqslant \sigma^2$.

For the proof, see the appendix. Because of the stochastic monotonicity that holds in many queueing networks, we expect that extended conditional Monte Carlo should also prove useful in situations where Q is vector-valued. For a result closely related to theorem 12, see Ross [25]. The efficiency of the method depends on the amount of variance reduction and the ease of computing f_e.

12. VRT 5: common random numbers

The idea underlying common random numbers (CRN's) is that to compare two queueing systems, it is "fairer" to drive the two systems using the same stream of random numbers. If CRN's are used and one system is treated "unfairly" by receiving a particularly unlucky string of random inputs, then both are. Of course, in making this statement, we are implicitly assuming that both systems respond in a similar manner to the driving inputs. This suggests that the method of CRN's depends on monotonicity.

Assume that X_1, X_2 are real-valued SIMC's ($X_i = (X_{in}\colon n \geqslant 0)$) and let f_1, f_2 be non-decreasing functions. Let $Y_{in} = f_i(X_{in})$ and suppose that $(Y_{in}\colon n \geqslant 0)$ ($i = 1, 2$) is ergodic in the sense that

$$\bar{Y}_i(n) \Rightarrow \mu_i$$

as $n \to \infty$. Assuming that μ_i represents the "performance" of system i, the simulator is often interested in estimating the difference $\alpha = \mu_1 - \mu_2$ in performance between the two queueing systems. The naive estimation approach would involve simulating X_1 and X_2 independently and estimating α via $\alpha_I(n) = \bar{Y}_1(n) - \bar{Y}_2(n)$. Assuming that there exists finite σ_1, σ_2 such that

$$n^{1/2}\big(\bar{Y}_i(n) - \mu_i\big) \Rightarrow \sigma_i N(0, 1), \tag{12.1}$$

it is easily seen that

$$n^{1/2}\big(\alpha_I(n) - \alpha\big) \Rightarrow \sigma N(0, 1)$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Suppose now that the method of CRN's is used. Specifically, suppose that $X_{1,n+1} = h_1(X_{1,n}, \eta_{n+1})$ and $X_{2,n+1} = h_2(X_{2,n}, \nu_{n+1})$ where $(\eta_n\colon n \geqslant 1)$ is i.i.d. independent of $X_{1,0}$ and $(\nu_n\colon n \geqslant 1)$ is i.i.d. independent of $X_{2,0}$. However, we now require that:

$$\{X_{1,0}, X_{2,0}, \eta_n, \nu_n\colon n \geqslant 1\} \tag{12.2}$$

is an associated family of r.v.'s. A standard approach to obtaining nontrivial association between the η_n's and ν_n's is to generate both η_n and ν_n by inversion from the same r.v. U_n:

$$\eta_n = F^{-1}(U_n)$$

$$\nu_n = G^{-1}(U_n),$$

where F^{-1}, G^{-1} are the inverse distribution functions for η, ν respectively.

However, in many settings, one can obtain the required association without explicit use of inversion. For example, suppose that $P\{\nu_n \in dx\} = F(dx/\theta_2)$ and $P\{\eta_n \in dx\} = F(dx/\theta_1)$, where θ_1, θ_2 are positive. Thus, the distributions of η_n and ν_n differ by a change of "scale". Suppose that the sequence $(\eta_n\colon n \geqslant 1)$ is generated arbitrarily (e.g. by acceptance-rejection). If the ν_n's are derived from the η_n's via

$$\nu_n = (\theta_2/\theta_1)\eta_n,$$

then (12.2) follows. Thus, it is *not* necessary, in this change-of-scale setting, to use inversion in order to apply CRN's. In a similar fashion, explicit inversion can be avoided when the distribution of ν_n differs from that of η_n by a change of "location". Changes of scale and location arise frequently when making comparison of stochastic systems.

Let $\alpha_c(n) = \overline{Y}_1(n) - \overline{Y}_2(n)$ be the CRN estimator, in which the input sequences $(\eta_n\colon\ n \geqslant 1)$ and $(\nu_n\colon\ n \geqslant 1)$ are dependent as in (12.2). Assume that

$$n^{1/2}(\alpha_c(n) - \alpha) \Rightarrow \sigma_c N(0, 1) \tag{12.3}$$

and that

$$\left\{n\left[\left(\overline{Y}_1(n) - \mu_1\right)^2 + \left(\overline{Y}_2(n) - \mu\right)^2\right]\colon\ n \geqslant 1\right\} \tag{12.4}$$

is a uniformly integrable family of r.v.'s. The following theorem is easily proved by the methods of Heidelberger and Iglehart [14].

THEOREM 13

Suppose X_1, X_2 are SIMC's with f_1, f_2 non-decreasing. If (12.1)–(12.4) are in force, then $\sigma_c^2 \leqslant \sigma_I^2$.

Because of the stochastic monotonicity present in many queueing systems, this suggests that a variance reduction will often be achieved by using CRN's to compare the performance of two queueing systems. An efficiency improvement will result if the variance is reduced and if the computational effort required to associate the η_n's and ν_n's is not too large. Note that if inversion was used to generate the η_n's and ν_n's independently, one can continue to use inversion in order to associate the η_n's and ν_n's, provided that the same stream of common uniforms is used for both families of r.v.'s. In this case, the effort required to associate the sequences is no greater than the effort required in the independent case.

An idea similar to CRN's can be used to estimate $\mu = \mu_1$ (see (12.1)). Henceforth, for the remainder of this section, we drop the subscript from the first SIMC. Note that now we are dealing with a single SIMC so that no comparison is being made. Suppose that the η_n's defining X are generated by inversion, so that $\eta_n = F^{-1}(U_n)$ when $U_1, U_2, \ldots$ are i.i.d. uniform r.v.'s. Observing that the *antithetic* r.v.'s $1 - U_1, 1 - U_2, \ldots$ are also i.i.d. uniforms, we can define the antithetic chain X_n' via $X_0' = X_0$ and $X_{n+1}' = h(X_n', F^{-1}(1 - U_{n+1}))$. Let $Y_n' = f(X_n')$ and consider the antithetic estimator

$$\alpha_a(n) = \tfrac{1}{2}(\overline{Y}(n) + \overline{Y}'(n)).$$

To compare the variance of this estimator with that of $\overline{Y}(\cdot)$, note that $\alpha_a(n)$ corresponds to simulating the dynamics of the SIMC for a total of $2n$ transitions. Thus, $\alpha_a(n)$ must be compared to $\overline{Y}(2n)$. From (12.1), the relative variance quantity is $\sigma^2/2$. Suppose that

$$n^{1/2}(\alpha_a(n) - \mu) \Rightarrow \sigma_a N(0, 1) \tag{12.5}$$

as $n \to \infty$. The following result can be proved by the methods of Heidelberger and Iglehart [14].

THEOREM 14

Suppose X is a SIMC and f is monotone. If (12.1), (12.4), (12.5) are in force, then $\sigma_a^2 \leqslant \sigma^2/2$.

Again, due to the stochastic monotonicity present in many networks, we expect that the method of antithetics will often provide a variance reduction in the queueing context. A disadvantage of antithetics, as opposed to CRN's, is that it appears inversion *must* be used to generate the corresponding input streams of r.v.'s.

13. VRT 6: simulating the error in heavy traffic

As indicated in section 6, the amount of computer time t required to simulate the steady-state mean waiting time EW must satisfy $t \gg (1-\rho)^{-2}$ as $\rho \nearrow 1$, in order to estimate EW to a fixed degree of relative precision ϵ. Thus, the problem of estimating the quantity EW gets harder, in a relative sense, as $\rho \nearrow 1$. Minh and Sorli [21] developed an elegant method for avoiding this difficulty when EW is the steady-state mean waiting time in the GI/G/1/∞ queue.

Let V_0, U_1 be the service time of the zero'th customer and the first interarrival time, respectively, in the GI/G/1/∞ queue. Then, Marshall [20] showed that if $EV_0 < EU_1$, $E(V_0^2 + U_1^2) < \infty$, then

$$EW = -\frac{EZ^2}{2EZ} - \frac{EI^2}{EI}$$

where $Z = V_0 - U_1$ and I is the length of the first idle period in a GI/G/1/∞ queue in which the zero'th customer arrives at $t = 0$. Thus, estimating $\alpha = EW$ is equivalent to estimating $\alpha = g(EI^2, EI)$, where $g(x, y) = -EZ^2/2EZ - x/y$; the methods of section 4 can then be used to estimate α. This was the basic idea of Minh and Sorli [21].

Noting that $EW \sim -EZ^2/2EZ$ as $\rho \nearrow 1$ is the classical diffusion approximation for EW, we recognize that the approach taken here is to use simulation only to estimate the error term in the diffusion approximation to EW. Extensions of this idea to the GI/G/s/∞ queue appear in Minh [22].

14. VRT 7: importance sampling

Consider the problem of estimating the parameter $\alpha = P\{W > w\}$, where W is the steady-state mean waiting time of a customer in a GI/G/1 queue in which $\rho < 1$. If w is large, the event $\{W > w\}$ occurs rarely, so that conventional simulation is inefficient. The idea underlying *importance sampling* is to alter the dynamics of the queueing system via a "change of measure" so that the event, under the altered dynamics, occurs more frequently (i.e. the "important" event occurs more often). Of course, in order to account for the new dynamics, the estimator must be suitably modified.

To illustrate, suppose that $P\{W > w\}$ is estimated by using the fact that $W \overset{\mathscr{D}}{=} \max(S_k\colon k \geqslant 0)$, where $(S_n\colon n \geqslant 0)$ is the random walk with negative drift and $S_0 = 0$ associated with the GI/G/1/∞ queue. Then, α can be re-expressed as $\alpha = P\{T < \infty\}$ where $T = \inf\{n \geqslant 0\colon S_n > w\}$. Suppose the i.i.d. summands of the random walk have common distribution F, where F has a moment generating function ϕ which converges in a neighborhood of the origin. Since $(S_n\colon n \geqslant 0)$ has negative drift, it follows that $\phi'(0) < 0$. Thus, if there exists a positive root to $\phi(\theta) = 1$, it must be unique (by the convexity of ϕ). Suppose that such a unique root θ^* exists.

Rather than generating the summands of the random walk from the distribution F, consider generating the summands from the distribution

$$\tilde{F}(\mathrm{d}x) = \mathrm{e}^{\theta^* x} F(\mathrm{d}x).$$

Then, it is straightforward to verify that

$$\alpha = P\{T < \infty\} = \tilde{E}\left\{\mathrm{e}^{-\theta^* S_T};\ T < \infty\right\},$$

where $\tilde{E}(\cdot)$ corresponds to the expectation in which the summands have distribution $\tilde{F}$. Since $T < \infty$ a.s. under $\tilde{P}$ (the random walk has positive drift under $\tilde{P}$),

$$\alpha = \tilde{E} \exp(-\theta^* S_T). \tag{14.1}$$

Siegmund [27] and Asmussen [1] show that simulation based on (14.1) is much more efficient than simulation under F. The key is that $\tilde{F}$ turns the drift positive, so that the event $\{T < \infty\}$ becomes certain. The factor $\exp(-\theta^* S_T)$ is introduced to compensate for the new measure $\tilde{F}$.

This area is currently active, from a research viewpoint. For an application of importance sampling to networks of queues, see Parekh and Walrand [23].

15. Gradient estimation for queueing systems

Consider a queueing system in which the dynamics depend on a parameter $\theta \in \mathbb{R}^m$. Specifically, let $Q = (Q(t)\colon t \geqslant 0)$ be a GSMP in which the clock-setting distributions, and routing probabilities depend on θ; i.e., under parameter θ, the clocks are driven by distributions $F(\theta, \cdot;\, s', e', s, e)$ and customers are routed via $p(\theta, s';\, s, e)$. Clearly, the queueing parameter α to be estimated by the simulation algorithm then depends on θ, i.e. $\alpha = \alpha(\theta)$. A current area of vigorous research concerns the efficient estimation of $\nabla\alpha(\theta)$. The interest in the problem derives from the fact that gradient evaluations form an important ingredient of most efficient optimization routines.

Two basic methods have been suggested for attacking this problem. The first technique, called *infinitesimal perturbation analysis* (IPA), typically assumes that the routing probabilities are independent of θ, so that all θ-dependence is incorporated via the distributions $F(\theta, \cdot; s', e', s, e)$. Also, the parameter $\alpha = \alpha(\theta)$ usually takes the form $\alpha(\theta) = g(E_\theta Y)(E_\theta(\cdot)$ denotes expectation under the θ-dynamics), where

$$E_\theta Y = E_\theta \int_0^{A(N)} h(Q(s))\,\mathrm{d}s, \tag{15.1}$$

N is an integer-valued r.v., $h: S \to \mathbb{R}^d$, $g: \mathbb{R}^d \to \mathbb{R}$. Suppose, for simplicity, that $\theta \in \mathbb{R}$. The idea is to observe that if $f(\theta, \cdot; s', e', s, e)$ is a scale family in θ (i.e. $F(\theta, \cdot; s', e', s, e) = F(\cdot/\theta; s', e', s, e)$), the clock-setting variates associated with the θ-system can be generated as $\theta R_1(s', e', s, e)$, $\theta R_2(s', e', s, e), \ldots$ where the sequence $(R_n(s', e', s, e)$: $n \geqslant 1)$ is that associated with $F(\cdot; s', e', s, e)$. Let $Q(\theta) = (Q(\theta, t)$: $t \geqslant 0)$ be the GSMP in which the clock readings are set by the sequence $(\theta R_n(s', e', s, e)$: $n \geqslant 1)$. The key observation is that for Y of the form (15.1), $E_\theta Y = EY(\theta)$ where

$$Y(\theta) = \int_0^{A(\theta, N(\theta))} h(Q(\theta, s))\,\mathrm{d}s = \sum_{k=0}^{N(\theta)-1} h(S_k(\theta))t^*(S_k(\theta), C_k(\theta)). \tag{15.2}$$

Given θ_0, the r.v.'s $(N(\theta), S_k(\theta)$: $1 \leqslant k < N(\theta))$ are typically constant in some neighborhood of θ_0. Therefore, it follows from (15.2) that in computing the path-by-path derivative $Y'(\theta)$ of $Y(\theta)$, we need only to compute the derivative of $t^*(S_k(\theta_0), C_k(\theta))$ with respect to θ. In many queueing systems, this calculation can easily be done. This allows us to calculate $\alpha'(\theta)$ as $\alpha'(\theta) = \nabla g(EY(\theta))^t EY'(\theta)$ (note that $\alpha'(\theta) = \tilde{g}(E\tilde{Y}(\theta))$ is a transient parameter of the type described in section 4, where $\tilde{g}(y_1, y_2) = \nabla g(y_1)^t y_2$ and $\tilde{Y}(\theta) = (Y(\theta), Y'(\theta)))$, provided that

$$\frac{\mathrm{d}}{\mathrm{d}\theta} EY(\theta) = EY'(\theta). \tag{15.3}$$

A similar method works for location families (i.e. $F(\theta, \cdot; s', e', s, e) = F(\cdot - \theta; s', e', s, e))$ and even more generally.

While the above method works efficiently on certain classes of queueing systems, it is not valid universally. In particular, there is a large class of non-pathological queueing networks for which (15.3) is false; see Heidelberger et al. [15] for details. For further background on IPA, see Zazanis and Suri [30].

The second technique for evaluating gradients uses the method of *likelihood ratios*. Here, the θ-dependence may be reflected in both the clock-setting distributions $F(\theta, \cdot, s', e', s, e)$ and routing probabilities $p(\theta, s'; s, e)$. However, the following density conditions need to be in force:

(i) $F(\theta, \mathrm{d}x; s', e', s, e) = f(\theta, x, s', e', s, e) \cdot F(\theta_0, \mathrm{d}x; s', e', s, e)$ where $f(\theta, \cdot; s', e', s, e)$ is positive and continuously differentiable in θ.

(ii) $p(\theta, s'; s, e)$ is either identically zero or identically positive as a function of θ, for each triplet (s', s, e). (15.4)

Given a transient parameter of the form (4.1) (note that Y need not be of the form (15.1) here), the density assumption (15.4) permits us to estimate $\alpha(\theta) = g(E_\theta Y)$ by using the likelihood ratio identity

$$E_\theta Y = E_{\theta_0} Y L(\theta), \tag{15.5}$$

where $L(\theta) = \mathrm{d}P_\theta / \mathrm{d}P_{\theta_0}$ is the Radon-Nikodym derivative of the distribution P_θ with respect to P_{θ_0} (existence of $\mathrm{d}P_\theta / \mathrm{d}P_{\theta_0}$ is assured by (15.4)). Hence, if θ is scalar, this suggests that $\alpha'(\theta_0) = \nabla g(E_{\theta_0} Y)^t E_{\theta_0} Y L'(\theta_0)$, provided that

$$\frac{\mathrm{d}}{\mathrm{d}\theta} E_{\theta_0} Y L(\theta) = E_{\theta_0} Y L'(\theta_0). \tag{15.6}$$

The derivative interchange (15.6) is typically valid for queueing systems of arbitrary structure. As in the case of IPA, it is easily seen that $\nabla g(E_{\theta_0} Y)^t E_{\theta_0} Y L'(\theta_0)$ is itself a transient parameter of the form (4.1) so that the methodology of section 4 may be used to develop an efficient estimator for $\alpha'(\theta_0)$. For more details on this approach, including a formula for $L'(\theta_0)$, see Glynn [10].

Appendix

Proof of proposition 2

Let $N(t) = \max\{n \geqslant 0: \Lambda(n) \leqslant t\}$. Since $n^{-1}\Lambda(n) \to \mu_2$ a.s., it follows that $N(t)/t \to 1/\mu_2$ a.s. Also, it is apparent that

$$\left\| \frac{1}{t} \int_0^t h(Q(s))\, \mathrm{d}s - \frac{1}{t} \sum_{k=0}^{N(t)} h(S_k) t^*(S_k, C_k) \right\| \leqslant \| h(S_{N(t)}) \| \, t^*(S_{N(t)}, C_{N(t)})/t \tag{A.1}$$

$$\left| \frac{1}{t} \sum_{k=0}^{N(t)} t^*(S_k, C_k) - 1 \right| \leqslant t^*(S_{N(t)}, C_{N(t)})/t. \tag{A.2}$$

Since $N(t) \to \infty$ a.s., evidently $N(t)^{-1} \Sigma_{k=0}^{N(t)} \| h(S_k) \| \, t^*(S_k, C_k) \to \mu_3$ a.s., which implies that $\| h(S_{N(t)}) \| \, t^*(S_{N(t)}, C_{N(t)})/N(t) \to 0$ a.s. But $N(t)/t \to 1/\mu_2$ a.s., so the right-hand side (RHS) of (A.1) converges to 0 a.s. Similarly, we see that the RHS of (A.2) goes to 0 a.s. Inequalities (A.1) and (A.2) then yields

$$\frac{1}{t}\int_0^t h(Q(s))\,\mathrm{d}s - \frac{\sum_{k=0}^{N(t)} h(S_k)t^*(S_k,\, C_k)/N(t)}{\sum_{k=0}^{N(t)} t^*(S_k,\, C_k)/N(t)} \to 0 \text{ a.s.} \tag{A.3}$$

The hypothesized strong laws then can be applied to the RHS of (A.3) to obtain the result.

Proof of proposition 3

By theorem 4, p. 84, of Chung [7], $E\,\|\,Y(y)\,\|^2 < \infty$ for all y if $E\,\|\,Y(x)\,\|^2 < \infty$. It is well known that $\mu = E_y \int_0^{\Lambda(\tau(y))} h(Q(s))\,\mathrm{d}s / E_y \Lambda(\tau(y))$ is independent of y. Write $Y(y) = (Y'(y),\, T(y))$, where $T(y) = \Lambda(\tau(y))$. Then, an easy calculation shows that $\nabla g(EY(y))^t C(y) \nabla g(EY(y)) = \operatorname{var}_y Z(y)/(E_y T(y))^2$ where $Z(y) = \nabla k(\mu)^t [Y'(y) - (E_y Y'(y)/E_y T(y)) T(y)]$. By theorem 1, p. 99, of Chung [7], $\operatorname{var}_y Z(y)/E_y T(y)$ is independent of y. Thus, the proof is complete if we show that $E_y \beta_1(y)/E_y T(y)$ is independent of y. But $E\beta_1(y)/ET(y) = -\Sigma\chi(z)\nu(z)/\Sigma A_{zz}^{-1}\nu(z)$, where ν is the stationary distribution of the embedded chain of Q and A is the generator of Q; this latter ratio is independent of y.

Proof of proposition 5

Put $G(\cdot) = P\{\eta_n \leqslant \cdot\}$ and $F(x,\, \cdot) = P\{h(x,\, \eta_n) \leqslant \cdot\}$. Let $F^{-1}(x,\, y)$ be the inverse function defined by $F^{-1}(x,\, y) = \sup\{z\colon F(x,\, z) \leqslant y\}$. Set $X_0' = \mathrm{X}_0$ and define $(X_n'\colon n \geqslant 1)$ via the recursion $X_{n+1}' = h'(X_n',\, \eta_{n+1})$, where $h'(x,\, y) = F^{-1}(x,\, G(y))$. Note that X' is a MC with transition distribution

$$\begin{aligned} P\{X_{n+1}' \leqslant y \mid X_n' = x\} &= P\{h'(x,\, \eta_{n+1}) \leqslant y\} \\ &= P\{F^{-1}(x,\, U) \leqslant y\}; \end{aligned}$$

$G(\eta) = U$ is a uniform r.v. since G is continuous. But $P\{F^{-1}(x,\, U) \leqslant y\} = F(x,\, y)$ because $F^{-1}(x,\, \cdot)$ is an inverse function. Thus, it follows that $X \overset{\mathscr{D}}{=} X'$. Monotonicity of h' in the second co-ordinate is easy. For the first co-ordinate, use the fact that $F(\cdot,\, y)$ is non-increasing to conclude that $F^{-1}(\cdot,\, y)$ is non-decreasing.

Proof of Theorem 12

Because of the uniform integrability, it suffices to show that

$$\operatorname{var} \sum_{k=1}^{n} f_e(X_k) \leqslant \operatorname{var} \sum_{k=2}^{n+1} f(X_k);$$

i.e.

$$\sum_{k=1}^{n} \operatorname{var} f_e(X_k) + 2\sum_{k\neq j} \operatorname{cov}\big(f_e(X_{k+1}),\, f(X_{j+1})\big).$$

$$\leqslant \sum_{k=1}^{n} \operatorname{var} f(X_{k+1}) + 2\sum_{h\neq j} \operatorname{cov}\big(f(X_{k+1}),\, f(X_{j+1})\big).$$

Note that $\operatorname{var} f_e(X_k) = \operatorname{var} E\{f(X_{k+1}) \mid X_k\} \leqslant \operatorname{var} f(X_{k+1})$ by the principle of conditional Monte Carlo. To complete the proof, we therefore need to show that $Ef_e(X_k)f_e(X_j) \leqslant Ef(X_{k+1})f(X_{j+1})$ for $k<j$. By the Markov property, $Ef(X_{k+1})f(X_{j+1}) = Ec_{j-k}(X_k)$, where $c_l(x) = E\{f(X_1)f(X_{l+1}) \mid X_0 = x\}$. Notice that if $X_0 = x$, X_n is a non-decreasing function of the independent r.v.'s $\eta_1, \ldots, \eta_n$. Since f is monotone, it follows that $f(X_1)$ and $f(X_{l+1})$ are associated r.v.'s, conditional on $X_0 = x$. Thus, a standard inequality (see Barlow and Proschan [4], p. 31) yields

$$\begin{aligned} c_l(x) &\geqslant E\{f(X_1) \mid X_0 = x\} \cdot E\{f(X_{l+1}) \mid X_0 = x\} \\ &= f_e(x) \cdot E\{f_e(X_l) \mid X_0 = x\}. \end{aligned}$$

If we integrate the above inequality with respect to $P\{X_k \in \mathrm{d}x\}$, we obtain $Ef(X_{k+1})f(X_{j+1}) \geqslant Ef_e(X_k)f_e(X_j)$.

References

[1] S. Asmussen, Conjugate processes and the simulation of ruin problems, Stochastic Process. Appl. 20 (1985) 213–229.

[2] S. Asmussen, Regenerative simulation in heavy traffic, Technical Report, Aalborg University Centre, Denmark, 1987.

[3] K.B. Athreya and P. Ney, A new approach to the limit theory of recurrent Markov chains, Trans. Amer. Math. Soc. 245 (1978) 493–501.

[4] R.E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing* (Holt, Rinehart and Winston, New York, 1975).

[5] K.W. Bauer, S. Venkataraman and J.R. Wilson, Estimation procedures based on control variates with known covariance matrix, *Proc. of the 1987 Winter Simulation Conference* (1987) 334–341.

[6] P. Bratley, B.L. Fox and L.E. Schrage, *A Guide to Simulation* (Springer-Verlag, New York, 1987).

[7] K.L. Chung, *Markov Chains with Stationary Transition Probabilities* (Springer-Verlag, New York, 1967).

[8] B.L. Fox and P.W. Glynn, Discrete-time conversion for simulating semi-Markov processes, Oper. Res. Lett. 5 (1986) 191–196.

[9] P.W. Glynn, Regenerative simulation of Harris recurrent Markov chains, Technical Report 18, Dept. of Operations Research, Stanford University, Stanford, CA, 1982.

[10] P.W. Glynn, Likelihood ratio gradient estimation: An overview, in: *Proc. of the 1987 Winter Simulation Conference* (1987) 366–375.

[11] P.W. Glynn and W. Whitt, Efficiency of simulation estimators, In preparation, 1988.

[12] P.W. Glynn and W. Whitt, Sequential stopping in simulation estimation. In preparation, 1988.

[13] P.W. Glynn and W. Whitt, Indirect estimation via $L = \lambda W$, Oper. Res. (1988) to appear.

[14] P. Heidelberger and D.L. Iglehart, Comparing stochastic systems using regenerative simulation with common random numbers, Adv. Appl. Prob. 11 (1979) 804–819.

[15] P. Heidelberger, X.-R. Cao, M.A. Zazanis and R. Suri, Convergence properties of infinitesimal perturbation analysis estimates, Manage. Sci. (1988) to appear.

[16] A. Hordijk, D.L. Iglehart and R. Schassberger, Discrete-time methods of simulating continuous-time Markov chains, Adv. Appl. Prob. 8 (1976) 772–788.

[17] S.S. Lavenberg, T.L. Moeller and P.D. Welch, Statistical results on multiple control variables with application to queueing network simulation, Oper. Res. 30 (1982), 182–202.

[18] A.M. Law, Efficient estimators for simulated queueing systems, Manage. Sci. 212 (1975) 30–41.

[19] N. Maigret, Théorème de limite centrale functionnel pour une chaine de Markov récurrente au sens de Harris et positive, Annales de l'Institut Henri Poincaré B 11 (1978) 187–197.

[20] K.T. Marshall, Some relationships between the distributions of waiting time, idle time and interoutput time in the GI/G/1 queue, SIAM J. Appl. Math. 16 (1968) 324–327.

[21] D.L. Minh and R.M. Sorli, Simulating the GI/G/1 queue in heavy traffic, Oper. Res. 31 (1983) 966–971.

[22] D.L. Minh, Simulating GI/G/k queues in heavy traffic, Manage. Sci. 33 (1987) 1192–1199.

[23] S. Parekh and J. Walrand, Quick simulation of excessive backlogs in networks of queues, *Proc. of the 28'th Conference on Decision and Control* (1986) 979–986.

[24] D. Revuz, *Markov Chains* (North-Holland, New York, 1984).

[25] S.M. Ross, Simulating average delay–variance reduction by conditioning, *Probability in the Engineering and Informational Sciences* (1988) to appear.

[26] R.J. Serfling, *Approximation Theorems for Mathematical Statistics* (John Wiley, New York, 1980).

[27] D. Siegmund, Importance sampling in the Monte Carlo study of sequential tests, Ann. Stat. 4 (1976) 673–684.

[28] W. Whitt, Continuity of generalized semi-Markov processes, Math. Oper. Res. 5 (1980) 494–501.

[29] W. Whitt, Planning queueing simulations, Manage. Sci. (1988) to appear.

[30] M.A. Zazanis and R. Suri, Comparison of perturbation analysis with conventional sensitivity estimates for stochastic systems, Oper. Res. (1988) to appear.