Philip Heidelberger

# Variance Reduction Techniques for the Simulation of Markov Processes, I: Multiple Estimates

*A method for reducing the variance of simulation-generated estimates is proposed and discussed. The method may be applied to the estimation of steady state parameters of discrete and continuous time Markov chains, semi-Markov processes, and regenerative discrete time Markov processes on a general state space (such as the waiting time process in a multiple-server queue). The method is similar to the technique of control variables, but differs in that the means of the controls need not be explicitly known. Numerical results for a variety of simple queueing models are presented.*

## 1. Introduction and summary

In recent years computer simulation has become a very important tool for analyzing the behavior of stochastic processes. As the structures of widely used processes become increasingly complex, analytic results become more difficult to obtain. Frequently simulation is the only computationally feasible method to study a process.

Unfortunately simulation can be a very expensive tool to use. It is therefore desirable to develop methods that can reduce the run lengths (and hence cost) of simulation without a decrease in the accuracy of estimates. Such methods are called variance reduction techniques. This paper will propose and test a new variance reduction technique for the special case when the stochastic process being simulated is a Markov process. A subsequent paper will describe several other related techniques applicable when the Markov process has a finite state space.

As an example of how expensive simulations can be, consider estimating via simulation $E[W]$, the expected stationary waiting time in an M/M/1 queue. The M/M/1 queue is not something that one would ordinarily simulate since analytic results for it are readily available. However, despite its simplicity the waiting time process for this queue can be a very expensive process to simulate. It is therefore a good candidate for testing simulation methodologies. Let $\rho$ be the usual traffic intensity of the queue and let $\overline{W}_N$ be the average of the first $N$ waiting times. It is known that if $\rho < 1$ (see Crane and Iglehart [1] or Iglehart [2]), then $\overline{W}_N$ has an asymptotically normal distribution with mean $E[W]$ and variance $\sigma^2/N$ for some constant $\sigma^2$ $(0 < \sigma^2 < \infty)$. The variance term $\sigma^2/N$ includes the effect of correlation between the waiting times. From this central limit theorem confidence intervals for $E[W]$ may be formed.

A major problem faced by simulators is how long to run a simulation. One possible stopping criterion is to run the simulation until the half length of a confidence interval is some prespecified fraction of the quantity to be estimated. Table 1 provides an indication of the run lengths needed for the M/M/1 queue when such a stopping rule is used. It is seen that as $\rho$ increases (beyond $\rho = 0.3$), the required run lengths increase rapidly until for large values of $\rho$ one must simulate such an enormous number of customers to obtain "decent" estimates that simulation is no longer a feasible alternative. It is run lengths such as these that variance reduction techniques are designed to shorten.

One of the most effective variance reduction techniques is that of control variables. A good introduction to this technique is given in Gaver and Thompson [3]. Recent studies involving control variables may be found in Carson [4], Gaver and Shedler [5], Iglehart and Lewis [6], Lavenberg [7], Lavenberg, Moeller, and Sauer [8], Lav-

enberg, Moeller, and Welch [9, 10], Lavenberg and Shedler [11], and Lavenberg and Welch [12]. Since the technique about to be proposed is closely related to this method, we now present a brief outline of control variables before proceeding. Let $\{X_n, n \geq 0\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with unknown mean $r = \mathrm{E}[X_n]$. We shall be interested in estimating $r$ via simulation. Let $\{C_n, n \geq 0\}$ be another sequence of i.i.d. random variables with known mean $r_c$ and assume that $X_n$ and $C_n$ are correlated (usually achieved by simulating $X_n$ and $C_n$ with the same random number streams). Let $\beta$ be some constant and set $Z_n(\beta) = X_n + \beta(C_n - r_c)$. Then $\{Z_n(\beta), n \geq 0\}$ are i.i.d. with mean $r$ and variance which will be denoted by $\sigma^2(\beta)$. Defining

$$\overline{Z}_N(\beta) = \sum_{n=1}^{N} Z_n(\beta)/N,$$

by the strong law of large numbers

$$\lim_{N\to\infty} \overline{Z}_N(\beta) = r \qquad \text{almost surely (a.s.)},$$

and by the central limit theorem

$$\frac{\sqrt{N}(\overline{Z}_N(\beta) - r)}{\sigma(\beta)} \Rightarrow N(0, 1) \qquad \text{as } N \to \infty.$$

Here $\Rightarrow$ denotes weak convergence or convergence in distribution (see Billingsley [13]) and $N(0, 1)$ is a normally distributed random variable with mean 0 and variance 1. The choice $\beta = 0$ corresponds to straightforward simulation and $\sigma^2(0) = \mathrm{var}\,(X_n)$. We now pick $\beta = \beta^*$ to minimize the variance term $\sigma^2(\beta)$. It is easy to show that

$$\beta^* = -\mathrm{cov}\,(X_n, C_n)/\mathrm{var}\,(C_n),$$

$$\sigma^2(\beta^*) = [1 - \rho^2(X_n, C_n)]\,\mathrm{var}\,(X_n),$$

where $\rho(X_n, C_n)$ is the coefficient of correlation between $X_n$ and $C_n$. Since $0 \leq \rho^2(X_n, C_n) \leq 1$, a reduction in variance has been obtained and we are thus able to form shorter confidence intervals for $r$. $C_n$ is called a control variable for $X_n$. The method can be extended to allow multiple controls (see [8]).

The key things to observe about this method are that $r_c = \mathrm{E}[C_n]$ must be known and that $X_n$ and $C_n$ must be highly correlated to get large variance reductions. It is often very difficult to devise good controls, particularly if the stochastic process being simulated is quite complicated. The method to be proposed in this paper circumvents this difficulty by devising controls which will usually be highly correlated with the process of interest and for which the means of the controls need not be explicitly known. The reason for this is that the controls are chosen in such a way that their means actually equal the parameter of interest.

**Table 1** Sample sizes for the M/M/1 queue.

| $\rho$ | E[$W$] | $\sigma^2$ | $N$ |
|---|---|---|---|
| 0.10 | 0.111 | 0.375 | 8,200 |
| 0.20 | 0.250 | 1.39 | 6,020 |
| 0.30 | 0.429 | 3.96 | 5,830 |
| 0.40 | 0.667 | 10.6 | 6,430 |
| 0.50 | 1.00 | 290 | 7,850 |
| 0.60 | 1.50 | 88.5 | 10,600 |
| 0.70 | 2.33 | 335 | 16,700 |
| 0.80 | 4.00 | 1,976 | 33,400 |
| 0.90 | 9.00 | 35,901 | 119,000 |
| 0.95 | 19.0 | 607,600 | 455,000 |
| 0.99 | 99.0 | $3.95 \times 10^8$ | $1.09 \times 10^7$ |

$N$ = Number of customers that must be simulated for a 90% confidence interval for E[$W$] to have a half length of $0.10\mathrm{E}[W] = (1.645\sigma/0.10\mathrm{E}[W])^{1/2}$; $\mu$ = service rate = 1; $\lambda$ = arrival rate; $\rho = \lambda/\mu$; $\mathrm{E}[W] = \lambda/\mu(\mu - \lambda)$.

We shall restrict ourselves to studying controls for functionals of the stationary distribution of irreducible, aperiodic, positive recurrent Markov chains. The method may be extended to continuous time Markov chains, semi-Markov processes, and regenerative discrete time Markov processes on a general state space such as the waiting time process in multiple server queues in light traffic. In Section 2 we introduce notation and state preliminary results for Markov chains upon which the technique is based. Section 3 contains a description and discussion of the variance reduction technique. Numerical examples taken from queueing theory which demonstrate the use of the method are presented in Section 4.

## 2. Markov chains

Let $\mathbf{X} = \{X_n, n \geq 0\}$ be an irreducible, aperiodic, positive recurrent Markov chain with state space $E = \{0, 1, 2, \cdots\}$, transition matrix $\mathbf{P} = \{p_{ij} : i, j \in E\}$ and initial distribution $\mu = \{\mu_i : i \in E\}$ (see Chung [14] for the definitions of these terms and a more detailed analysis of Markov chains). It is well known that there then exists a probability distribution $\pi = \{\pi_i : i \in E\}$ on $E$ and a random variable $X$, having distribution $\pi$, such that $X_n \Rightarrow X$ for any initial distribution $\mu$. If $\mathbf{P}^n = \{p_{ij}^n : i, j \in E\}$ is the matrix of $n$-step transition probabilities, then

$$\lim_{n\to\infty} p_{ij}^n = \pi_j$$

for all $i$ and $j$ in $E$. The stationary distribution of $\mathbf{X}$ is $\pi$ and it satisfies the stationary equations $\pi = \pi\mathbf{P}$, which may be written componentwise as

$$\pi_j = \sum_{i\in E} \pi_i p_{ij} \qquad \text{for all } j \in E. \tag{1}$$

Let $f$ be a real-valued function on $E$ and set

$$r = \mathrm{E}[f(X)] = \pi f = \sum_{i\in E} \pi_i f(i). \tag{2}$$

We shall be interested in finding $r$ for the given function $f$. To do so we could solve the stationary equations (1) and then apply (2), but if the state space is very large (and in many interesting cases it is infinite) these equations may be quite difficult to solve numerically. In this case it becomes necessary to estimate $r$ via simulation. It is the efficient estimation of such quantities that is our concern.

The regenerative structure of Markov chains will be exploited by the variance reduction technique. This structure is now outlined (see Crane and Iglehart [15] or [16] for a more complete discussion). Since the method requires the use of a multidimensional central limit theorem, results will be stated for the multidimensional case.

For some (typically small) integer $k$ let $f_\nu$, $\nu = 0, \cdots, k$ be real-valued functions on $E$ and let $r_\nu = \pi f_\nu$. Pick some state in $E$, say 0, called the return state. Set $X_0 = 0$ and $T_0 = 0$. Define

$$T_m = \inf \{n > T_{m-1} : X_n = 0\} \qquad m \geq 1,$$

and let $\tau_m = T_m - T_{m-1}$. Because $\mathbf{X}$ is positive recurrent, $X_n = 0$ infinitely often, $\mathrm{E}[\tau_m] < \infty$, $T_m < \infty$, and $T_m$ increases to $\infty$ as $m \to \infty$. Notice that $T_m$ is the $m$th time the process enters state 0. The process is said to be in the $m$th cycle between times $T_{m-1}$ and $T_m - 1$ and the length of the $m$th cycle is $\tau_m$. The process "regenerates" itself at each time $T_m$; *i.e.*, $\{X_n, n \geq T_m\}$ has the same distribution as $\{X_n, n \geq 0\}$ and furthermore is independent of $\{X_n, n < T_m\}$. A consequence of this is that the behavior of $\mathbf{X}$ during a cycle has the same distribution and is independent of the behavior of $\mathbf{X}$ during any other cycle. The importance of this in a simulation context is that for such a process a single simulation run can be broken up into randomly spaced i.i.d. blocks, or cycles. This allows the use of classical statistical techniques in analyzing the output of the simulation. Define

$$Y_m(\nu) = \sum_{n=T_{m-1}}^{T_m - 1} f_\nu(X_n) \qquad m \geq 1, \quad \nu = 0, \cdots, k.$$

Then, by the regenerative property, $\{(\tau_m, Y_m(0), \cdots, Y_m(k)), m \geq 1\}$ is a sequence of i.i.d. random vectors. Define

$$\hat{x}_\nu(N) = \sum_{n=0}^{N} f_\nu(X_n)/N + 1,$$

$$\hat{r}_\nu(M) = \sum_{m=1}^{M} Y_m(\nu) \Big/ \sum_{m=1}^{M} \tau_m.$$

The following proposition, the proof of which may be found in [16], shows how point estimates for $r_\nu$ may be found.

- *Proposition 1*

If $\pi|f_\nu| = \sum_{i \in E} \pi_i |f_\nu(i)| < \infty$, then

$$r_\nu = \mathrm{E}[Y_m(\nu)]/\mathrm{E}[\tau_m],$$

$$\lim_{N \to \infty} \hat{x}_\nu(N) = r_\nu \qquad \text{a.s.},$$

$$\lim_{M \to \infty} \hat{r}_\nu(M) = r_\nu \qquad \text{a.s.} \tag{3}$$

Notice that $\hat{x}_\nu(N)$ estimates $r_\nu$ based on a simulation run of $N$ transitions, whereas $\hat{r}_\nu(M)$ estimates $r_\nu$ based on $M$ cycles.

We now turn to the formation of confidence intervals for $r_\nu$. Let

$$Z_m(\nu) = Y_m(\nu) - r_\nu \tau_m \qquad m \geq 1, \quad \nu = 0, \cdots, k.$$

By (3), $\mathrm{E}[Z_m(\nu)] = 0$. Let

$$\sigma_{ij} = \mathrm{E}[Z_m(i) Z_m(j)] \qquad 0 \leq i, \quad j \leq k,$$

$\mathbf{\Sigma}_k = \{\sigma_{ij} : 0 \leq i, j \leq k\}$ and $\sigma_i^2 = \sigma_{ii}$. We will assume that

$$\mathrm{E}\left[\left(\sum_{n=0}^{T_i - 1} |f_\nu(X_n)|\right)^2\right] < \infty, \tag{4}$$

$$\mathrm{E}[\tau_1^2] < \infty, \text{ and} \tag{5}$$

$$\mathbf{\Sigma}_k \text{ is positive definite.} \tag{6}$$

The elements of $\mathbf{\Sigma}_k$ exist and are finite by (4) and (5). In general $\mathbf{\Sigma}_k$ will be positive semidefinite; we assume (6) because the variance reduction technique requires the existence of $\mathbf{\Sigma}_k^{-1}$. In most simulations these assumptions will not be a restriction.

Let $\mathbf{N(0, A)}$ denote a random vector having the multivariate normal distribution with means 0 and covariance matrix $\mathbf{A} = \{a_{ij}\}$. Let $\mathbf{A}'$ denote the transpose of $\mathbf{A}$ and for any $a \neq 0$ let $\mathbf{A}/a$ be the matrix with elements $a_{ij}/a$. Let $\hat{\mathbf{x}}\mathbf{(N)}$, $\hat{\mathbf{r}}\mathbf{(M)}$ and $\mathbf{r}$ be $(k+1)$-dimensional column vectors with $\nu$th entries $\hat{x}_\nu(N)$, $\hat{r}_\nu(M)$, and $r_\nu$, respectively. The following multidimensional central limit theorems are a direct consequence of the i.i.d. structure of cycles.

- *Proposition 2*

If $\pi|f_\nu| < \infty$ for $\nu = 0, \cdots, k$ and if (4), (5) and (6) hold, then

$$\sqrt{N}(\hat{\mathbf{x}}(N) - \mathbf{r}) \Rightarrow \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_k/\mathrm{E}[\tau_1]) \qquad \text{as } N \to \infty,$$

$$\sqrt{M}(\hat{\mathbf{r}}(M) - \mathbf{r}) \Rightarrow \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_k/\mathrm{E}[\tau_1]^2) \qquad \text{as } M \to \infty. \tag{7}$$

*Proof* Apply the Cramér-Wold device described on page 48 of [13] to the one-dimensional central limit theorem given in [16] to yield the multidimensional central limit theorem. □

Now let $\boldsymbol{\beta}$ be a $(k+1)$-dimensional row vector. The following proposition is a corollary to Proposition 2.

- *Proposition 3*

Let $\sigma_k^2(\boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{\Sigma}_k\boldsymbol{\beta}' = \sum_{i=0}^{k}\sum_{j=0}^{k}\beta(i)\,\sigma_{ij}\beta(j)$.

Under the hypotheses of Proposition 2,

$$\frac{\sqrt{N}(\boldsymbol{\beta}\hat{\mathbf{x}}(N) - \boldsymbol{\beta}\mathbf{r})}{\sigma_k(\boldsymbol{\beta})/\mathrm{E}[\tau_1]^{1/2}} \Rightarrow N(0,1) \qquad \text{as } N \to \infty,$$

$$\frac{\sqrt{M}(\boldsymbol{\beta}\hat{\mathbf{r}}(M) - \boldsymbol{\beta}\mathbf{r})}{\sigma_k(\boldsymbol{\beta})/\mathrm{E}[\tau_1]} \Rightarrow N(0,1) \qquad \text{as } M \to \infty. \tag{8}$$

Note that, for example,

$$\boldsymbol{\beta}\mathbf{r} = \sum_{\nu=0}^{k}\beta(\nu)r_\nu.$$

In a simulation the covariance matrix $\boldsymbol{\Sigma}_k$ is usually unknown and must be estimated. In addition $\boldsymbol{\beta}$ may be unknown (as is usually the case in control variable schemes). However, the central limit theorems in (8) remain valid if $\mathrm{E}[\tau_1]$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\beta}$ are replaced by any strongly consistent estimates (see Heidelberger [17]). By setting $\beta(\nu) = 0$ for $\nu \neq i$ and $\beta(i) = 1$, central limit theorems for individual $r_i$'s are obtained from (8). Confidence intervals for $r_i$ can then be formed based on this central limit theorem. While the regenerative method was described here for discrete time Markov chains, analogous results hold true for continuous time Markov chains, semi-Markov processes, and regenerative discrete time Markov processes on a general state space.

## 3. Variance reduction techniques

In this section we apply the results of Section 2 and take further advantage of the structure of Markov chains to obtain variance reductions. Let $f$ now be a fixed real-valued function on $E$ and as before let $r = \pi f$. Our goal is to obtain both point estimates and "short" confidence intervals for $r$. This will be achieved by forming several estimates for $r$ and then taking the (asymptotic) minimum variance linear combination of these estimates which is strongly consistent. In order to form these multiple estimates some additional calculations must be done both before and during the simulation, but hopefully their cost will not be so great as to prohibit the use of this method.

The multiple estimates for $r$ are formed by defining new functions $f_\nu$ on $E$ so that $\pi f_\nu = r$ for each value of $\nu$. The values of $\nu$ for which this is done will typically be small and labeled by $\{0, 1, \cdots, k\}$. Once the $f_\nu$'s have been computed the process is simulated for, say, $M$ cycles. Define $Y_m(\nu)$, $\tau_m$, and $\hat{r}_\nu(M)$ as in Section 2. Since, assuming $\pi|f_\nu| < \infty$, $\hat{r}_\nu(M) \to \mathrm{E}[Y_m(\nu)]/\mathrm{E}[\tau_m] = \pi f_\nu = r$, each $\hat{r}_\nu(M)$ is a strongly consistent estimator for $r$ so that any one of them could be used to estimate $r$. Actually we can do significantly better than that by using $\{\hat{r}_0(M), \cdots, \hat{r}_k(M)\}$ simultaneously to estimate $r$. If $\{\beta(\nu) : \nu = 0, \cdots, k\}$ are any constants so that

$$\sum_{\nu=0}^{k}\beta(\nu) = 1 \tag{9}$$

and $\hat{r}_{\boldsymbol{\beta}}(M)$ is defined by

$$\hat{r}_{\boldsymbol{\beta}}(M) = \sum_{\nu=0}^{k}\beta(\nu)\hat{r}_\nu(M), \tag{10}$$

then $\hat{r}_{\boldsymbol{\beta}}(M) \to r$ a.s. as $M \to \infty$. The values of $\beta(\nu)$ are then chosen to minimize the asymptotic variance of $\hat{r}_{\boldsymbol{\beta}}(M)$. Details of the choice of the $\beta(\nu)$'s will be presented later.

We now turn to the selection of the functions $f_\nu$. In this paper we will concentrate on only one (actually the simplest) way to choose the $f_\nu$'s. In a subsequent paper we will study alternate methods of choosing the $f_\nu$'s in the case when the state space is finite.

As our current choice of $f_\nu$, let $f_\nu = \mathbf{P}^\nu f$, for $\nu = 0, \cdots, k$, so that

$$f_\nu(i) = \sum_{j\in E} p_{ij}^\nu f(j). \tag{11}$$

Recall that $\mathbf{P}^\nu$ is the $\nu$-step transition matrix of the process. If, as before, $r_\nu = \pi f_\nu$, then we must show that $r_\nu = r$ for each value of $\nu$. Under the assumption that $\pi|f| < \infty$ (see [17]), we have

$$\begin{aligned} r_\nu = \pi f_\nu &= \pi(\mathbf{P}^\nu f) \\ &= \pi\mathbf{P}(\mathbf{P}^{\nu-1}f) \qquad \text{(assumption allows interchange)} \\ &= \pi(\mathbf{P}^{\nu-1}f) \qquad \text{(since } \pi\mathbf{P} = \pi\text{)} \\ &= \pi f_{\nu-1} = r_{\nu-1}, \end{aligned}$$

so all the $r_\nu$'s are equal. Noting that $f_0 = \mathbf{P}^0 f = f$, then $\pi f_0 = \pi f = r$, and so $r_\nu = r$ for all $\nu$.

We now return to the minimization of the asymptotic variance of the estimator $\hat{r}_{\boldsymbol{\beta}}(M)$ defined in Eq. (10). By Proposition 3 the following central limit theorem holds:

$$\frac{\sqrt{M}\left(\sum_{\nu=0}^{k}(\beta(\nu)\hat{r}_\nu(M) - \beta(\nu)r)\right)}{\sigma_k(\boldsymbol{\beta})/\mathrm{E}[\tau_1]} \Rightarrow N(0,1) \text{ as } M \to \infty,$$

where $\sigma_k(\boldsymbol{\beta})$ is defined in Proposition 3. Since the $\beta(\nu)$'s sum to one [Eq. (9)], this may be rewritten as

$$\frac{\sqrt{M}(\hat{r}_{\boldsymbol{\beta}}(M) - r)}{\sigma_k(\boldsymbol{\beta})/\mathrm{E}[\tau_1]} \Rightarrow N(0,1),$$

so that confidence intervals for $r$ can be formed based on

$\hat{r}_{\boldsymbol{\beta}}(M)$. Since we are free to pick $\boldsymbol{\beta}$ in any way we please [subject to (9)], we select $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^*$ minimizes the variance term $\sigma_k^2(\boldsymbol{\beta})$. This will produce the smallest possible confidence intervals for $r$. Note that there is no reason to restrict $\boldsymbol{\beta}$ to be nonnegative; *i.e.*, $\hat{r}_{\boldsymbol{\beta}}(M)$ need not be a convex combination of the $r_\nu(M)$'s. To minimize the variance the following nonlinear programming problem must be solved:

$$\text{minimize } \boldsymbol{\beta}\boldsymbol{\Sigma}_k\boldsymbol{\beta}'$$

$$\text{subject to } \mathbf{e}\boldsymbol{\beta}' = 1, \tag{12}$$

where $\mathbf{e}$ denotes a $(k+1)$-dimensional row vector each of whose components is one. It is straightforward (using Lagrange multipliers) to show that

$$\boldsymbol{\beta}^* = \mathbf{e}\boldsymbol{\Sigma}_k^{-1}/\mathbf{e}\boldsymbol{\Sigma}_k^{-1}\mathbf{e}', \tag{13}$$

$$\sigma_k^2(\boldsymbol{\beta}^*) = 1/\mathbf{e}\boldsymbol{\Sigma}_k^{-1}\mathbf{e}'. \tag{14}$$

The general idea of combining multiple estimates for the same quantity in this particular manner is outlined on page 19 of Hammersley and Handscomb [18]. This basic technique has been successfully applied to estimate work rates and waiting times in closed queueing networks (see [4, 7, 11]). The contribution of this paper is to describe how this general technique may be applied to an entire class of simulations. This general technique may also be reformulated as a standard control variable application with $k-1$ controls with mean zero, the $j$th control being $\hat{r}_j(M) - \hat{r}_0(M)$ (see, *e.g.*, [12]). The relationship between control variables and linear regression is explored in [10].

We have dealt here with the formation of short confidence intervals based on a run length of $M$ cycles. The same technique applies to a run of $N$ transitions. In this case the nonlinear program (12) must still be solved because the variance terms in the central limit theorems (8) differ only by a constant multiple. Equations (13) and (14) are therefore still valid in this case.

Since the covariance matrix is in general unknown it becomes necessary to estimate $\boldsymbol{\Sigma}_k$. If $\hat{\boldsymbol{\Sigma}}_k(M)$ is any estimate such that $\hat{\boldsymbol{\Sigma}}_k(M) \to \boldsymbol{\Sigma}_k$ a.s. as $M \to \infty$, then $\hat{\boldsymbol{\Sigma}}_k(M)^{-1} \to \boldsymbol{\Sigma}_k^{-1}$. Letting

$$\hat{\boldsymbol{\beta}}^*(M) = \mathbf{e}\boldsymbol{\Sigma}_k(M)^{-1}/\mathbf{e}\boldsymbol{\Sigma}_k(M)^{-1}\mathbf{e}',$$

it is clear that $\hat{\boldsymbol{\beta}}^*(M) \to \boldsymbol{\beta}^*$ a.s. as $M \to \infty$. Thus by the results of Section 2 the asymptotic normality can be maintained even when $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\beta}^*$ must be estimated. For the simulations reported in Section 4, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\beta}^*$ were estimated from data collected over the entire run. It has been suggested (in [8]) that in control variable schemes the multipliers should be estimated from only a fraction of the cycles simulated. While this will generally result in less variance reduction, the confidence intervals will tend to cover $r$ a greater (truer) percentage of the time. This effect has been quantified for nonregenerative simulations in [10].

In order to apply this method, the functions $f_\nu$ must be computed (usually before the start of the simulation). For computational efficiency $f_\nu$ can be defined recursively by $f_0 = f$ and $f_\nu = \mathbf{P}f_{\nu-1}$ for $\nu \geq 1$. This avoids having to compute the $\nu$-step transition function $\mathbf{P}^\nu$, a potentially great computational savings. If the state space is finite and the transition matrix is sparse, the work involved in calculating $f_\nu$ for a few values of $\nu$ may not be too great. If the $f_\nu$'s are computed before the start of the simulation they must be stored; this may be a considerable problem if the state space is very large.

We note that to form the estimates $\hat{x}_\nu(N)$ [or $\hat{r}_\nu(M)$], $f_\nu(X_n)$ must be evaluated for each value of $\nu$ and each transition $n$. This will tend to increase the amount of time needed for each transition simulated. However, if the variance reduction obtained is sufficiently large, the potential savings in the number of transitions that need to be simulated will more than offset the extra work per transition. This will be discussed in greater detail in Section 4. We also note that additional work must be done at the end of each cycle to update the estimates of the covariance matrix $\boldsymbol{\Sigma}_k$ (using no variance-reducing technique only the estimate for $\sigma_0^2$ need be updated). However, this computation will usually be insignificant compared to that of the simulation.

It should be mentioned that if one has a choice of more than one return state, the variance reductions that are obtained using this method are (in theory) independent of the return state. The reader should consult page 99 of [14] for this point. For practical reasons it is recommended that, if possible, the return state be chosen so that cycles are not excessively long.

This method can be extended to certain types of continuous time processes such as continuous time Markov chains and semi-Markov processes. This may be accomplished either by transforming the continuous time process into an appropriate discrete time Markov chain (see Hordijk, Iglehart, and Schassberger [19] for a description of this transformation) or by working with the continuous time process directly. The interested reader should consult [17] for details of this extension. The method may also be applied to discrete time regenerative Markov processes on a general state space. In this case integrals replace the summations in Eqs. (1), (2), (3), and (11). Again $f_\nu = \mathbf{P}^\nu f$, but now

$$f_\nu(x) = \int_E \mathbf{P}^\nu(x, dy)\, f(y) \qquad \text{for all } x \in E,$$

where $\mathbf{P}^\nu(x, A) = \mathbf{P}\{X_{n+\nu} \in A | X_n = x\}$ for measurable sets $A$. Section 4 reports numerical results for two such processes, the waiting time processes in the M/M/1 and M/M/2 queues.

We now examine the selection of the parameter $k$. As the value of $k$ increases, $\sigma_k^2(\boldsymbol{\beta}^*)$ decreases. This is because the $k$th minimization problem is the same as the $(k+1)$st problem which has the additional constraint that $\beta(k+1) = 0$. This means that as we do more computation we can get increasingly accurate estimates of $r$.

If the state space is finite, then

$$f_k(i) = \sum_{j \in E} p_{ij}^k f(j) \to \sum_{j \in E} \pi_j f(j) = r \qquad \text{as } k \to \infty,$$

so that for large values of $k$, the value of $f_k(i)$ will be close to $r$ for each state $i$. The point estimate $\hat{x}_k(N)$ is then the average of $N+1$ terms each of which is close to $r$ so that, for large $k$, $\hat{x}_k(N)$ will have a smaller variance than $\hat{x}_0(N)$. In fact it can be shown that $\sigma_k^2 \to 0$ as $k \to \infty$. By placing all weight on $\hat{x}_k(N)$, *i.e.*, by setting $\beta(k) = 1$ and $\beta(\nu) = 0$ for $\nu \neq k$, then $\sigma_k^2(\boldsymbol{\beta}^*) \leq \sigma_k^2(0, 0, \cdots, 1) = \sigma_k^2 \to 0$. (Thus if an infinite amount of work is performed in advance, there is no need to simulate at all.)

For many types of Markov chains substantial variance reductions can be expected even when $k$ is relatively small (say 2 or 3). If the Markov chain makes transitions only to "neighboring" states and if $f(j)$ is close to $f(i)$ for $j$ "near" to $i$, then for small $k$, $f_k(i)$ and $f_0(i)$ should be nearly the same. This means that $\hat{x}_k(N)$ and $\hat{x}_0(N)$ will be highly correlated, a condition that generally results in good variance reduction. Many Markovian queueing networks exhibit this special type of structure.

Ideally one would like to select the optimal value of $k$ in the sense that for a given computer budget the value of $k$ which yields the smallest confidence intervals for $r$ is picked (part of the budget must be allocated to the computation of $f_0, \cdots, f_k$). This is an open and seemingly difficult problem. It is felt that the inability to predict the variance reductions in advance is the major drawback of the method (this is also a drawback in other more standard control variable techniques). The simulator must be very careful to apply the method only in those situations when it is computationally efficient to do so. Generally speaking the success of this technique depends on one's ability to efficiently compute and store the functions $f_\nu$.

## 4. Examples

To gain insight into the use of the method, four test problems were chosen for numerical studies. These problems all come from the area of queueing theory. They are the queue length process in the finite-capacity M/M/1 queue, the queue length process in the repairman problem with spares, and the waiting time processes in both the M/M/1 and M/M/2 queues. These processes were chosen because analytic results are readily available, thereby making a comparison between analytic and simulation results possible. Despite their simplicity they are by no means "easy" processes to simulate, particularly the heavily loaded queues which require very long run lengths (see Table 1) to get good simulation estimates. For all four processes substantial variance reductions have been realized (although in the M/M/2 queue with $\rho = 0.9$ the reduction in variance is not enough to justify the use of the method). While it is difficult to predict the variance reductions that will be obtained for a particular stochastic process, it is felt that this method shows a great deal of promise and deserves serious consideration when a simulation experiment is being planned. The remainder of this section will be devoted to a more detailed description of the examples and a presentation of their numerical results.

### *Birth and death processes*

The first two examples, the queue length processes in the finite-capacity M/M/1 queue and the repairman problem with spares, are both birth and death processes. The M/M/1 queue has birth and death parameters

$$\lambda_i = \begin{cases} \lambda & 0 \leq i \leq C-1, \\ 0 & i \geq C, \end{cases}$$

$$\mu_i = \mu \qquad 1 \leq i \leq C,$$

where $0 < \lambda, \mu < \infty$ and $C$ is the (finite) capacity of the system. For this example we were interested in estimating $\mathrm{E}[X]$, the expected stationary number of customers in the queue, so the appropriate function $f$ is $f(i) = i$. Let $\rho = \lambda/\mu$.

The repairman problem has parameters

$$\lambda_i = \begin{cases} n\lambda & 0 \leq i \leq m, \\ (n+m-i)\lambda & m < i \leq n+m, \end{cases}$$

$$\mu_i = \begin{cases} i\mu & 1 \leq i \leq s, \\ s\mu & s < i \leq m+n, \end{cases}$$

where $n$ is the number of operating units, $m$ is the number of spare units, $s$ is the number of repairmen, and $\lambda$ and $\mu$ are the failure and repair rates, respectively, of the units. For this process we chose $f(i) = i^2$ so that $r = \mathrm{E}[X^2]$. Both

**Table 2** Calculated variance reductions for finite-capacity M/M/1 queue: $C = 14$, $r = \mathrm{E}[X]$.

| $\rho$ | $\mathrm{E}[X]$ | $\sigma_0^2$ | $R_1^2$<br>$R_1$ | $R_2^2$<br>$R_2$ | $R_3^2$<br>$R_3$ |
|---|---|---|---|---|---|
| 0.10 | 0.1111 | 0.0244 | 0.0454 | 0.0045 | 0.0005 |
| | | | 0.2136 | 0.0674 | 0.0213 |
| 0.20 | 0.2500 | 0.2812 | 0.0926 | 0.0185 | 0.0037 |
| | | | 0.3043 | 0.1361 | 0.0609 |
| 0.30 | 0.4286 | 1.469 | 0.1413 | 0.0424 | 0.0127 |
| | | | 0.3759 | 0.2058 | 0.1126 |
| 0.40 | 0.6667 | 5.888 | 0.1905 | 0.0756 | 0.0297 |
| | | | 0.4364 | 0.2749 | 0.1725 |
| 0.50 | 0.9995 | 21.59 | 0.2341 | 0.1121 | 0.0524 |
| | | | 0.4838 | 0.3347 | 0.2288 |
| 0.60 | 1.493 | 76.57 | 0.2601 | 0.1352 | 0.0681 |
| | | | 0.5100 | 0.3677 | 0.2610 |
| 0.70 | 2.262 | 250.9 | 0.2884 | 0.1395 | 0.0683 |
| | | | 0.5370 | 0.3734 | 0.2613 |
| 0.80 | 3.453 | 670.2 | 0.4094 | 0.1623 | 0.0692 |
| | | | 0.6399 | 0.4028 | 0.2631 |
| 0.90 | 5.111 | 1262 | 0.6050 | 0.2659 | 0.1148 |
| | | | 0.7778 | 0.5156 | 0.3388 |
| 0.95 | 6.052 | 1476 | 0.6880 | 0.3425 | 0.1607 |
| | | | 0.8294 | 0.5852 | 0.4009 |
| 0.99 | 6.813 | 1548 | 0.7404 | 0.4056 | 0.2047 |
| | | | 0.8605 | 0.6369 | 0.4525 |

$R_k$ = percent reduction in confidence interval width given equal run length; $R_k^2$ = percent reduction in run length given equal confidence interval width.

of these continuous time processes were transformed into discrete time using the methods of [19] and systems of linear equations were solved to find $r$, $\boldsymbol{\Sigma}_k$, $\boldsymbol{\beta}^*$, and $\sigma_k^2(\boldsymbol{\beta}^*)$. Thus Tables 2 and 3 report theoretically calculated variance reductions, not simulation results.

For the numerical results that follow, the definition of $\sigma_k^2(\boldsymbol{\beta}^*)$ has been slightly modified to make it the asymptotic variance term in the central limit theorem for

$$\hat{x}_{\boldsymbol{\beta}^*}(N) = \sum_{\nu=0}^{k} \beta^*(\nu)\hat{x}_\nu(N).$$

It therefore takes into account all constant multiples such as $\mathrm{E}[\tau_1]^{1/2}$. Let $R_k^2 = \sigma_k^2(\boldsymbol{\beta}^*)/\sigma_0^2$. To obtain confidence intervals of equal length, if we use the estimator $\hat{x}_{\boldsymbol{\beta}^*}$ we need only simulate $R_k^2$ times as many transitions as would be needed using no variance reduction technique (that is, if we used just the regular point estimate $\hat{x}_0$). For a fixed (large) number of simulated transitions $N$, the length of the confidence interval for $r$ using $\hat{x}_{\boldsymbol{\beta}^*}(N)$ divided by the length of the confidence interval for $r$ using $\hat{x}_0(N)$ is $R_k = \sigma_k(\boldsymbol{\beta}^*)/\sigma_0$. The quantities $R_k^2$ and $R_k$ are the usual efficiency measures of a variance reduction technique.

Tables 2 and 3 list $r$, $\sigma_0^2$, $R_k^2$, and $R_k$ for $k = 1, 2, 3$. For each $k$, $R_k$ is listed directly below $R_k^2$. As an example, in Table 2, for $\rho = 0.5$ we see that to obtain confidence intervals for $\mathrm{E}[X] = 0.9995$ of equal length, we need only simulate 5.24% as many transitions using $\hat{x}_{\boldsymbol{\beta}^*}$ (with $k = 3$) rather than using just $\hat{x}_0$. For the same number of transitions simulated the ratio of the lengths of confidence intervals is 0.2288. Notice that in Table 2 the value of $\rho$ influences the variance reductions. As the traffic intensity $\rho$ increases, the variance reduction obtained decreases. For the repairman problem the variance reductions are more or less constant over the entire range of parameters tested. Variance reductions for different functions $f$ are reported in [17]. They follow a pattern similar to those in Tables 2 and 3.

### *Waiting time process in an M/M/1 queue*

We now turn to an example of a regenerative Markov process with an uncountable state space $E = [0, \infty)$. Let $W_n$ and $S_n$ be the waiting and service times, respectively, of the $n$th customer in a GI/G/1 queue and let $\{A_n, n \geq 0\}$ be the sequence of i.i.d. interarrival times. Set $X_n = S_{n-1} - A_n$. Assuming the queue is initially empty, the waiting time process $\{W_n, n \geq 0\}$ is defined by

$$W_n = \begin{cases} 0 & n = 0, \\ (W_{n-1} + X_n)^+ & n \geq 1, \end{cases}$$

where for any real number $a$, $a^+$ denotes the maximum of 0 and $a$.

It is known that if the traffic intensity $\rho < 1$, there exists an infinite number of indices $n$ such that $W_n = 0$ and the expected time between any two such consecutive indices

**Table 3** Calculated variance reductions for repairman problem: $n = 10$, $m = 4$, $\lambda = 1$, $r = E[X^2]$.

| $(s, \mu)$ | $E[X^2]$ | $\sigma_0^2$ | $R_1^2$<br>$R_1$ | $R_2^2$<br>$R_2$ | $R_3^2$<br>$R_3$ |
|---|---|---|---|---|---|
| 1, 12 | 13.46 | 9,610 | 0.0836 | 0.0156 | 0.0067 |
| | | | 0.2892 | 0.1251 | 0.0819 |
| 2, 6 | 15.06 | 9,377 | 0.0674 | 0.0172 | 0.0083 |
| | | | 0.2595 | 0.1313 | 0.0911 |
| 3, 4 | 17.28 | 9,009 | 0.0532 | 0.0177 | 0.0104 |
| | | | 0.2306 | 0.1329 | 0.1018 |
| 4, 3 | 20.01 | 8,568 | 0.0445 | 0.0217 | 0.0108 |
| | | | 0.2109 | 0.1473 | 0.1039 |
| 1, 9 | 31.66 | 28,346 | 0.1659 | 0.0351 | 0.0134 |
| | | | 0.4073 | 0.1874 | 0.1157 |
| 2, 4.5 | 32.85 | 26,287 | 0.1420 | 0.0290 | 0.0117 |
| | | | 0.3768 | 0.1702 | 0.1083 |
| 3, 3 | 34.51 | 23,787 | 0.1160 | 0.0231 | 0.0108 |
| | | | 0.3406 | 0.1521 | 0.1041 |
| 4, 2.25 | 36.59 | 21,153 | 0.0928 | 0.0199 | 0.0113 |
| | | | 0.3046 | 0.1411 | 0.1061 |
| 1, 6 | 69.25 | 33,944 | 0.1897 | 0.0629 | 0.0248 |
| | | | 0.4356 | 0.2509 | 0.1573 |
| 2, 3 | 69.43 | 33,120 | 0.1804 | 0.0563 | 0.0208 |
| | | | 0.4247 | 0.2373 | 0.1443 |
| 3, 2 | 69.74 | 31,904 | 0.1676 | 0.0481 | 0.0165 |
| | | | 0.4094 | 0.2193 | 0.1285 |
| 4, 1.5 | 70.21 | 30,281 | 0.1523 | 0.0397 | 0.0129 |
| | | | 0.3902 | 0.1992 | 0.1134 |

$R_k$ = percent reduction in confidence interval width given equal run length; $R_k^2$ = percent reduction in run length given equal confidence interval width.

is finite. Thus 0 is chosen to be the return state and regenerations occur whenever a customer arrives at an empty queue. Therefore, for $\rho < 1$ there exists a random variable $W$ such that $W_n \Rightarrow W$. For more details on this queue see for example [2].

We shall be interested in estimating $E[W]$, which is finite if $E[S_n^2] < \infty$. The appropriate function $f$ is then $f(x) = x$. To calculate $f_\nu = \mathbf{P}^\nu f$ we need to find the transition function of the process. We illustrate this for the M/M/1 queue. For M/M/1 the calculations are straightforward. The approach generalizes easily to the GI/G/1 queue, although one's ability to carry out the computations in practice depends on the distributions of the service and interarrival times. Numerical integration techniques may be of use here although, in theory, one must calculate $\mathbf{P}^\nu f(x)$ for all values of $x \geq 0$.

For the M/M/1 queue it is easy to show that

$$P\{X_n \leq y\} = \begin{cases} \dfrac{\mu}{\lambda + \mu} e^{\lambda y} & \text{for } y < 0, \\ 1 - \dfrac{\lambda}{\lambda + \mu} e^{-\mu y} & \text{for } y \geq 0, \end{cases}$$

where $\lambda$ is the arrival rate and $\mu^{-1}$ is the mean service time. Thus $g(y) = (d/dy)\, P\{X_n \leq y\}$ exists for all $y$ and we write $P\{X_n \in dy\} = g(y)dy$ where

$$g(y) = \begin{cases} g_-(y) = \dfrac{\lambda\mu}{\lambda + \mu} e^{\lambda y} & y < 0, \\ g_+(y) = \dfrac{\lambda\mu}{\lambda + \mu} e^{-\mu y} & y \geq 0. \end{cases}$$

Now to evaluate $f_1(x)$,

$$\begin{aligned} f_1(x) &= \int_{[0,\infty)} \mathbf{P}(x, dy) f(y) \\ &= \int_{[0,\infty)} y\mathbf{P}\{(W_n + X_{n+1})^+ \in dy | W_n = x\} \\ &= \int_{(0,\infty)} y\mathbf{P}\{x + X_{n+1} \in dy\} = \int_{(0,\infty)} yg(y - x)dy \\ &= \int_{y=0}^{x} yg_-(y - x)dy + \int_{y=x}^{\infty} yg_+(y - x)dy. \end{aligned}$$

Evaluation of these integrals is straightforward and we find that

$$f_1(x) = x + \frac{\lambda - \mu}{\mu\lambda} + \frac{\mu}{\lambda(\lambda + \mu)} e^{-\lambda x}. \tag{15}$$

The calculation of $f_2(x)$ is similar; however, care must be taken to include the atom at 0. It is found that

**Table 4** Calculated variance reductions for the waiting time process in an M/M/1 queue: $r = E[W]$, $\mu = 1$, $\lambda = \rho$, $E[W] = \lambda/\mu(\mu - \lambda)$.

| $\rho$ | $E[W]$ | $\sigma_0^2$ | $R_1^2$<br>$R_1$ | $R_2^2$<br>$R_2$ |
|---|---|---|---|---|
| 0.10 | 0.1111 | 0.375 | 0.0070<br>0.0838 | 0.0001<br>0.0008 |
| 0.20 | 0.250 | 1.39 | 0.0265<br>0.1628 | 0.0009<br>0.0306 |
| 0.30 | 0.429 | 3.96 | 0.0568<br>0.2383 | 0.0045<br>0.0672 |
| 0.40 | 0.667 | 10.6 | 0.0968<br>0.3111 | 0.0137<br>0.1173 |
| 0.50 | 1.00 | 29.0 | 0.1457<br>0.3817 | 0.0327<br>0.1808 |
| 0.60 | 1.50 | 88.5 | 0.2029<br>0.4505 | 0.0664<br>0.2577 |
| 0.70 | 2.33 | 336 | 0.2678<br>0.5175 | 0.1214<br>0.3485 |
| 0.80 | 4.00 | 1,976 | 0.3397<br>0.5828 | 0.2055<br>0.4533 |
| 0.90 | 9.00 | 35,901 | 0.4175<br>0.6461 | 0.3280<br>0.5727 |
| 0.95 | 19.00 | 607,601 | 0.4582<br>0.6769 | 0.4072<br>0.6381 |
| 0.99 | 99.00 | $3.96 \times 10^8$ | 0.4904<br>0.7003 | 0.4686<br>0.6845 |

$R_k$ = percent reduction in confidence interval width given equal run length; $R_k^2$ = percent reduction in run length given equal confidence interval width.

$$f_2(x) = f_1(x) + \frac{\lambda - \mu}{\mu\lambda} + \left(\frac{\mu}{\lambda + \mu}\right)^2 e^{-\lambda x}\left(\frac{1}{\lambda} + \frac{1}{\lambda + \mu} + x\right). \quad (16)$$

In this case it is possible to calculate exactly the covariance matrix $\Sigma_2$ so that exact results for the variance reductions can be obtained (see [17]). Table 4 lists the calculated variance reductions for this queue. Again substantial variance reductions are realized although the method is less effective for high values of $\rho$.

These theoretical computations were then compared with results obtained in actual simulations. For $\rho = 0.5$ the queue was simulated for a total of 200,000 cycles (the expected number of customers simulated for this run is 400,000). The random number generator described in Learmonth and Lewis [20] was used. These 200,000 cycles were then broken up into $R$ independent replications of $C$ cycles per replication for several combinations of $R$ and $C(RC = 200{,}000)$. At the end of each replication, point estimates for the various parameters of interest were formed. The figures reported in Table 5 are then the sample averages of the point estimates taken over the $R$ independent replications. Approximate 95% confidence intervals for each parameter (formed in the usual manner using the $R$ i.i.d. replications) are given directly below the point estimates. As an example for $R = 200$ and $C = 1000$ a 95% confidence interval for $r = E[W] = 1.000$ based on the estimate $\hat{r}_0$ is $(1.015 - 0.016, 1.015 + 0.016)$.

At the end of each replication 95% confidence intervals for $r$ were formed based on each point estimate for $r$ using the central limit theorem of Proposition 3. The fraction of these confidence intervals that actually contained $r$ is given in Table 6. If valid conference intervals are being formed, this fraction (called a coverage) should be approximately 0.95. Directly below each coverage is a 95% confidence interval for the coverage based on the normal approximation to the binomial distribution (see Appendix 3 of Lavenberg and Sauer [21]). Generally speaking, the coverage increases with the run length $C$. Because of this behavior care must be taken that the run length is not too short; otherwise unjustified confidence may be placed in the estimates. This suggests that the method may be better used to produce very tight confidence intervals from moderately long run lengths rather than to reduce the run length by a significant factor. The sequential techniques developed in [21] for determining run lengths are applicable and may be of practical value in this area.

In order to use this method the functions $f_0(W_n)$, $f_1(W_n)$, and $f_2(W_n)$ must be evaluated for each customer $n$. To get a measure of the computational savings (if any) of the method it is important to determine how much additional work is required for each customer. From (15) and (16) it

**Table 5** Simulation results for waiting time process in M/M/1 queue with $\rho = 0.5$, point estimates and 95% confidence intervals.

| *Parameter* | *True value* | *R = 200 C = 1000* | *R = 100 C = 2000* | *R = 50 C = 4000* | *R = 1 C = 200,000* |
|---|---|---|---|---|---|
| $\hat{r}_0$ | 1.000 | 1.015 | 1.016 | 1.018 | 1.018 |
| | | 0.016 | 0.017 | 0.015 | |
| $\hat{r}_1$ | 1.000 | 1.013 | 1.014 | 1.015 | 1.015 |
| | | 0.013 | 0.014 | 0.012 | |
| $\hat{r}_2$ | 1.000 | 1.011 | 1.011 | 1.012 | 1.013 |
| | | 0.011 | 0.011 | 0.010 | |
| $\hat{r}_{\hat{\beta}^*}(k = 1)$ | 1.000 | 0.999 | 1.000 | 1.003 | 1.003 |
| | | 0.006 | 0.007 | 0.006 | |
| $\hat{r}_{\hat{\beta}^*}(k = 2)$ | 1.000 | 0.996 | 0.998 | 0.999 | 1.001 |
| | | 0.003 | 0.003 | 0.003 | |
| $R_1^2$ | 0.146 | 0.106 | 0.116 | 0.125 | 0.135 |
| | | 0.005 | 0.007 | 0.007 | |
| $R_2^2$ | 0.033 | 0.015 | 0.018 | 0.022 | 0.026 |
| | | 0.001 | 0.002 | 0.003 | |
| $\sigma_0^2$ | 29.0 | 30.23 | 30.56 | 30.83 | 30.98 |
| | | 2.53 | 2.67 | 2.34 | |
| $\sigma_{01}$ | 23.67 | 24.67 | 24.95 | 25.18 | 25.31 |
| | | 2.21 | 2.34 | 2.05 | |
| $\sigma_{02}$ | 19.48 | 20.30 | 20.53 | 20.73 | 20.83 |
| | | 1.93 | 2.04 | 1.80 | |
| $\sigma_1^2$ | 19.48 | 20.30 | 20.53 | 20.73 | 20.84 |
| | | 1.94 | 2.05 | 1.81 | |
| $\sigma_{12}$ | 16.14 | 16.79 | 16.99 | 17.16 | 17.25 |
| | | 1.69 | 1.79 | 1.59 | |
| $\sigma_2^2$ | 13.44 | 13.95 | 14.12 | 14.27 | 14.34 |
| | | 1.48 | 1.56 | 1.39 | |
| $\sigma_1^2(\boldsymbol{\beta}^*)$ | 4.23 | 3.46 | 3.74 | 3.98 | 4.18 |
| | | 0.40 | 0.46 | 0.46 | |
| $\sigma_2^2(\boldsymbol{\beta}^*)$ | 0.948 | 0.505 | 0.603 | 0.699 | 0.792 |
| | | 0.079 | 0.104 | 0.119 | |

is seen that to evaluate $f_1$ and $f_2$ one exponential and several multiplications and additions must be computed for each customer. From CPU times we estimated that on the average each customer using multiple estimates requires 5/3 as much CPU time as a customer on a run using no variance reduction technique. Since, for $\rho = 0.5$, we need only simulate 0.03 times as many customers to get confidence intervals of equal length (see Table 4), our computational savings, which is defined as the ratio of CPU times needed to obtain equally accurate estimates, is 0.05 ($= 0.03 \times 5/3$). For $\rho = 0.9$ we estimate the computational savings to be 0.55 ($= 5/3 \times 0.33$). On the other hand, if the CPU time is fixed to be the same for each method, what is the statistical savings (defined to be the ratio of confidence interval lengths for equal CPU times)? Suppose for a specified CPU time we can simulate $N_1$ customers using no variance reduction technique (method 1) and $N_2$ customers using multiple estimates (method 2). Since each method 2 customer requires 5/3 as much CPU time as a method 1 customer, we must have $N_2 = 3/5\, N_1$. The ratio of the lengths of the confidence intervals is then

**Table 6** Point estimates and 95% confidence intervals for 95% confidence interval coverages for E[*W*] in an M/M/1 queue with $\rho = 0.5$.

| *Estimator* | *R = 200 C = 1000* | *R = 100 C = 2000* | *R = 50 C = 4000* |
|---|---|---|---|
| $\hat{r}_0$ | 0.95 | 0.96 | 0.98 |
| | (0.91, 0.97) | (0.90, 0.98) | (0.90, 1.00) |
| $\hat{r}_1$ | 0.95 | 0.94 | 0.96 |
| | (0.91, 0.97) | (0.88, 0.97) | (0.87, 0.99) |
| $\hat{r}_2$ | 0.94 | 0.94 | 0.96 |
| | (0.90, 0.97) | (0.88, 0.97) | (0.87, 0.99) |
| $\hat{r}_{\hat{\beta}^*}(k = 1)$ | 0.87 | 0.88 | 0.94 |
| | (0.82, 0.91) | (0.80, 0.93) | (0.84, 0.98) |
| $\hat{r}_{\hat{\beta}^*}(k = 2)$ | 0.74 | 0.79 | 0.90 |
| | (0.68, 0.80) | (0.70, 0.86) | (0.79, 0.96) |

$$\frac{\sigma_0/N_1^{1/2}}{\sigma_2(\boldsymbol{\beta}^*)/N_2^{1/2}} = \frac{\sigma_0}{\sigma_2(\boldsymbol{\beta}^*)} \times (5/3)^{1/2}.$$

This ratio is 0.23 and 0.74 for $\rho = 0.5$ and 0.9, respectively.

• *Waiting time process in an M/M/2 queue*

The structure of this queue, which is described in Kiefer and Wolfowitz [22] and [23], is quite a bit more complicated than that for GI/G/1. Nevertheless, it is possible (though tedious) to perform the necessary calculations to apply the method, particularly if $k$ is small. For a two-server queue the state space $E = \{\mathbf{w} = (w_1, w_2) : 0 \leq w_i \leq w_2\}$, and in the M/M/2 case

$$f_1(\mathbf{w}) = w_1 + \left(\frac{1}{\lambda} - \frac{1}{\mu}\right) + \frac{\mu}{\lambda(\lambda+\mu)} e^{-\lambda w_1} + e^{\mu(w_1-w_2)}\left(\frac{e^{-\lambda w_2}}{\lambda+\mu}\frac{1}{\mu}\right).$$

The M/M/2 queue was simulated for $\rho = 0.5$ and 0.90. On the basis of these runs we estimated $R_1^2$ to be 0.32 and 0.58 for $\rho = 0.5$ and 0.9, respectively (complete simulation results are given in [17]).

Notice that to evaluate the function $f_1$ three exponentials must be computed. Since this must be done for each customer, the CPU time required for the simulation is substantially increased. In fact we estimate that each customer requires 2.25 times as much CPU time as in straightforward simulation. Since, for $\rho = 0.9$, $R_1^2$ is greater than 0.5 it must be concluded that in this case the method is not computationally efficient. By this we mean that for a fixed amount of CPU time more accurate estimates can be obtained by not using the variance reduction technique. For this reason the method is not recommended for the heavily loaded GI/G/$c$ queue ($c > 1$) unless the value of $k$ can be increased and the functions $f_1, \cdots, f_k$ can be evaluated cheaply.

## 5. Conclusions

In this paper a variance reduction technique for a wide class of stochastic processes has been proposed. The method differs from most other control variable methods in that the means of the control variables do not need to be known explicitly. The method is capable of producing substantial variance reductions. Because the method requires additional computations to be done both before and during the simulation, care must be taken so that the method is used only when it is computationally advantageous to do so; that is, it should only be used when for a fixed amount of computer time more accurate estimates can be obtained by using the method than by not using it. In the case of Markov chains it is likely that the method will be most effective when the transition matrix of the process is sparse, and specially structured, in which case the preliminary calculations can be carried out with relative ease. For example, a direct numerical analysis of closed Markovian queueing networks which do not satisfy the conditions necessary for a product form stationary distribution is often infeasible since the number of states may be enormous. Thus simulation (or sometimes approximation) is the only feasible alternative. However, for these queueing systems the elements of the transition matrix are often very simple functions of only a few parameters. Thus the transition matrix need never be stored and the functions $f_v(i)$ may be easily evaluated whenever the simulation enters state $i$. In such a case the method could reasonably be expected to work well.


## Acknowledgments

I wish to thank Donald L. Iglehart for his many valuable suggestions throughout the course of this research. I would also like to thank Stephen S. Lavenberg for his comments and careful reading of the manuscript. Most of this work was done at Stanford University and was supported by National Science Foundation Grant MCS-23607 and Office of Naval Research Contract N00014-76-0578.


## References


1. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multi-Server Queues," *J. ACM* **21,** 103-113 (1974).
2. D. L. Iglehart, "Functional Limit Theorems for the Queue GI/G/1 in Light Traffic," *Adv. Appl. Prob.* **3,** 269-281 (1971).
3. D. P. Gaver and G. L. Thompson, *Programming and Probability Models in Operations Research*, Brooks/Cole Publishing Co., Monterey, CA, 1973.
4. J. S. Carson, "Variance Reduction Techniques for Simulated Queuing Processes," Ph.D. Thesis, Department of Industrial Engineering, University of Wisconsin, Madison, WI, 1978.
5. D. P. Gaver and G. S. Shedler, "Control Variable Methods in the Simulation of a Model of a Multiprogrammed Computer System," *Naval Res. Logist. Quart.* **18,** 435-450 (1971).
6. D. L. Iglehart and P. A. W. Lewis, "Regenerative Simulation with Internal Controls," *J. ACM* **26,** 271-282 (1979).
7. S. S. Lavenberg, "Efficient Estimation of Work-Rates in Closed Queueing Networks," *Proceedings in Computational Statistics*, Physica Verlag, Vienna, 1974, pp. 353-362.
8. S. S. Lavenberg, T. L. Moeller, and C. H. Sauer, "Concomitant Control Variables Applied to the Regenerative Simulation of Queueing Systems," *Oper. Res.* **27,** 134-160 (1979).
9. S. S. Lavenberg, T. L. Moeller, and P. D. Welch, "Control Variables Applied to the Simulation of Queueing Models of Computer Systems," *Computer Performance*, North-Holland Publishing Co., Amsterdam, 1977, pp. 459-467.
10. S. S. Lavenberg, T. L. Moeller, and P. D. Welch, "Statistical Results on Multiple Control Variables with Application to Variance Reduction in Queueing Network Simulation," *Research Report RC7423*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1978.
11. S. S. Lavenberg and G. S. Shedler, "Derivation of Confidence Intervals for Work Rate Estimators in a Closed Queueing Network," *SIAM J. Comput.* **4,** 108-124 (1975).
12. S. S. Lavenberg and P. D. Welch, "A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations," *Research Report RC8161*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1980.
13. P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, Inc., New York, 1968.

14. K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd Ed., Springer-Verlag, Berlin, 1967.
15. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, II: Markov Chains," *J. ACM* **21,** 114-123 (1974).
16. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems, III: Regenerative Processes and Discrete-Event Simulations," *Oper. Res.* **23,** 33-45 (1975).
17. P. Heidelberger, "Variance Reduction Techniques for the Simulation of Markov Processes, I: Multiple Estimates," *Technical Report 42*, Department of Operations Research, Stanford University, Stanford, CA, 1977.
18. J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, Methuen and Co., Ltd., London, 1964.
19. A. Hordijk, D. L. Iglehart, and R. Schassberger, "Discrete Time Methods for Simulating Continuous Time Markov Chains," *Adv. Appl. Prob.* **8,** 772-788 (1976).
20. G. P. Learmonth and P. A. W. Lewis, "Naval Postgraduate School Random Number Generator Package LLRANDOM," *Report NP555LW73061A*, Naval Postgraduate School, Monterey, CA, 1973.
21. S. S. Lavenberg and C. H. Sauer, "Sequential Stopping Rules for the Regenerative Method of Simulation," *IBM J. Res. Develop.* **21,** 545-558 (1977).
22. J. Kiefer and J. Wolfowitz, "On the Theory of Queues with Many Servers," *Trans. Amer. Math. Soc.* **78,** 1-18 (1955).
23. J. Kiefer and J. Wolfowitz, "On the Characteristics of the General Queueing Process with Applications to Random Walks," *Ann. Math. Statist.* **27,** 147-161 (1956).

*Received March 5, 1980*

*The author is located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.*