Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

VARIANCE REDUCTION IN SIMULATION

James R. Wilson
Mechanical Engineering Department
The University of Texas
Austin, TX 78712

ABSTRACT

In the design and analysis of simulation experiments, it is generally difficult to estimate model performance parameters with adequate precision at an acceptable computing cost. This paper surveys the main variance reduction techniques that have been developed to improve the efficiency of simulation-based performance statistics.

INTRODUCTION

From both a theoretical and practical standpoint, experimentation with a simulation model is frequently the only feasible means for performing systems analysis on a large-scale problem. However, the computing cost associated with direct simulation of a complex stochastic system can be a major drawback. In particular, excessive sample sizes may be required to yield acceptable precision in simulation-based estimators of relevant system parameters. This paper gives an up-to-date account of the variance redcution techniques that have been developed for or adapted to stochastic simulation.

We represent the basic simulation model under discussion as a response function $\phi(\cdot)$ whose input process $\{U_i : i \geqslant 1\}$ consists of independent random numbers. Following Hammersley and Handscomb [1], we reserve the term <u>random number</u> to refer to a variate that is uniformly distributed on the unit interval (0, 1). For simplicity we also assume that there is a finite upper bound m on the number of inputs sampled in one replication of the model; thus the response function $\phi(\cdot)$ has for its input the m×1 random vector $\underset{\sim}{U} = [U_1, \ldots, U_m]'$ that is uniformly distributed over the m-dimensional unit cube

$$I^m \equiv \prod_{j=1}^{m} (0, 1) \qquad (1)$$

with probability density

$$f_0(\underset{\sim}{u}) = \begin{cases} 1, & \underset{\sim}{u} \in I^m \\ 0, & \underset{\sim}{u} \in R^m - I^m \end{cases} . \qquad (2)$$

In terms of the random variable $Y = \phi(\underset{\sim}{U})$, the estimand of interest is

$$\theta = E(Y) = \int_{R^m} \phi(\underset{\sim}{u}) f_0(\underset{\sim}{u}) d\underset{\sim}{u} = \int_{I^m} \phi(\underset{\sim}{u}) d\underset{\sim}{u} \qquad (3)$$

The Variance Reduction Problem

<u>Direct simulation</u> simply computes the sample mean response $\bar{Y}_n$ over n independent replications of the basic model to yield an unbiased estimator of θ with $\mathrm{Var}(\bar{Y}_n) = \mathrm{Var}(Y)/n$. For a fixed sample size n, the problem is to apply an appropriate variance reduction technique (VRT) to the basic model in order to obtain an alternative estimator $\hat{\theta}_n$ with

$$E(\hat{\theta}_n) = \theta \quad \text{and} \quad \mathrm{Var}(\hat{\theta}_n) < \mathrm{Var}(\bar{Y}_n). \qquad (4)$$

Now different variance reduction techniques require different amounts of computing time to execute one replication of the simulation model; and some VRTs inherently require a random run length or a random replication count (for example, see the discussion below on importance sampling in the time domain). To take these phenomena into account, a more comprehensive formulation of the variance reduction problem is required.

In general let $\hat{\theta}$ denote an estimator for θ that is based on a VRT whose total computing time is $C(\hat{\theta})$. Relative to the direct simulation estimator $\bar{Y}_n$ with computing time $C(\bar{Y}_n)$, the <u>efficiency</u> of $\hat{\theta}$ is

$$\eta(\hat{\theta}:\bar{Y}_n) \equiv \mathrm{Var}(\bar{Y}_n)\cdot E[C(\bar{Y}_n)]/\{\mathrm{Var}(\hat{\theta})\cdot E[C(\hat{\theta})]\}. \qquad (5)$$

The general variance reduction problem is to construct $\hat{\theta}$ such that

$$E(\hat{\theta}) = \theta \quad \text{and} \quad \eta(\hat{\theta}:\bar{Y}_n) > 1. \qquad (6)$$

Of course the efficiency $\eta(\hat{\theta}:\bar{Y}_n)$ of $\hat{\theta}$ relative to direct simulation should be as large as possible.

Classification of Variance Reduction Techniques

Following Kohlas [2], we classify all variance reduction techniques into two major categories -- correlation methods and importance methods. For a fundamentally different taxonomy of VRTs, see Nelson and Schmeiser [3, 4, 5].

The <u>correlation methods</u> include three techniques that take advantage of linear correlation among simulation responses to yield efficiency increases. The techniques of <u>common random numbers</u> and <u>antithetic variates</u> respectively require the experimenter to induce positive and negative response correlations within blocks of simulation runs by forcing an appropriate functional dependence among the input vectors $\{\underset{\sim}{U}_j : 1 \leqslant j \leqslant n\}$ used on those runs. In contrast to this approach, the <u>control variates</u> technique uses regression methods to exploit any inherent correlation between the output Y and a selected concomitant random vector $\underset{\sim}{X}$ with known mean $\underset{\sim}{\mu}_X$ that is observed during each run.

The *importance methods* include four techniques that achieve improved efficiency by ultimately concentrating the sampling effort in those subregions of the input domain I^m that make the greatest contribution to the integral (3). Among the variants of the technique called *importance sampling*, *Russian roulette* and *splitting* are most easily adapted to discrete-event simulation because they are applied in the time domain. Whereas the control variates technique is effective when there is a strong linear association between the response Y and some auxiliary random vector $\underset{\sim}{X}$ with known mean, *stratified sampling* exploits prior knowledge of the distribution of $\underset{\sim}{X}$ to yield an efficiency increase when there is a complex nonlinear relationship between Y and $\underset{\sim}{X}$. As an alternative means of ensuring that the importance regions of I^m (or of the time domain) are adequately sampled, *systematic sampling* forces uniform sampling throughout the domain of interest by partitioning that domain into directly congruent strata so that a single point randomly sampled in one stratum automatically identifies a corresponding sample point in each stratum. Finally, the technique of *conditional Monte Carlo* achieves an efficiency increase by converting an estimation problem expressed as a conditional expectation (respectively, as an unconditional expectation) into another problem expressed in terms of an unconditional expectation (respectively, a conditional expectation).

CORRELATION METHODS

Induced Correlation Methods

If $Y_1 = \phi_1(\underset{\sim}{U}_1)$ and $Y_2 = \phi_2(\underset{\sim}{U}_2)$ are the responses of a pair of simulation experiments for which the difference $E(Y_1)-E(Y_2)$ is to be estimated, then the technique of common random numbers is used to reduce the variance of $Z = Y_1-Y_2$

$$\mathrm{Var}(Y_1-Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) - 2\mathrm{Cov}(Y_1,Y_2) \qquad (7)$$

by inducing $\mathrm{Cov}(Y_1,Y_2) \geqslant 0$. If instead $E(Y_1)$ is to be estimated and Y_1, Y_2 represent replicates, then the method of antithetic variates is used to reduce the variance of $Z = \frac{1}{2}(Y_1+Y_2)$

$$\begin{aligned}\mathrm{Var}[\tfrac{1}{2}(Y_1+Y_2)] &= \tfrac{1}{4}\mathrm{Var}(Y_1) + \tfrac{1}{4}\mathrm{Var}(Y_2) + \tfrac{1}{2}\mathrm{Cov}(Y_1,Y_2)\\ &= \tfrac{1}{2}\mathrm{Var}(Y_1) + \tfrac{1}{2}\mathrm{Cov}(Y_1,Y_2) \qquad (8)\end{aligned}$$

by inducing $\mathrm{Cov}(Y_1,Y_2) \leqslant 0$. In each case, the alternative direct approach would be simply to execute two independent runs in order to yield a single observation of Z. For both of the induced correlation techniques, the final estimator $\hat{\theta}_n$ is the average $\bar{Z}_{n/2}$ of a random sample of n/2 observations of the associated variate Z.

To induce $\mathrm{Cov}(Y_1,Y_2) \geqslant 0$, the technique of common random numbers consists of taking

$$\underset{\sim}{U}_2 = \underset{\sim}{U}_1. \qquad (9)$$

Similarly, to induce $\mathrm{Cov}(Y_1, Y_2) \leqslant C$, the simplest version of the method of antithetic variates uses the relation

$$\underset{\sim}{U}_2 = \underset{\sim}{1}_m-\underset{\sim}{U}_1, \text{ where } \underset{\sim}{1}_m \equiv [1, \ldots, 1]' \text{ is } m{\times}1. \qquad (10)$$

Although statements frequently appear in the literature to the effect that (9) and (10) are not guaranteed to produce the desired correlations in complex simulations, it should be noted that efficiency gains are ensured in the important special case that the response functions ϕ_1 and ϕ_2 are *concordant* for each of the random numbers constituting their input sequences. This means that with respect to each input coordinate, the functions ϕ_1 and ϕ_2 must be monotone in the same direction; however, both functions may be monotone nondecreasing in one coordinate and monotone nonincreasing in another coordinate. Bratley, Fox and Schrage [6] gave a definitive treatment of the methods of common random numbers and antithetic variates based on these considerations.

Control Variates

To construct a controlled estimator for $\theta = E(Y)$, we must identify a q-dimensional column vector of concomitant random variables $\underset{\sim}{X} = [X_1, \ldots, X_q]'$ having both a known expectation $\underset{\sim}{\mu}_X$ and a strong linear association with Y. In essence we try to predict and counteract the unknown deviation Y-θ by subtracting from Y an appropriate linear combination of the known deviations $\underset{\sim}{X}-\underset{\sim}{\mu}_X$:

$$Y(\underset{\sim}{b}) = Y - \underset{\sim}{b}(\underset{\sim}{X}-\underset{\sim}{\mu}_X). \qquad (11)$$

The controlled response $Y(\underset{\sim}{b})$ is unbiased for any fixed q-dimensional row vector $\underset{\sim}{b}$ of control coefficients. Let σ_Y^2 denote the variance of Y, let $\underset{\sim}{\Sigma}_X$ denote the covariance matrix of $\underset{\sim}{X}$ (assumed to be nonsingular), and let $\underset{\sim}{\sigma}_{YX}$ denote the row vector of covariances between Y and the components of $\underset{\sim}{X}$. The variance of the controlled response

$$\mathrm{Var}[Y(\underset{\sim}{b})] = \sigma_Y^2 - 2\underset{\sim}{\sigma}_{YX}\underset{\sim}{b}' + \underset{\sim}{b}\underset{\sim}{\Sigma}_X\underset{\sim}{b}' \qquad (12)$$

is minimized by the optimal control coefficient vector

$$\underset{\sim}{\beta} = \underset{\sim}{\sigma}_{YX}\underset{\sim}{\Sigma}_X^{-1}, \qquad (13)$$

which yields the minimum variance

$$\mathrm{Var}[Y(\underset{\sim}{\beta})] = \sigma_Y^2\cdot(1-\bar{R}^2_{Y\cdot X}) \qquad (14)$$

where $\bar{R}_{Y\cdot X}$ is the multiple correlation coefficient between Y and $\underset{\sim}{X}$.

In practice $\underset{\sim}{\sigma}_{YX}$ and $\underset{\sim}{\Sigma}_X$ are usually unknown and hence $\underset{\sim}{\beta}$ must be estimated. Let $\{(Y_j,\underset{\sim}{X}_j) : 1 \leqslant j \leqslant n\}$ denote the results observed on n independent replications of the simulation. In terms of the statistics

$$\bar{Y}_n = n^{-1}\sum_{j=1}^{n} Y_j, \quad S_Y^2 = (n-1)^{-1}\sum_{j=1}^{n}(Y_j-\bar{Y}_n)^2, \qquad (15)$$

$$\bar{\mathbf{X}}_n = n^{-1}\sum_{j=1}^{n}\mathbf{X}_j, \quad \mathbf{S}_X = (n-1)^{-1}\sum_{j=1}^{n}(\mathbf{X}_j-\bar{\mathbf{X}}_n)(\mathbf{X}_j-\bar{\mathbf{X}}_n)', \qquad (16)$$

and $$\mathbf{S}_{YX} = (n-1)^{-1}\sum_{j=1}^{n}(Y_j-\bar{Y}_n)(\mathbf{X}_j-\bar{\mathbf{X}}_n)', \qquad (17)$$

the sample analogue of (13) is

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{YX}\mathbf{S}_X^{-1}. \qquad (18)$$

Thus a point estimator of θ is

$$\hat{\theta} \equiv \bar{Y}(\hat{\boldsymbol{\beta}}) \equiv \bar{Y}_n - \hat{\boldsymbol{\beta}}(\bar{\mathbf{X}}_n-\boldsymbol{\mu}_X). \qquad (19)$$

Under the assumption that Y and $\mathbf{X}$ have a joint normal distribution

$$\begin{bmatrix} Y \\ \mathbf{X} \end{bmatrix} \sim N_{q+1}\left(\begin{bmatrix} \theta \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_{YX} \\ \boldsymbol{\sigma}'_{YX} & \boldsymbol{\Sigma}_X \end{bmatrix}\right), \qquad (20)$$

an exact 100(1-γ)% confidence interval for θ is given by

$$\bar{Y}(\hat{\boldsymbol{\beta}}) \pm t_{1-\gamma/2}(n-q-1)\cdot\hat{\sigma}_{Y\cdot X}\cdot\Delta, \quad \text{where} \qquad (21)$$

$$\hat{\sigma}^2_{Y\cdot X} \equiv (n-q-1)^{-1}\cdot(n-1)\cdot(S_Y^2 - \mathbf{S}_{YX}\mathbf{S}_X^{-1}\mathbf{S}'_{YX}), \qquad (22)$$

$$\Delta^2 \equiv n^{-1} + (n-1)^{-1}\cdot(\bar{\mathbf{X}}_n-\boldsymbol{\mu}_X)'\mathbf{S}_X^{-1}(\bar{\mathbf{X}}_n-\boldsymbol{\mu}_X), \qquad (23)$$

and $t_{1-\gamma/2}(n-q-1)$ denotes the $(1-\gamma/2)^{th}$ quantile of Student's t-distribution with n-q-1 degrees of freedom.

Now the use of $\hat{\boldsymbol{\beta}}$ rather than $\boldsymbol{\beta}$ means that the minimum variance (14) is not achieved. To measure the efficiency loss arising from estimation of the optimal control coefficients, Lavenberg, Moeller and Welch [7] derived the <u>loss factor</u>

$$\mathrm{Var}[\bar{Y}(\hat{\boldsymbol{\beta}})]/\mathrm{Var}[\bar{Y}(\boldsymbol{\beta})] = (n-2)/(n-q-2). \qquad (24)$$

Combining (14) and (24), we have

$$\mathrm{Var}[\bar{Y}(\hat{\boldsymbol{\beta}})] = \mathrm{Var}(\bar{Y}_n)\cdot\left[(1-\bar{R}^2_{Y\cdot X})\cdot\frac{n-2}{n-q-2}\right], \qquad (25)$$

from which it is clear that a variance <u>increase</u> can result from using too many control variates (q) relative to the replication count (n).

Rubinstein and Markus [8] carried out a development similar to (11) through (25) for the case of a multivariate response -- that is, for a p-dimensional output vector $\mathbf{Y} = [Y_1, \ldots, Y_p]'$. For a univariate response, Nozari, Arnold and Pegden [9] extended the preceding development in another direction so as to apply multiple controls to the estimation of a general linear model. Portanova and Wilson [10] further extended the work of Nozari, Arnold and Pegden to handle a multivariate response.

In all of the foregoing discussion of control variates, the output analysis was based on the method of independent replications. For the case of a univariate response, Lavenberg and Welch [11] also discussed the application of control variates in conjunction with the method of batch means and the regenerative method of simulation analysis.

Several types of control variates have recently been proposed for different classes of simulation models. Wilson and Pritsker [12, 13] developed a set of asymptotically stable "standardized service-time" controls for use in queueing network simulations. Grant and Solberg [14] and Venkatraman [15] devised effective controls for stochastic activity networks. In the context of nonlinear regression problems, Swain and Schmeiser [16] sought to characterize the sampling distribution of the nonlinear parameter estimators. Using as controls the linear approximators to the regression solution, Swain and Schmeiser were able to achieve substantial efficiency increases when estimating the moments of the nonlinear parameter estimators.

IMPORTANCE METHODS

Importance Sampling in the Input Domain I^m

This technique requires the input vector $\mathbf{U}$ to be sampled from an alternative density f(·) instead of the uniform density $f_0(\cdot)$. To compensate for this distortion of the input so as to achieve condition (4), the original response $Y = \phi(\mathbf{U})$ is replaced by the variate $Z = \phi(\mathbf{U})/f(\mathbf{U})$. The importance estimator $\hat{\theta}_n$ is then taken to be the sample mean $\bar{Z}_n$ computed over n independent replications of the new response Z. When the importance density f(·) closely mimics $\phi(\cdot)$, the ratio $\phi(\mathbf{U})/f(\mathbf{U})$ is nearly constant, and a substantial variance reduction is achieved. Sampling from the optimal importance density

$$f^*(\mathbf{u}) = |\phi(\mathbf{u})|\cdot f_0(\mathbf{u})/\int_{R^m}|\phi(\mathbf{w})|\cdot f_0(\mathbf{w})d\mathbf{w}, \quad \mathbf{u}\varepsilon R^m \qquad (26)$$

minimizes $\mathrm{Var}(\hat{\theta}_n)$; see Kleijnen [17]. This technique has been successfully applied to many distribution sampling experiments. In such situations, there is no notion of a stochastic process evolving over time; and thus the general behavior of the response function $\phi(\cdot)$ is relatively easy to explore. However, in complex discrete-event simulations exhibiting dynamic behavior over (simulated) time, it is almost impossible to arrange even a general similarity between the functions $\phi(\cdot)$ and f(·). It should be noted that this technique is not guaranteed to yield a variance redcution: with a poorly chosen importance density, large variance increases can occur (Bratley, Fox and Schrage [6]).

Importance Sampling in the Time Domain

Kioussis and Miller [18] successfully applied a variant of importance sampling to estimate the probability of system failure in a fault-tolerant computer system. In such systems, failure is an extremely <u>rare event</u>; typically its probability of occurrence is 0.0001 or less. Clearly an effective variance reduction technique is essential to the feasibility of simulation-based reliability analyses of this type. Although the basic idea of importance sampling is relevant to the estimation of rare-event probabilities, the complexity of the response function $\phi(\cdot)$ precludes the approach outlined in the preceding subsection. Using the techniques of Russian roulette and splitting, Kioussis and Miller gave a general formulation of importance sampling in the time domain for transient simulations.

Let $\{\underset{\sim}{M}(t) : 0 \leq t \leq t^{\dagger}\}$ be a stochastic process with state space S such that $\underset{\sim}{M}(t)$ denotes the status of the simulation model at time t, and $t^{\dagger}$ is the simulation stopping time. Let $\{B_i : 1 \leq i \leq b\}$ denote disjoint subsets of S that are "bad" in the following sense: if $\underset{\sim}{M}(t)$ hits a state in B_i at time t, a degradation in system status occurs; and the system failure event F is more likely to occur in the remaining time interval $(t, t^{\dagger}]$. We also have the disjoint subsets $\{G_i : 1 \leq i \leq g\}$ that are "good" in the opposite sense: if $\underset{\sim}{M}(t)$ hits a state in G_i at time t, then the event F is less likely to occur in the remaining time interval $(t, t^{\dagger}]$.

The Russian roulette technique is applied when the sample path of the process $\{\underset{\sim}{M}(t)\}$ hits a "good" subset G_i at time t: (a) The path is continued with probability $p_i(t)$, and it is terminated with probability $1-p_i(t)$ (so that the associated simulation outputs are discarded); (b) If the path is continued, then its weight is increased by the factor $1/p_i(t)$ [this means that the indicator function I_F for the event F is replaced by $I_F/p_i(t)$]. The splitting technique is applied when the sample path of $\{\underset{\sim}{M}(t)\}$ hits a "bad" subset B_i at time t: (a) The path is split into $s_i(t)$ separate paths, each of which continues in time independently of the others (conditional on the common history up to time t); (b) The weight of each path is decreased by the factor $1/s_i(t)$ [this means that I_F is replaced by $I_F/s_i(t)$]. Independent replications of this entire procedure can be used to construct point and interval estimators of the failure probability $\theta \equiv \Pr\{F\} = E(I_F)$. In terms of the efficiency measure (5), Kioussis and Miller reported results in the range $1.67 \leq \eta(\hat{\theta}:\bar{Y}_n) \leq 2.55$ for a fault-tolerant computer system with four parallel processors.

In the context of steady-state simulation, Hopmans and Kleijnen [19] applied time-domain importance sampling in conjunction with regenerative analysis to estimate the proportion θ of calls coming into a telephone exchange that encounter a busy signal and thus are blocked (lost). Hopmans and Kleijnen reported results in the range $0.859 \leq \eta(\hat{\theta}:\bar{Y}_n) \leq 1.06$; thus in some instances the net efficiency increase required by (6) was not achieved.

Stratified Sampling

Prestratification. In some instances the effective use of an auxiliary variate $\underset{\sim}{X} = \underset{\sim}{\beta}(\underset{\sim}{U})$ to estimate $\theta = E(Y)$ may require partitioning the space of $\underset{\sim}{X}$ into, say, L strata $\{S_h : 1 \leq h \leq L\}$ with known weights

$$\pi_h \equiv \Pr\{\underset{\sim}{X}\varepsilon S_h\} = \int_{\underset{\sim}{\beta}^{-1}(S_h)} f_0(\underset{\sim}{u})d\underset{\sim}{u}, \quad 1 \leq h \leq L. \tag{27}$$

If in stratum h we randomly sample n_h pairs $\{(Y_{hj},\underset{\sim}{X}_{hj}): 1 \leq j \leq n_h\}$ (where n_h is fixed in advance) and calculate the associated mean response

$$\bar{Y}_h = n_h^{-1} \sum_{j=1}^{n_h} Y_{hj} \quad \text{for} \quad 1 \leq h \leq L, \tag{28}$$

then the prestratified estimator of θ is given by

$$\hat{\theta}_n = \sum_{h=1}^{L} \pi_h \bar{Y}_h. \tag{29}$$

In effect we are forcing a random sample of n_h inputs $\{\underset{\sim}{U}_j\}$ to fall in the subregion $\underset{\sim}{\beta}^{-1}(S_h)$ of I^m. Let

$$\mu_{Yh} \equiv E(Y|\underset{\sim}{X}\varepsilon S_h) \quad \text{and} \quad \sigma^2_{Yh} \equiv E(Y^2|\underset{\sim}{X}\varepsilon S_h) - \mu^2_{Yh} \tag{30}$$

respectively denote the mean and variance of the response Y within the h^{th} stratum. With the proportional allocation

$$n_h^{\circ} \equiv n\cdot\pi_h, \quad 1 \leq h \leq L, \tag{31}$$

the variance of the corresponding prestratified estimator $\hat{\theta}_n^{\circ}$ is

$$\mathrm{Var}(\hat{\theta}_n^{\circ}) = \mathrm{Var}(\bar{Y}_n) - n^{-1} \sum_{h=1}^{L} \pi_h(\mu_{Yh}-\theta)^2. \tag{32}$$

See Cochran [20] for a comprehensive treatment of prestratification.

Poststratification. It is usually awkward to implement prestratified sampling in discrete-event simulation. By contrast, poststratification merely requires the experimenter to make n independent replications of his original simulation model in order to generate a random sample $\{(Y_j,\underset{\sim}{X}_j) : 1 \leq j \leq n\}$ of size n. After the jth replication, the observed auxiliary variate $\underset{\sim}{X}_j$ is used to classify the corresponding response Y_j into its appropriate stratum. Let $\{Y_{hj} : 1 \leq j \leq N_h\}$ denote the final subsample of random size N_h falling in stratum h, $1 \leq h \leq L$. Subject to the condition that all strata are nonempty (that is, $N_h > 0$ for all h), the poststratified estimator for θ is

$$\hat{\theta}_n = \sum_{h=1}^{L} \pi_h(N_h^{-1} \sum_{j=1}^{N_h} Y_{hj}). \tag{33}$$

$$\Longrightarrow E(\hat{\theta}_n) = \theta \quad \text{and} \tag{34}$$

$$\mathrm{Var}(\hat{\theta}_n) = n^{-1} \sum_{h=1}^{L} [\pi_h + (1-\pi_h)/n]\sigma^2_{Yh} + o(n^{-2}). \tag{35}$$

After some manipulation of (32) and (35), we see that the efficiency of poststratification is asymptotically equivalent to that of prestratification with proportional allocation. Wilson and Pritsker [12, 13] developed poststratified point and interval estimators for the regenerative method of simulation analysis as well as for the method of independent replications.

Systematic Sampling

Based on a partition of the input domain I^m into n directly congruent strata $\{S_j\}$, systematic sampling requires the selection of a single observation from each stratum. Although this technique appears to resemble stratified sampling, it does not involve a separate randomization within each stratum. Once the first sample point $\underset{\sim}{U}_1 \varepsilon S_1$ is determined, all of the other points $\{\underset{\sim}{U}_j \varepsilon S_j : 2 \leqslant j \leqslant n\}$ are chosen to occupy the same relative position within their respective strata. The systematic sampling estimator of θ is

$$\hat{\theta}_n = n^{-1} \sum_{j=1}^{n} \phi(\underset{\sim}{U}_j) \ . \tag{36}$$

This technique is attractive because it is both simpler and easier to apply than the other importance methods

In the simulation of a Markov chain with rewards, Fishman [21, 22] applied systematic sampling in the time domain. Observe that if ω is a fixed "rotation angle" in the unit interval [0, 1] and if U is a random number, then the translation modulo 1 of U by ω (that is, the rotation through the angle ω of the random point U on the circle with unit circumference)

$$U \oplus \omega \equiv \begin{cases} U+\omega & \text{if } U+\omega \leqslant 1 \\ U+\omega-1 & \text{if } U+\omega > 1 \end{cases} \tag{37}$$

yields a random number. To generate n parallel correlated replications of a given Markov chain, Fishman proposed that all of the state transitions starting from state s at time i should be simulated using rotations of a single independently sampled random number U_{si} by a set of regularly-spaced rotation angles. If K_{si} denotes the number of replications of the chain that reside in state s at time i and if $K_{si} > 0$, then the next transition for the jth such replication is sampled using the random number input

$$U_{sij} = U_{si} \oplus [(j-1)/K_{si}] \quad \text{for} \quad 1 \leqslant j \leqslant K_{si}. \tag{38}$$

For a broad class of infinite-state Markov chains, Fishman established the following properties of the resulting rotation sampling estimator $\hat{\theta}^{\circ}_n$ and its associated cost $C(\hat{\theta}^{\circ}_n)$:

$$\mathrm{Var}(\hat{\theta}^{\circ}_n) = O\{[\ln(n)]^4/n^2\}, \tag{39}$$

$$E[C(\hat{\theta}^{\circ}_n)] = O\{[\ln(n)]^2\}. \tag{40}$$

$$\Longrightarrow \quad 1/n(\hat{\theta}^{\circ}_n : \bar{Y}_n) = O\{[\ln(n)]^6/n^2\} = o(n^{-1}). \tag{41}$$

Relations (39) through (41) are remarkable results. Clearly systematic sampling merits further attention.

Conditional Monte Carlo

Unconditioning by a Change of Variable. The conditional Monte Carlo technique was originally developed to estimate the conditional expectation of the response $Y = \phi(\underset{\sim}{U})$ given a fixed value $\underset{\sim}{x}_0$ for some auxiliary random vector $\underset{\sim}{X} = \underset{\sim}{\beta}(\underset{\sim}{U})$. To reformulate the estimand

$$\theta = E(Y|\underset{\sim}{X}=\underset{\sim}{x}_0) = \int_{I^m} \phi(\underset{\sim}{u}) \cdot f_0(\underset{\sim}{u}|\underset{\sim}{\beta}(\underset{\sim}{u})=\underset{\sim}{x}_0)d\underset{\sim}{u} \tag{42}$$

as an unconditional expectation taken with respect to the original input density $f_0(\cdot)$, the key step of conditional Monte Carlo consists of finding suitable spaces A, B and a continuously differentiable map

$$\underset{\sim}{\omega} : \underset{\sim}{u} \ \varepsilon \ I^m \longrightarrow [\underset{\sim}{\alpha}(\underset{\sim}{u}), \underset{\sim}{\beta}(\underset{\sim}{u})] \ \varepsilon \ A \times B \tag{43}$$

with continuously differentiable inverse $\underset{\sim}{\tau} = \underset{\sim}{\omega}^{-1}$. Thus if we generate a random sample $\{\underset{\sim}{U}_j : 1 \leqslant j \leqslant n\}$ from f_0 and apply the input transformation

$$\underset{\sim}{T} : \underset{\sim}{u} \ \varepsilon \ I^m \longrightarrow \underset{\sim}{\tau}[\underset{\sim}{\alpha}(\underset{\sim}{u}), \underset{\sim}{x}_0] \ \varepsilon \ \underset{\sim}{\beta}^{-1}(\underset{\sim}{x}_0), \tag{44}$$

then the conditional Monte Carlo estimator of (42) is

$$\hat{\theta}_n = n^{-1} \sum_{j=1}^{n} \phi[\underset{\sim}{T}(\underset{\sim}{U}_j)] \cdot W(\underset{\sim}{U}_j), \tag{45}$$

where the weight function W(·) is chosen to satisfy (4). Let $J_\tau(\underset{\sim}{z}, \underset{\sim}{x})$ denote the Jacobian of $\underset{\sim}{\tau}$ evaluated at the point $(\underset{\sim}{z}, \underset{\sim}{x}) \ \varepsilon \ A \times B$. Moreover, let $f_Z(\underset{\sim}{z})$ and $F_X(\underset{\sim}{x})$ respectively denote the marginal densities of the random vectors $\underset{\sim}{Z} = \underset{\sim}{\alpha}(\underset{\sim}{U})$ and $\underset{\sim}{X} = \underset{\sim}{\beta}(\underset{\sim}{U})$ when $\underset{\sim}{U}$ is sampled from $f_0(\cdot)$. Granovsky [23] showed that the variance of $\hat{\theta}_n$ is minimized by the optimal weight function

$$W^*(\underset{\sim}{u}) = \frac{f_0\{\underset{\sim}{\tau}[\underset{\sim}{\alpha}(\underset{\sim}{u}), \underset{\sim}{x}_0]\} \cdot |J_\tau[\underset{\sim}{\alpha}(\underset{\sim}{u}), \underset{\sim}{x}_0]|}{f_Z[\underset{\sim}{\alpha}(\underset{\sim}{u})] \cdot f_X(\underset{\sim}{x}_0)} \ . \tag{46}$$

Swindles Based on Conditioning. In recent years the term conditional Monte Carlo has also been used to refer to a class of specialized techniques for which we obtain a more precise estimate of an unconditional expectation θ = E(Y) by conditioning on an appropriate auxiliary variate $\underset{\sim}{X} = \underset{\sim}{\beta}(\underset{\sim}{U})$ at some stage of the estimation procedure; see Bratley, Fox and Schrage [6]. The law of total probability ensures that the new response

$$Z \equiv E(Y|\underset{\sim}{X}) = E[Y|\underset{\sim}{\beta}(\underset{\sim}{U})] \tag{47}$$

is unbiased. Moreover, the variance decomposition

$$\mathrm{Var}(Y) = \mathrm{Var}[\ E(Y|\underset{\sim}{X})\ ] + E[\ \mathrm{Var}(Y|\underset{\sim}{X})\ ] \tag{48}$$

reveals that Var(Z) < Var(Y) unless Y has a strict functional dependence on $\underset{\sim}{X}$ (that is, unless $\mathrm{Var}(Y|\underset{\sim}{X}) = 0$ with probability one). Thus with n independent observations of the pair $(Y, \underset{\sim}{X})$, the corresponding conditional Monte Carlo estimator $\hat{\theta}_n = \bar{Z}_n$ is more precise than the direct simulation estimator $\bar{Y}_n$:

$$\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}(Z)/n < \mathrm{Var}(\bar{Y}_n). \qquad (49)$$

However, when the pairs $\{(Y_j, \mathbf{X}_j) : 1 \leq j \leq n\}$ are not independent, the left-hand equality in (49) breaks down and variance increases can occur unless some additional assumptions are made about the nature of the functional relationship $Z = Z(\mathbf{U})$.

REFERENCES

[1] Hammersley, J. M. and Handscomb, D. C., *Monte Carlo Methods*, Methuen & Co., Ltd., London, 1964.

[2] Kohlas, J., *Stochastic Methods of Operations Research*, Cambridge University Press, Cambridge, 1982.

[3] Nelson, B. L. and Schmeiser, B. W., "A Mathematical-Statistical Framework for Variance Reduction, Part I: Simulation Experiments," *Research Memorandum 84-4*, School of Industrial Engineering, Purdue University, 1984.

[4] Nelson, B. L. and Schmeiser, B. W., "A Mathematical-Statistical Framework for Variance Reduction, Part II: Classes of Transformations," *Research Memorandum 84-5*, School of Industrial Engineering, Purdue University, 1984.

[5] Nelson, B. L. and Schmeiser, B. W., "Decomposition of Some Well-Known Variance Reduction Techniques," *Research Memorandum 84-6*, School of Industrial Engineering, Purdue University, 1984.

[6] Bratley, P., Fox, B. L., and Schrage, L. E., *A Guide to Simulation*, Springer-Verlag, New York, 1983.

[7] Lavenberg, S. S., Moeller, T. L. and Welch, P. D., "Statistical Results on Control Variables With Application to Queueing Network Simulation," *Operations Research*, **30**, 182-202, 1982.

[8] Rubinstein, R. Y. and Markus, R., "Efficiency of Multivariate Control Variates in Monte Carlo Simulation," *Technical Report No. I 339*, National Research Institute for Mathematical Sciences, Pretoria, South Africa, 1981; to appear in *Operations Research*.

[9] Nozari, A., Arnold, S. F., and Pegden, C. D., "Control Variates for Multipopulation Simulation Experiments," *IIE Transactions*, **16**, 159-169, 1984.

[10] Portanova, A. and Wilson, J. R., "Using Multiple Control Variables to Estimate Multivariate Metamodels for Simulation Experiments," *Technical Report*, Mechanical Engineering Department, The University of Texas, Austin, Texas, 1984.

[11] Lavenberg, S. S. and Welch, P. D., "A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations," *Management Science*, **27**, 322-335, 1981.

[12] Wilson, J. R. and Pritsker, A. A. B., "Variance Reduction in Queueing Simulation Using Generalized Concomitant Variables," *Journal of Statistical Computation and Simulation*, **19**, 129-153, 1984.

[13] Wilson, J. R. and Pritsker, A. A. B., "Experimental Evaluation of Variance Reduction Techniques for Queueing Simulations Using Generalized Concomitant Variables," *Management Science* (to appear).

[14] Grant, F. H. and Solberg, J. J., "Variance Reduction Techniques in Stochastic Shortest Route Analysis: Applications, Procedures, and Results," *Mathematics and Computers in Simulation*, **XXV**, 366-375, 1983.

[15] Venkatraman, S., "Application of the Control Variate Technique to Multiple Simulation Output Analysis," *Technical Report*, Mechanical Engineering Department, The University of Texas, Austin, Texas, 1983.

[16] Swain, J. J. and Schmeiser, B. W., "Monte Carlo Estimation of the Sampling Distribution of Nonlinear Parameter Estimators," *Technical Report C-83-1*, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, 1983.

[17] Kleijnen, J. P. C., *Statistical Techniques in Simulation, Part I*, Marcel Dekker, New York, 1974.

[18] Kioussis, L. C. and Miller, D. R., "An Importance Sampling Scheme for Simulating the Degradation and Failure of Complex Systems During Finite Missions," *Proceedings of the 1983 Winter Simulation Conference*, Institute of Electrical and Electronic Engineers, Washington, D. C., 1983.

[19] Hopmans, A.C.M. and Kleijnen, J. P. C., "Importance Sampling in Systems Simulation: A Practical Failure?" *Mathematics and Computers in Simulation*, **XXI**, 209-220, 1979.

[20] Cochran, W. G., *Sampling Techniques*, Third Edition, John Wiley and Sons, New York, 1977.

[21] Fishman, G. S., "Accelerated Accuracy in the Simulation of Markov Chains," *Operations Research*, **31**, 466-487, 1983.

[22] Fishman, G. S., "Accelerated Convergence in the Simulation of Countably Infinite State Markov Chains," *Operations Research*, **31**, 1074-1089, 1983.

[23] Granovsky, B. L., "Optimal Formulae for the Conditional Monte Carlo," *SIAM Journal on Algebraic and Discrete Methods*, **2**, 289-294, 1981.