COMPUTATIONAL AND STATISTICAL ISSUES IN DISCRETE-EVENT SIMULATION

by

Peter W. Glynn and Donald L. Iglehart

TECHNICAL REPORT No. 33

March 1989

Prepared under the Auspices
of
U.S. Army Research Contract
DAAL-88-K-0063
-03-

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

COMPUTATIONAL AND STATISTICAL ISSUES IN DISCRETE-EVENT SIMULATION

by

Peter W. Glynn

and

Donald L. Iglehart

Department of Operations Research
Stanford University
Stanford, CA 94305

Abstract

Discrete-event simulation is one of the most important techniques available for studying complex stochastic systems. In this paper we review the principal methods available for analyzing both the transient and steady-state simulation problems in sequential and parallel computing environments. Next we discuss several of the variance reduction methods designed to make simulations run more efficiently. Finally, a short discussion is given of the methods available to study system optimization using simulation.

Keywords: stochastic simulation, output analysis, variance reduction, parallel computation, and system optimization.

1. Introduction.

Computer simulation of complex stochastic systems is an important technique for evaluating system performance. The starting point for this method is to formulate the time varying behavior of the system as a basic stochastic process $\boldsymbol{Y} \equiv \{Y(t) : t \geq 0\}$, where $Y(\cdot)$ may be vector-valued. [Discrete time processes can also be handled.] Next a computer program is written to generate sample realizations of $\boldsymbol{Y}$. Simulation output is then obtained by running this program. Our discussion in this paper is centered on the analysis of this simulation output. The goal being to develop sound probabilistic and statistical methods for estimating system performance.

Two principal problems arise: the transient simulation problem and the steady-state simulation problem. Let T denote a stopping time and $X \equiv h\{Y(t) : 0 \leq t \leq T\}$, where h is a given real-valued function. The transient problem is to estimate $\alpha \equiv E\{X\}$. Examples of α include the following:

$$\begin{aligned} \alpha &= E\{f(Y(t_0)\}, \\ &= E\left\{\frac{1}{t_0}\int_0^{t_0} f(Y(s))ds\right\}, \end{aligned}$$

and

$$\alpha = P\{Y \text{ does not enter } A \text{ before } t_0\}.$$

Here t_0 is a fixed time (> 0), f is a given real-valued function, and A is a given subset of the state-space of $\boldsymbol{Y}$. The transient problem is relevant for systems running for a limited (but possibly random) length of time that cannot be expected to reach a steady-state. Our goal here is to provide both point and interval estimates for α.

For the steady-state problem we assume the $\boldsymbol{Y}$ process is asymptotically stationary in the sense that

$$\frac{1}{t}\int_0^t f(Y(s))ds \Rightarrow \alpha$$

as $t \to \infty$. Here $\Rightarrow$ denotes weak convergence and f is a given real-valued function defined on the state-space of $\boldsymbol{Y}$. The easiest example to think about here is an irreducible, positive recurrent, continuous time Markov chain. In this case $Y(t) \Rightarrow Y$ as $t \to \infty$ and $\alpha \equiv E\{f(Y)\}$. Examples of α in this case include the following:

$$\alpha = E\{Y^k\} \quad (\text{when } Y \text{ is real-valued}),$$

$$\alpha = P\{Y \in A\},$$

and

$$\alpha = E\{c(Y)\},$$

where c is a given cost function. Again as in the transient case, we wish to construct both point and interval estimates for α.

2. Transient Problem.

Assume we have a computational budget of t time units with which to simulate the process $\boldsymbol{Y}$ and estimate $\alpha \equiv E\{X\}$, as defined in Section 1. In a sequential computing environment we would generate independent, identically distributed (iid) copies

$$(X_1, \tau_1), (X_2, \tau_2), \cdots,$$

where the X_i's are copies of X and τ_i is the computer time required to generate X_i. Let $N(t)$ denote the number of copies of X generated in time t; this is just the renewal process associated with the iid τ_i's. A natural point estimator for α is

$$\bar{X}_{N(t)} \equiv \begin{cases} \frac{1}{N(t)} \sum_{i=1}^{N(t)} X_i & , \ N(t) > 0 \\ 0 & , \ N(t) = 0. \end{cases}$$

The standard asymptotic results for $\bar{X}_{N(t)}$ are the strong law of large numbers (SLLN) and the central limit theorem (CLT).

STRONG LAW OF LARGE NUMBERS. If $E\{\tau_1\} < \infty$ and $E\{| X_1 |\} < \infty$, then $\bar{X}_{N(t)} \to \alpha$ *a.s.* as $t \to \infty$.

CENTRAL LIMIT THEOREM. If $E\{\tau_1\} < \infty$ and $\text{var}\{X_1\} < \infty$, then

$$t^{1/2}[\bar{X}_{N(t)} - \alpha] \Rightarrow (E\{\tau_1\} \cdot \text{var}\{X_1\})^{1/2} \cdot N(0,1),$$

where $N(0,1)$ is a mean zero, variance one normal random variable. The SLLN follows from the SLLN for iid summands and the SLLN for renewal processes. The CLT result can be found in BILLINGSLEY (1968), Section 17.

From the SLLN we see that $\bar{X}_{N(t)}$ is a strongly consistent point estimator for α. Thus for large t we would use $\bar{X}_{N(t)}$ as our point estimate. On the other hand, the CLT can be used in the standard manner to construct a confidence interval for α. Here the constant $E\{\tau_1\} \cdot var\{X_i\}$ appearing in the CLT would have to be estimated.

Suppose now that we are in a parallel computing environment with p independent processors. Now we wish to estimate α for a fixed t as $p \to \infty$. On the p processors we generate iid copies of (X, τ):

$$\begin{array}{llll} (X_{11}, \tau_{11}), (X_{12}, \tau_{12}) & , \cdots , & (X_{1N_1(t)}, \tau_{1N_1(t)}) \\ (X_{21}, \tau_{21}), (X_{22}, \tau_{22}) & , \cdots , & (X_{2N_2(t)}, \tau_{2N_2(t)}) \\ & \vdots & \\ (X_{p1}, \tau_{p1}), (X_{p2}, \tau_{p2}) & , \cdots , & (X_{pN_p(t)}, \tau_{p,N_p(t)}). \end{array}$$

A number of estimators have been proposed for estimating $\alpha = E\{X\}$. The most natural estimator to consider first is that obtained by averaging the realizations of X across each processor and then averaging these sample means. This leads to

$$\alpha_1(p,t) \equiv \frac{1}{p} \sum_{i=1}^{P} \bar{X}^{(i)}_{N_i(t)},$$

where

$$\bar{X}^{(i)}_{N_i(t)} \equiv \begin{cases} \frac{1}{N_i(t)} \sum_{j=1}^{N_i(t)} X_{ij} & , \ N_i(t) > 0 \\ 0 & , \ N_i(t) = 0. \end{cases}$$

Here the processing ends on all processors at time $T_p = t$. If $E\{\tau_1\} < \infty$ and $E\{\mid X \mid\} < \infty$, then for all $t > 0$

$$\alpha_1(p,t) \to E\{\bar{X}_{N(t)}\} = E\{X \cdot 1_{\{\tau \le t\}}\} \quad a.s.$$

as $p \to \infty$. Here 1_A is the indicator function of the set A. Unfortunately, $E\{X\} \neq E\{X \cdot 1_{\{\tau \le t\}}\}$ and so $\alpha_1(p,t)$ is not strongly consistent for α as $p \to \infty$.

The next estimator for α was proposed by HEIDELBERGER (1987). For this estimator we let all processors complete the replication in process at time t. The estimator is

$$\alpha_2(p,t) \equiv \frac{\sum_{i=1}^{p} \sum_{j=1}^{N_i(t)+1} X_{ij}}{\sum_{i=1}^{p} [N_i(t)+1]}.$$

Here all processors complete by time

$$T_p = \max_{1 \le 1 \le p} [\tau_{i1} + \tau_{i2} + \cdots + \tau_{iN_i(t)+1}].$$

Unfortunately, $T_p \to +\infty$ *a.s.* as $p \to \infty$. However, $\alpha_2(p,t)$ is strongly consistent for α. To see this, note that if $E\{|X|\} < \infty$ and $P\{\tau > 0\} > 0$, then as $p \to \infty$

$$\alpha_2(p,t) \to \frac{E\left\{\sum_{j=1}^{N_i(t)+1} X_{ij}\right\}}{E\{N_i(t)+1\}} = E\{X\} \qquad a.s.$$

The equality about is simply Wald's equation. Finally, since $\alpha_2(p,t)$ is a ratio estimator, a CLT is also available from which a confidence interval can be constructed.

The last estimator we consider was proposed by HEIDELBERGER and GLYNN (1987). Here we set

$$\alpha_3(p,t) = \frac{1}{p}\sum_{i=1}^{p} \tilde{X}^{(i)}_{N_i(t)},$$

where

$$\tilde{X}^{(i)}_{N_i(t)} = \bar{X}^{(i)}_{N_i(t)} + X_{i1}1_{\{\tau_{i1} > t\}}.$$

Given $N(t) \ge 1$, Heidelberger and Glynn show that the pairs of random variables (X_1, τ_1), $\ldots, (X_{N(t)}, \tau_{N(t)})$ are exchangeable. Using this fact, they prove that $E\{\tilde{X}^{(i)}_{N_i(t)}\} = E\{X_1\}$. Since the $\tilde{X}^{(i)}_{N_i(t)}$'s are iid, we see that $\alpha_3(t)$ is strongly consistent for $\alpha = E\{X_1\}$. Since the summands in $\alpha_3(p,t)$ are iid, the standard CLT holds (under appropriate variance assumptions) and can be used to develop a confidence interval for α. Note that the definition of $\tilde{X}^{(i)}_{N_i(t)}$ requires the ith processor to complete the replication in process at time t, if no observations have been completed by time t; i.e., $\tau_{i1} > t$. Thus the completion time for all p processors is given by

$$T_p = \max_{1 \le i \le p} \{\max(t, \tau_{i1})\}.$$

While $T_p \to \infty$ *a.s.* as $p \to \infty$ (if $P\{\tau_{i1} > t\} > 0$), T_p goes to infinity at a much slower rate than is the case for $\alpha_2(p,t)$. They also show that the following CLT holds:

$$t^{1/2}[\tilde{X}_{N(t)} - \alpha] \Rightarrow \sigma E^{1/2}\{\tau_1\} \cdot N(0,1)$$

as $t \to \infty$. where we assume $0 < \sigma^2 = var\{X_1\} < \infty$ and $0 < E\{\tau_1\} < \infty$. Thus $\bar{X}_{N(t)}$ can also be used in a sequential environment to estimate α.

3. Steady-State Problem.

The steady-state estimation problem is considerably more difficult than the transient estimation problem. This difficulty stems from the following considerations: (i) need to estimate long-run system behavior from a finite length simulation run; (ii) an initial bias (or transient) usually is present since the process being simulated is non-stationary; and (iii) strong autocorrelations are usually present in the process being simulated. While classical statistical methods can often be used for the transient estimation problem, these methods generally fail for the steady-state estimation problem for the reasons mentioned above.

Assume our simulation output process is $\boldsymbol{Y} \equiv \{Y(t) : t \geq 0\}$ and for a given real-valued function f

$$\alpha(t) \equiv \frac{1}{t}\int_0^t f[Y(s)]ds \Rightarrow \alpha. \tag{1}$$

As stated above, we wish to construct point and interval estimators for α. In addition to (1), many methods also assume that a positive constant σ exists such that the following CLT holds:

$$\sqrt{t}[\alpha(t) - \alpha] \Rightarrow \sigma \cdot N(0,1) \tag{2}$$

as $t \to \infty$. From (1) and (2) we can construct a point estimate and confidence interval for α provided we can estimate σ. Estimating σ is generally the hardest problem.

A variety of methods have been developed to address the steady-state estimation problem. In Figure 1 we have given a break-down of these methods. Most of the methods are single replicate methods, since multiple replicate methods tend to be inefficient because of the initial bias problem.

Here we only consider single replicate methods. These methods are of two types: those that consistently estimate σ and those in which σ is cancelled out.

For consistent estimation of σ, we need a process $\{s(t) : t \geq 0\}$ such that $s(t) \Rightarrow \sigma$.

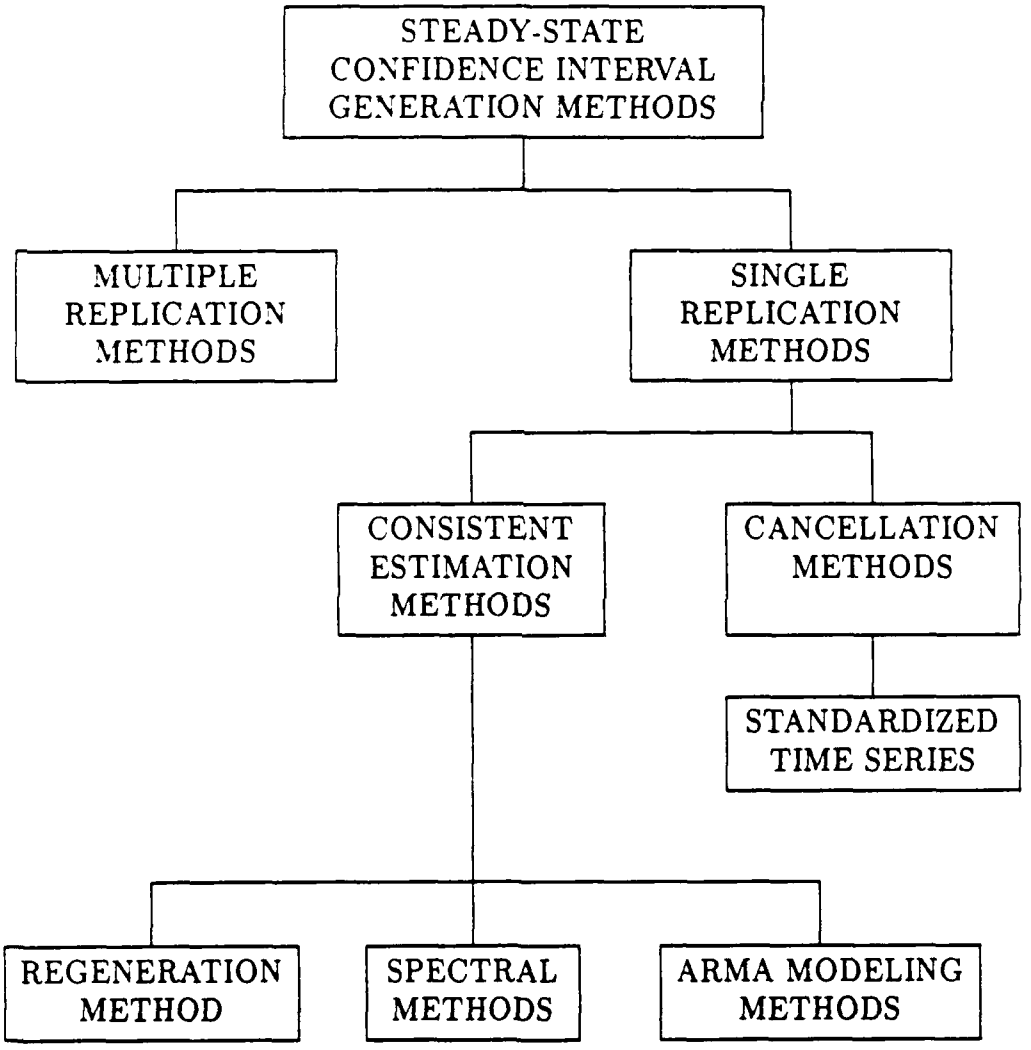
STEADY-STATE CONFIDENCE INTERVAL GENERATION METHODS
MULTIPLE REPLICATION METHODS
SINGLE REPLICATION METHODS
CONSISTENT ESTIMATION METHODS
CANCELLATION METHODS
STANDARDIZED TIME SERIES
REGENERATION METHOD
SPECTRAL METHODS
ARMA MODELING METHODS

Figure 1

In which case (2) leads to a $100(1-\delta)$ % confidence interval for α given by

$$[\alpha(t) - z(1-\delta/2)s(t)/t^{1/2}, \alpha(t) + z(1-\delta/2)s(t)/t^{1/2}],$$

where $\Phi(z(1-\delta/2)) = 1-\delta/2$ and Φ is the standard normal distribution function.

On the other hand, the canceling out methods require a non-vanishing process $\{Z(t): t \geq 0\}$ such that

$$[t^{1/2}(\alpha(t)-\alpha), Z(t)] \Rightarrow [\sigma N(0,1), \sigma Z]$$

as $t \to \infty$. Then using the continuous mapping theorem (cf., BILLINGSLEY (1968), p. 30) we have

$$t^{1/2}(\alpha(t)-\alpha)/Z(t) \Rightarrow N(0,1)/Z \tag{3}$$

as $t \to \infty$. Note from (3) that σ has been cancelled out in a manner reminiscent of the t-statistic.

First we discuss one of the methods in which σ is consistently estimated, namely, the regenerative method; see IGLEHART (1978) for a discussion of this method plus other background material. Here we assume that the simulation output process $\mathbf{Y}$ is a regenerative process. We are given a real-valued function f and wish to estimate $\alpha(f) \equiv E\{f(Y)\}$, where $Y(t) \Rightarrow Y$ as $t \to \infty$. Again it is convenient to think of $\mathbf{Y}$ as an irreducible, positive recurrent, continuous time Markov chain. Let $T(0) = 0, T_1, T_2, \ldots$ be the regeneration times for $\mathbf{Y}$ and set $\tau_i = T_i - T_{i-1}, i \geq 1$. The τ_i's are the lengths of the regenerative cycles. Next define the areas under the $\mathbf{Y}$ process in the kth regenerative cycle by

$$Y_k(f) = \int_{\tau_{k-1}}^{\tau_k} f[Y(s)]ds.$$

The following basic facts provide the foundation for the regenerative method:

(i) the pairs $\{(Y_k(f), \tau_k): k \geq 1\}$ are iid;

(ii) if $E\{|\, f(Y)\,|\} < \infty$, then $\alpha(f) = E\{Y_1(f)\}/E\{\tau_1\}$.

The regenerative method can be developed on either the intrinsic time scale (t) or on the random time scale (n) corresponding to the number of regenerative cycles simulated. On the intrinsic time scale our point estimate for α is given by

$$\alpha(t,f) \equiv \frac{1}{t}\int_0^t f(Y(s))ds,$$

where t is the length of time the simulation is run. On the random time scale our point is given by

$$\alpha_n(f) \equiv \bar{Y}_n(f)/\bar{\tau}_n,$$

where $\bar{Y}_n(f)$ (respectively, $\bar{\tau}_n$) is the sample mean of $Y_1(f), \ldots, Y_n(f)$ $(\tau_1, \ldots, \tau_n)$. Here the $\boldsymbol{Y}$ process is simulated to the completion of n regenerative cycles. Using the basic facts (i) and (ii) above, it can be shown that both $\alpha(t, f)$ and $\alpha_n(f)$ are strongly consistent for $\alpha(f)$ as t and n respectively tend to infinity. Next we define $Z_k \equiv Y_k(f) - \alpha(f)\tau_k$ and assume that $var\{Z_k\} \equiv \sigma^2 < \infty$. Then it can be shown that the following two CLT's hold as $t \to \infty$ and $n \to \infty$:

$$t^{1/2}[\alpha(t, f) - \alpha(f)] \Rightarrow (\sigma/E^{1/2}\{\tau_1\})N(0, 1),$$

and

$$n^{1/2}[\alpha_n(f) - \alpha(f)] \Rightarrow (\sigma/E\{\tau_1\})N(0, 1).$$

These two CLT's can then be used to construct confidence intervals for $\alpha(f)$ provided both σ^2 and $E\{\tau_1\}$ can be estimated. The mean $E\{\tau_1\}$ is easily estimated by $\bar{\tau}_n$ and σ^2 can be estimated from its definition in terms of $Y_1(f)$ and τ_1.

Next we turn to a discussion of the principal method available for canceling out σ. This is the method of standardized time series developed by SCHRUBEN (1983). Our discussion is based on the paper GLYNN and IGLEHART (1989) and uses some results from weak convergence theory; see BILLINGSLEY (1968) for background on this theory. From our output process $\boldsymbol{Y}$ we form the random elements of $C[0, 1]$, the space of real-valued continuous functions on the interval $[0, 1]$, given by

$$\bar{Y}_n(t) \equiv \frac{1}{n}\int_0^{nt} Y(s)ds$$

and

$$X_n(t) \equiv n^{1/2}[\bar{Y}_n(t) - \alpha t],$$

where $0 \leq t \leq 1$ and $n \geq 1$. Now we make the basic assumption that a finite, positive constant σ exists such that

$$X_n \Rightarrow \sigma B \text{ as } n \to \infty, \tag{4}$$

where B is standard Brownian motion. This assumption holds for a wide class of output processes. To find the scaling process $\{Z(t) : t \geq 0\}$ consider the class $\mathcal{M}$ of functions $g : C[0,1] \to \mathbb{R}$ such that

(i) $g(\alpha x) = \alpha g(x)$ for all $\alpha > 0$ and $x \in C[0,1]$;

(ii) $g(B) > 0$ with probability one;

(iii) $g(x + \beta k) = g(x)$ for all real β and $x \in C[0,1]$, where $k(t) = t$;

(iv) $P\{B \in D(g)\} = 0$, where $D(g)$ is the set of discontinuities of g.

The process

$$S_n(t) \equiv \frac{\bar{Y}_n(t) - \alpha t}{g(\bar{Y}_n)}, \quad 0 \leq t \leq 1,$$

is called a standardized time series. Using weak convergence arguments it is easy to show from (4) that

$$S_n(1) \Rightarrow B(1)/g(B) \tag{5}$$

as $n \to \infty$. Unfolding this CLT we have the following $100(1-\delta)\%$ confidence interval for α:

$$[\bar{Y}_n(1) - z(1-\delta/2)g(\bar{Y}_n), \bar{Y}_n(1) + z(\delta/2)g(\bar{Y}_n)],$$

where $P\{B(1)/g(B) \leq z(\alpha)\} = \alpha$ for $0 \leq \alpha \leq 1$. Thus each $g \in \mathcal{M}$ gives rise to a confidence interval for α provided we can find the distribution of $B(1)/g(B)$. Fortunately, this can be done for a number of interesting g functions.

One of the g functions leads to the batch means method, perhaps the most popular method for steady-state simulation. We conclude our discussion of the method of standardized time series by displaying this special g function. To this end we first define the Brownian bridge mapping $\Gamma : C[0,1] \to C[0,1]$ as

$$(\Gamma x)(t) = x(t) - tx(1), \quad x \in C[0,1], \quad 0 \leq t \leq 1$$

Now think of partitioning our original output process $\boldsymbol{Y}$ into $m \geq 2$ intervals of equal length and define the mapping $b_m : C[0,1] \to \mathbb{R}$ by

$$b_m(x) = \left[\left(\frac{m}{m-1} \right) \sum_{i=1}^{m} (x(i/m) - x((i-1)/m)^2 \right]^{1/2},$$

for $x \in C[0,1]$. Finally, the g function of interest is $g_m = b_m \circ \Gamma$. To see that g_m corresponds to the batch means method we observe that

$$g_m(\bar{Y}_n) = m^{-1/2}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(Z_i(n) - \frac{1}{m}\sum_{j=1}^{m} Z_j(n)\right)^2\right]^{1/2},$$

where

$$Z_i(n) = \int_{(i-1)n/m}^{in/m} Y(x)dx/(n/m)$$

is the ith batch mean of the process $\{Y(t) : 0 \le t \le n\}$. Specializing (5) to the function g_m we see that

$$[\frac{1}{m}\sum_{i=1}^{m} Z_i(n) - \alpha]/g_m(\bar{Y}_n) \Rightarrow t_{m-1}$$

as $n \to \infty$, where t_{m-1} is a Student's-t random variable with $m-1$ degrees of freedom. This follows from the fact that $B(1)/g_m(B)$ is distributed as t_{m-1} since B has independent normal increments. For other examples of functions $g \in \mathcal{M}$ for which the distribution of $B(1)/g(B)$ is known see GLYNN and IGLEHART (1989).

4. Variance Reduction Techniques.

Once a basic method is developed to produce point estimates and confidence intervals for a parameter of interest, we turn our attention to making these methods more efficient. Over the years a dozen or more techniques have been proposed to improve simulation efficiency. Good references for many of these techniques are BRATLEY, FOX, and SCHRAGE (1987), WILSON (1984). Here we have elected to outline three of these techniques.

As we have seen in Sections 2 and 3, confidence intervals for parameters being estimated are generally constructed from an associated CLT. Each CLT has an intrinsic variance constant, say, σ_1^2. The idea for many variance reduction techniques (VRT's) is to modify the original estimate in such a way as to yield a new CLT with a variance constant $\sigma_2^2 < \sigma_1^2$. This will, of course, lead to confidence intervals of shorter length, or alternatively, confidence intervals of the same length from a shorter simulation run. Frequently

VRT's are based on some analytic knowledge or structural properties of the process being simulated.

The first VRT we discuss is known as importance sampling. This idea was first developed in conjunction with the estimation of $E\{h(X)\} \equiv \alpha$, where h is a known real-valued function and X a random variable with density, say, f. Instead of sampling X from f, we sample X from a density g which has been selected to be large in the regions that are "most important", namely, where $|f|$ is largest. Then we estimate α by the sample mean of $h(X)f(X)/g(X)$; see HAMMERSLEY and HANDSCOMB (1964).

This same basic idea can be carried forward to the estimation of parameters associated with stochastic processes. We generate the process with a new probabilistic structure and estimate a modified parameter to produce an estimate of the original quantity of interest. The example we consider here is the $M/M/1$ queue with arrival rate λ, service rate μ, and traffic intensity $\rho \equiv \lambda/\mu < 1$. Let V denote the stationary virtual waiting time and consider estimating the quantity $\alpha \equiv P\{V > u\}$ for large u. When ρ is less than one, the virtual waiting time process has a negative drift and an impenetrable barrier at zero. Thus the chance of the process getting above a large u is small, and a long simulation would be required to accurately estimate α. The idea used here in importance sampling is to generate a so-called conjugate process obtained by reversing the roles of λ and μ. For the conjugate process the traffic intensity is greater than one, and the estimation problem becomes much easier. ASMUSSEN (1985) reports efficiency increases on the order of a factor of 3 to a factor of 400 over straight regenerative simulation depending on the values of ρ and u. In general, importance sampling can yield very significant variance reductions. Further work along these lines can be found in SIEGMUND (1976), GLYNN and IGLEHART (1989), SHAHABUDDIN et al. (1988), and WALRAND (1987).

The second VRT we discuss is known as indirect estimation. Assume we are interested in estimating $\alpha \equiv E\{X\}$, but happen to know that $E\{Y\} = aE\{X\} + b$ where a and b are known. Sometimes it happens that a CLT associated with the estimation of $E\{Y\}$ will have a smaller variance constant associated with it than does the CLT for estimating $E\{X\}$. In this case we would prefer to estimate $E\{Y\}$ and we use the affine transformation above to yield an estimate for $E\{X\}$. This idea has proved to be useful in queuing simulations where

the affine transformation is a result of Little's Law. In general, variance reductions realized using this method are not dramatic, being usually less than a factor of 2. For further results along these lines, see LAW (1975) and GLYNN and WHITT (1986). While the affine transformation works in queuing theory, it is conceivable that other transformations might arise in different contexts.

The third and final VRT we discuss here is known as discrete time conversion. Suppose that $\boldsymbol{X} = \{X(t) : t \geq 0\}$ is an irreducible, positive recurrent, continuous time Markov chain (CTMC). Then $X(t) \Rightarrow X$ as $t \to \infty$ and we may be interested in estimating $\alpha \equiv E\{f(X)\}$, where f is a given real-valued function. As we have discussed above, the regenerative method can be used to estimate α. A CTMC has two sources of randomness: the embedded discrete time jump chain and the exponential holding times in the successive states visited. The discrete time conversion method eliminates the randomness due to the holding times by replacing them by their expected values. It has been shown that this leads to a variance reduction when estimating α. Also, as an added side benefit computer time is saved since the exponential holding times no longer need to be generated. Gains in efficiency for this method can be substantial. Further discussion of this idea can be found in HORDIJK, IGLEHART, and SCHASSBERGER (1976), and FOX and GLYNN (1986).

5. System Optimization Using Simulation.

Consider a family of stochastic systems indexed by a parameter θ (perhaps vector-valued). Suppose $\alpha(\theta)$ is our performance criterion for system θ. Our concern here is to find that system, say θ_0, which optimizes the value of α. For a complex system it is frequently impossible to evaluate α analytically. Simulation may be the most attractive alternative. We could naively simulate the systems at a sequence of parameter settings $\theta_1, \theta_2, \cdots, \theta_k$ and select setting that optimizes $\alpha(\theta_i)$. In general this would not be very efficient, since k would have to be quite large. A better way would be to estimate the gradient of α and use this estimate to establish a search direction. Then stochastic approximation and ideas from non-linear programming could be used to optimize α.

Two general methods have been proposed to estimate gradients: the likelihood ratio method and the infinitesimal perturbation method. We will discuss both methods briefly.

Suppose $\boldsymbol{X} = \{X_n : n \geq 0\}$ is a discrete time Markov chain (DTMC) and that the cost of running system θ for $n+1$ steps is $g(\theta, X_0, \ldots, X_n)$. The expected cost of running system θ is then given by

$$\alpha(\theta) \equiv E_\theta\{g(\theta, X_0, \ldots, X_n)\}, \tag{6}$$

where E_θ is expectation relative to the probability measure $P(\theta)$ associated with system θ. If $E_\theta\{\cdot\}$ were independent of θ, we would simply simulate iid replicates of $\nabla g(\theta, X_0, \ldots, X_n)$. By introducing the likelihood function $\mathrm{L}(\theta, X_0, \ldots, X_n)$ it is possible to write $\alpha(\theta)$ as

$$\alpha(\theta) = E_{\theta_0}\{g(\theta, X_0, \cdots, X_n)L(\theta, X_0, \cdots, X_n)\}$$

for a fixed value of θ_0. Then we can write

$$\nabla\alpha(\theta) = E_{\theta_0}\{\nabla g(\theta, X_0, \cdots, X_n)L(\theta, X_0, \cdots, X_n)\},$$

where the interchange of ∇ and E_{θ_0} must be justified. A similar approach can be developed to estimate the gradient of a performance criterion for a steady-state simulation. For an overview of this approach see GLYNN (1987), and REIMAN and WEISS (1986).

The second method which has been proposed for estimating gradients is called the infinitesimal perturbation analysis (IPA) method. In this method a derivative, with respect to an input parameter, of a simulation sample path is computed. For example, we might be interested in estimating the mean stationary waiting time for a queueing system as well as its derivative with respect to the mean service time. Since we are taking a derivative of the sample path inside an expectation operator, the interchange of expectation and differentiation must be justified in order to produce an estimate for the gradient $\nabla\alpha(\theta)$, say. The IPA method assumes that if the change in the input parameter, θ, is small enough, then the times at which events occur get shifted slightly, but their order does not change. It has been shown that the IPA method yields strongly consistent estimates for the performance gradient in a variety of queueing contexts; see HEIDELBERGER, CAO, ZAZANIS, and SURI (1988) for details on the IPA method and a listing of queueing problems for which the technique works.

REFERENCES

ASMUSSEN, S. (1985). Conjugate processes and the simulation of ruin-problems. **Stoch. Proc. Appl.** 20, 213–229.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons, New York.

BRATLEY, P., FOX, B., and SCHRAGE, L. (1987). *A Guide to Simulation*. 2nd Ed. Springer-Verlag, New York.

FOX, B. and GLYNN. p. (1986). Discrete-time conversion for simulating semi-Markov processes. **Operations Research Letters** 5, 191–196.

GLYNN, P. and WHITT, W. (1989). Indirect estimation via $L = \lambda W$. **Operations Research** 37, 82–103.

GLYNN, p. (1987). Likelihood ratio gradient estimation: an overview. **Proceedings of the 1987 Winter Simulation Conference**, 366–375.

GLYNN, P. and HEIDELBERGER, P. (1987). Bias properties of budget constrained Monte Carlo simulations, I: estimating a mean. Technical Report, Department of Operations Research, Stanford University.

GLYNN, P. and IGLEHART, D. (1989). Simulation output analysis using standardized time series. To appear in **Math. of Operations Res.**

GLYNN, P. and IGLEHART, D. (1989). Importance sampling for stochastic simulations. To appear in **Management Sci.**

HAMMERSLEY, J. and HANDSCOMB, D. (1964). *Monte Carlo Methods*. Methuen, London.

HEIDELBERGER, P. (1987). Discrete event simulations and parallel processing: statistical properties. IBM Reserach Report RC 12733. Yorktown Heights, New York.

HEIDELBERGER, P., CAO, X-R, ZAZANIS, M. and SURI, R. (1988). Convergence properties of infinitesimal perturbation analysis estimates. **Management Sci.** 34, 1281–1302.

HORDIJK, A. IGLEHART, D. and SCHASSBERGER, R. (1976). Discrete-time methods for simulating continuous time Markov chains. **Adv. Appl. Prob.** 8, 772–788.

IGLEHART, D. (1978). The Regenerative method for simulation analysis. In *Current*

Trends in Programming Methodology - Software Modeling. (K. M. Chandy and R. T. Yeh, editors). Prentice-Hall, Englewood Cliffs, NJ, 52–71.

LAW, A. (1975). Efficient estimators for simulated queueing systems. **Management Sci.** 22, 30–41.

REIMAN, M. and WEISS, A. (1986). Sensitivity analysis via likelihood ratios. **Proceedings of the 1986 Winter Simulation Conference**, 285–289.

SHAHABUDDIN, P., NICOLA, V., HEIDELBERGER, P., GOYAL, A., and GLYNN, P. (1988). Variance reduction in mean time to failure simulations. **Proceedings of the 1988 Winter Simulation Conference**, 491–499.

SIEGMUND, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. **Ann. Statist.** 4, 673–684.

WALRAND, J. (1987). Quick simulation of rare event in queueing networks. **Proceedings of the Second International Workshop on Applied Mathematics and Performance/Reliability Models of Computer/Communication Systems**. G. Iazeolla, P. J. Courtois, and O. J. Boxina (eds.). North Holland Publishing Co., Amsterdam, 275–286.

WILSON, J. (1984). Variance reduction techniques for digital simulation. **Amer. J. Math. Management Sci.** 4, 277–312.

UNCLASSIFIED — MASTER COPY - FOR REPRODUCTION PURPOSES

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS:
2a SECURITY CLASSIFICATION AUTHORITY:	3 DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE:	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): Technical Report NO. 33	5 MONITORING ORGANIZATION REPORT NUMBER(S): ARO 25839.6-MA
6a NAME OF PERFORMING ORGANIZATION: Dept. of Operations Research; 6b OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: U. S. Army Research Office
6c ADDRESS (City, State, and ZIP Code): Stanford, CA 94305-4022	7b. ADDRESS (City, State, and ZIP Code): P. O. Box 12211 Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING / SPONSORING ORGANIZATION: U. S. Army Research Office; 8b. OFFICE SYMBOL (If applicable):	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DAAL03-88-K-0063
8c ADDRESS (City, State, and ZIP Code): P. O. Box 12211 Research Triangle Park, NC 27709-2211	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO; PROJECT NO; TASK NO; WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)
Computational and Statistical Issues in Discrete-Event Simulation

12 PERSONAL AUTHOR(S) Peter W. Glynn and Donald L. Iglehart

13a TYPE OF REPORT	13b TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Technical	FROM _____ TO _____	March 1989	16

16 SUPPLEMENTARY NOTATION The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17 COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	stochastic simulation, output analysis, variance reduction, parallel computation, and system optimization.

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Discrete-event simulation is one of the most important techniques available for studying complex stochastic systems. In this paper we review the principal methods available for analyzing both the transient and steady-state simulation problems in sequential and parallel computing environments. Next we discuss several of the variance reduction methods designed to make simulations run more efficiently. Finally, a short discussion is given of the methods available to study system optimization using simulation.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code) / 22c OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED