EFFICIENCY IMPROVEMENT AND VARIANCE REDUCTION

Pierre L'Ecuyer

Département d'IRO
Université de Montréal, C.P. 6128, Succ. A
Montréal, H3C 3J7, CANADA

ABSTRACT

We give an overview of the main techniques for improving the statistical efficiency of simulation estimators. Efficiency improvement is typically (but not always) achieved through variance reduction. We discuss methods such as common random numbers, antithetic variates, control variates, importance sampling, conditional Monte Carlo, stratified sampling, and some others, as well as the combination of certain of those methods. We also survey the recent literature on this topic.

1. INTRODUCTION

1.1. A Notion of Efficiency

Stochastic simulation is typically used to compute the value of a realization of a random variable X, taken as an estimator of some unknown quantity μ. Suppose that X is defined over some probability space $(\Omega, \mathcal{B}, P)$ and use E to denote a mathematical expectation. The *bias*, *variance*, and *mean square error* (MSE) of X are defined as

$$\begin{aligned} \beta &= E[X] - \mu; \\ \sigma^2 &= \mathrm{Var}(X) \ = \ E[(X - E[X])^2]; \\ \mathrm{MSE}[X] &= E[(X-\mu)^2] \ = \ \beta^2 + \sigma^2, \end{aligned}$$

respectively. We assume that the cost for computing X (e.g., cpu time) is also a random variable and we denote its mathematical expectation by $C(X)$. We define the *efficiency* of X by

$$\mathrm{Eff}(X) \ = \ \frac{1}{\mathrm{MSE}[X] \cdot C(X)}. \quad (1)$$

In this context, for two estimators X and Y, we say that X is *more efficient* than Y if $\mathrm{Eff}(X) < \mathrm{Eff}(Y)$. *Efficiency improvement* means finding another estimator Y which is more efficient than the currently used estimator X in the above sense. Often, both estimators are unbiased and are assumed to have roughly the same computational costs; then, improving the efficiency is equivalent to reducing the variance. For that reason, most textbooks and research papers talk about *variance reduction techniques* (VRTs). However, efficiency can sometimes be improved by *increasing* the variance; see Fishman and Kulkarni (1992) and Glynn and Whitt (1992) for examples.

This paper gives an overview of the main ideas and recent research developments on efficiency improvement, mainly through variance reduction. We give a long list of references, with pointers to the most recent or important (according to the judgement and knowledge of this author). The list is clearly not exhaustive and we make no attempt to trace back the historical developments and give the original references.

For the readers who want to go further, we would like to particularly recommend the nice survey papers of Glynn (1994a), Heidelberger (1993) and Wilson (1984). Good introductions on variance reduction can also be found in Bratley, Fox, and Schrage (1987), Hammersley and Handscomb (1964), and Law and Kelton (1991) (among others).

Remark 1 The efficiency criterion (1) is not the only possibility, but is often agreed upon, typically with the assumption of no bias. Without bias, one can generally sample twice as many independent copies of the estimator, thus cutting the variance in half but doubling the computational effort, so the efficiency is invariant with respect to the number of replications in this case. In the presence of bias, the latter no longer holds, but (1) implies that variance can be traded off for squared bias, and vice-versa, without essentially altering the statistical precision of the estimator.

1.2. Asymptotic Efficiency

Arguing that (1) is difficult to compute in practice, Glynn and Whitt (1992) propose to consider the efficiency of simulation estimators in the asymptotic sense, as the size of the computer budget increases to infinity. What we now describe is a much simplified version of their framework. Let $\bar{X}(t)$ be the estimator obtained with a budget t (here, we have $C(\bar{X}(t)) = t$). Typically (under a few technical conditions), there exists a constant γ and a random variable Z such that $t^\gamma(\bar{X}(t) - \mu) \Rightarrow Z$ (where $\Rightarrow$ denotes the convergence in distribution), and also $t^{2\gamma}\mathrm{MSE}[\bar{X}(t)] = \nu + o(1)$ for some constant ν, where $o(1) \to 0$ as $t \to \infty$. Then, the *asymptotically most efficient* estimator is the one with the largest value of γ and, in case of a tie, the one with the smallest value of ν. Often, $\gamma = 1/2$ and Z is a centered normal for all estimators of interest in a given class. In that case, those estimators are compared through their variance constants. Note that one often has $\gamma = 1/2$ even in the presence of bias. Examples where $\gamma \neq 1/2$ are discussed in Glynn and Whitt (1992) and Glynn (1994a). See also L'Ecuyer (1992), L'Ecuyer and Perron (1994) and L'Ecuyer and Yin (1994). Note that in the latter case, changing the number of replications may change the efficiency of an estimator.

As an illustration, suppose we want to estimate the total expected discounted cost over an infinite horizon, in a stochastic model with discounting, using a truncated-horizon estimator over horizon τ. For a computing budget t, we may perform $n = \lfloor t/\tau \rfloor$ runs of length τ. Then, the simulation cost per run typically increases linearly in τ, whereas the marginal decrease of the MSE (as a function of τ) damps out exponentially fast as $\tau \to \infty$. To maximize the asymptotic efficiency in that case, there is an optimal way of increasing τ as a function of t; that is, a tradeoff between the horizon length and the number of replications. See also Fox and Glynn (1989). Other important examples involve derivative estimators based on finite differences or on the likelihood ratio method, and stochastic approximation based on these methods.

1.3. Modifying an Estimator for Variance Reduction

To see how an estimator can be modified (in general), recall that X is a measurable function of the sample point ω, say $X = h(\omega)$, and that

$$\mathrm{MSE}(X) = \int_\Omega (h(\omega) - \mu)^2 dP(\omega).$$

Modifying the estimator means modifying the function h without altering the probability law P, or perhaps modifying P itself, or both. Nelson (1985, 1986, 1987a, 1987b) proposed a decomposition of the transformation h into several levels, say, $h(\omega) = T_3(T_2(T_1(\omega)))$. In his framework, the random variables (or vectors) T_1, T_2, and T_3 are called the inputs, the outputs, and the performance statistics, and are defined (directly) over probability spaces called the probability spaces of inputs, outputs, and performance statistics, respectively. Variance reduction techniques can then be classified according to the level(s) at which the transformation is modified. Nelson identified six mutually exclusive classes of elementary transformations (two classes at each level) and showed that any VRT is a composition of such elementary transformations. That framework was developed with the hope that (a) fundamentally new VRTs could be found by playing with those building blocks and that (b) this decomposition would facilitate the "automation" of variance reduction by enabling the construction of general software for that purpose. It appears that reaching these long term objectives is still far ahead.

The remainder of this paper is devoted to a discussion of several VRTs. The common random numbers (next section) are used for comparing two or more "related" systems, whereas the other methods can be used for estimating the performance measure of a single system. The methods discussed in the next four sections are *correlation-based*: correlation is induced and exploited between different random variables. Some of the others (like importance sampling or stratification) may be called *importance methods*: they improve the efficiency by concentrating the sampling effort in the most critical regions of the sample space.

2. COMMON RANDOM NUMBERS

The common random numbers (CRN) method is normally used when estimating the *difference* between the expected performance measures of two (or more) systems. It is perhaps the most widely used VRT method in practice. Suppose we want to estimate $\mu_1 - \mu_2$, where μ_1 and μ_2 are two unknown quantities, estimated by X_1 and X_2, respectively. Let $Z = X_1 - X_2$ and suppose that $E[Z] = \mu_1 - \mu_2$. The variance of Z is then

$$\mathrm{Var}[Z] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] - 2\mathrm{Cov}[X_1, X_2].$$

If X_1 and X_2 are generated independently, the covariance term disappears. But if we manage to induce a positive covariance between X_1 and X_2 without changing their individual distributions, then the variance (and MSE) of Z will be reduced. The standard way of inducing such a covariance is to use the same underlying uniform random numbers to drive the simulation for both X_1 and X_2, and to make sure that these random numbers are used at exactly the same place for both systems (the latter is called *synchronization*). If both systems react in a similar way to these uniforms, then that should work. The rationale is that with the same uniforms, the random noise (or "experimental conditions") will be the same for both systems; so the observed differences will be due only to the differences between the systems, and not to the fact that one has been more lucky than the other in picking its random numbers. As an analogy, using CRNs is like comparing two fertilizers by using each of them on the same piece of land, at the same time (this is impossible in real life, but simulation makes it possible). With the CRNs, independent replicates of Z can be obtained by simulation and a confidence interval for $\mu_1 - \mu_2$ computed as usual.

Example 1 Suppose we want to compare the FIFO service policy with another policy, in a single-server $GI/GI/1$ queue, with regards to the average waiting time. Here, X_1 and X_2 may represent the observed average waiting times under each of the two policies. If the interarrival and service time distributions are the same for both systems, then using CRNs with proper synchronization implies that both systems will see the same customers arriving at the same times and with the same service requirements (the service requirements may be permuted between the customers if they are generated when the service begins, but not if they are generated when the customer arrive). To facilitate the synchronization, one may use here two different random number generators: one for the interarrivals and one for the service times. In general, having several generators available, as well as software tools for resetting a generator to a previous state and for jumping ahead, is very handy for the application of CRN and other VRTs. A software package offering that is provided by L'Ecuyer and Côté (1991). Several other examples of CRN applications (with numerical illustrations) are given in Bratley, Fox, and Schrage (1987), Law and Kelton (1991), and the references therein.

CRNs do not always work: using the same uniforms does not guarantee that $\mathrm{Cov}(X_1, X_2) > 0$. In practice, the uniforms are transformed in very complicated ways and at several levels to produce the estimators, making that covariance extremely hard to evaluate a priori. A sufficient (but by no means necessary) condition for the covariance to be positive is that X_1 and X_2 are both monotone (both increasing or decreasing) with respect to any given underlying uniform (see Heidelberger and Iglehart (1979) and Theorem 5.1 of Bratley, Fox, and Schrage 1987). If the monotonicity condition is satisfied only for some of the uniforms, then one can use CRNs only for those, and independent random numbers for the other uniforms. However, the monotonicity conditions are not always easy to check. If the covariance turns out to be negative, then the variance of Z will actually be increased. In the best case, if the correlation is perfect, the variance is reduced to zero. In the worst case, the variance could be doubled compared with independent simulations.

A well-known heuristic for trying to keep the monotonicity is to generate all the nonuniform random variables in the model by inversion. If these nonuniform variables have different distributions for the two systems, inversion will ensure that they remain monotone. However, the further transformations applied to these variables for producing the estimates X_1 and X_2 may be non-monotone. For complex real-life models, to assess whether CRNs would work, one may make a pilot study: perform a number of replications *with* CRNs and check whether or not the (estimated) variance of Z is smaller than the sum of the variances of X_1 and X_2.

A situation where CRNs would work extremely well, even in the absence of monotonicity, is when a system is parameterized by some continuous parameter θ, reacts similarly to similar values of θ, and we want to compare the performance under two values of θ that are close to each other. More specifically, if the response is a "smooth" function of θ when the values of the underlying uniforms are fixed, and if X_j is the value of the response evaluated at θ_j, $j = 1, 2$, then the correlation between X_1 and X_2 will approach 1 as $|\theta_1 - \theta_2|$ approaches 0, so the CRNs will reduce the variance if θ_1 and θ_2 are close enough to each other. This (and related issues) is studied by Glasserman and Yao (1992). One important application of this property is the estimation of derivatives (or gradients) by finite differences. In that context, with independent random numbers, the variance of the derivative estimator increases to infinity as the size of the finite-difference interval shrinks to zero. But with CRNs and under appropriate smoothness conditions, the variance remains bounded; L'Ecuyer and Perron (1994) give formal proofs and numerical examples. This is important because making the size of the finite difference interval converge to zero is generally required to make the bias converge to zero.

CRNs are also effective for comparing multiple (more than 2) systems; however, the induced dependence makes the statistical analysis more difficult (e.g., for selecting the best system with high probability or for computing a simultaneous confidence region for all differences). There exist simple analysis methods, such as using the Bonferroni inequality to compute confidence intervals, but these are very conservative. Consider the specific problem of *multiple comparisons with the best* (MCB): perform simultaneous statistical inference on all the $\mu_j - \mu_{j_*}$ for $j \neq j_*$, where μ_j is the (unknown) performance measure of system j and j_* is the best system. For MCB, Yang and Nelson (1991) and Nelson and Hsu (1993) proposed linear regression models trying to "explain" the effect of CRNs on the output via control variates which are functions of the simulation inputs. Their analysis assumes that all of the dependence induced by the CRNs is explained by the control variates and that the residuals are iid normals. Such control variates that account well for the dependence are not always easy to select in practice. Nelson (1993) proposed another (more robust) approach, that can be used with or without the control variate model, and for which the control variates are not assumed to capture all of the dependence.

CRNs are not useful only for estimating a difference such as $\mu_1 - \mu_2$, they could be effective more generally for estimating a function of several means: $g(\mu_1, \ldots, \mu_d)$, where each μ_j is a mathematical expectation estimated by X_j. Inducing correlations between the X_j's by using CRNs may reduce the variance of the estimator $g(X_1, \ldots, X_d)$. A special case is when estimating a ratio of expectations: $g(\mu_1, \mu_2) = \mu_1/\mu_2$.

Besides the variance reduction, there are situations where the CRNs also make the computations less costly. The idea is that the random numbers need to be generated only once. When comparing similar related systems, the lower-level transformations (e.g., the generation of interarrival and service times in a queue) are sometimes exactly (or almost) the same for all systems of interest, and the systems differ only at a higher level. Then, a significant amount of computation may be common to all systems and could be performed only once. L'Ecuyer and Vázquez-Abad (1994) show how this idea could be exploited to efficiently estimate an entire function of a univariate continuous parameter.

3. ANTITHETIC VARIATES

The idea of antithetic variates (AV) resembles that of CRNs. Now, we want to estimate a *single* mathematical expectation μ, using a pair of unbiased estimators (X^1, X^2). The (unbiased) estimator of μ will be the average: $X = (X^1 + X^2)/2$, whose variance is:

$$\mathrm{Var}[X] = \frac{\mathrm{Var}[X^1] + \mathrm{Var}[X^2]}{4} + \frac{\mathrm{Cov}[X^1, X^2]}{2}.$$

Assume that $\mathrm{Var}[X^1] = \mathrm{Var}[X^2]$. If X^1 and X^2 are independent, then $\mathrm{Var}[X] = \mathrm{Var}[X^1]/2$. But if $\mathrm{Cov}[X^1, X^2] < 0$, then X has a smaller variance. A standard way (but not the only way) of inducing the negative correlation is to use a sequence of underlying iid uniforms $\omega_1 \equiv \{U_k,\ k \geq 1\}$ to drive the simulation for computing X^1, and use the *antithetic* sequence $1 - \omega_1 \equiv \{1 - U_k,\ k \geq 1\}$ to drive the simulation when computing X^2. The two estimators can then be written as $X_1 = h(\omega_1)$ and $X_2 = h(1 - \omega_1)$. The rationale is that disastrous events in the first simulation should be compensated by "antithetic" lucky events in the second one, thus reducing the variance of the average.

As with CRNs, that does not guarantee a negative covariance neither a variance reduction in general. A sufficient condition for a negative covariance is that h be monotone with respect to each underlying uniform (Bratley, Fox, and Schrage 1987; Avramidis and Wilson 1994). In fact, if h is monotone only with respect to a subset Ψ of its (uniform) arguments, then variance reduction is still guaranteed if we take AVs only for uniforms that are in Ψ and independent random numbers for the others. Proper synchronization is again important. The best possible situation occurs when the response is a linear function of all underlying uniforms: the variance is then reduced to zero. The worst case is when X^1 and X^2 are perfectly correlated: the AV method then doubles the variance.

More general versions of the AV method are anaylzed in Cheng (1982), Cheng (1984), Fishman and Wang (1983), Wilson (1983) and Wilson (1984).

4. LATIN HYPERCUBE SAMPLING

Avramidis and Wilson (1994) describe a negative correlation-induction framework that generalizes the AV method. In their framework, n dependent replications are performed, the ith replication using a sequence of iid uniforms denoted, say, by $\omega_i = \{U_{i,k},\ k \geq 1\}$. Negative correlation is induced across the components of the different ω_i's as follows: for each index k in some finite subset Ψ, the vector of random numbers $U^{(k)} = (U_{1,k}, \ldots, U_{n,k})$, which contains the kth random number of each replication, follows a multivariate distribution with the following properties: (a) each univariate marginal is $U(0,1)$ and (b) each bivariate marginal is negatively quadrant dependent (nqd). (A bivariate random vector (Y_1, Y_2) is called nqd if $P[Y_1 \leq y_1,\ Y_2 \leq y_2] \leq P[Y_1 \leq y_1] \cdot P[Y_2 \leq y_2]$ for all y_1 and y_2.) Variance reduction is again guaranteed if h is monotone with respect to each of the arguments that have been included in Ψ.

Special cases of that correlation-induction framework include AV and the latin hypercube sampling (LHS) method (Avramidis and Wilson 1994), which we now describe. Select a finite subset Ψ as above and for each $k \in \Psi$, generate a random permutation of the integers $\{1, \ldots, n\}$ (independently for the different indices k), and let $\pi_{i,k}$ denote the ith element of that permutation. Then, for each (i, k), generate $U_{i,k}$ uniformly over the interval $((\pi_{i,k}-1)/n,\ \pi_{i,k}/n)$. The other $U_{i,k}$'s, for $k \notin \Psi$, are generated independently from the $U(0,1)$ distribution. It is easily seen that each $\omega_i \equiv \{U_{i,k},\ k \geq 1\}$ is then a sequence of iid uniforms. On the other hand, for each $k \in \Psi$, the interval $(0,1)$ is partitioned into n equal pieces and across the n replications, the $U_{i,k}$'s form a stratified sample over $(0,1)$.

5. CONTROL VARIABLES

The control variates (CV) method exploits auxiliary information to figure out whether the random events have been more favorable or less favorable than usual in influencing the sample performance, and makes appropriate corrections. Let X be the default performance estimator and $Y = (Y^{(1)}, \ldots, Y^{(q)})'$ (the prime means "transpose") be a vector of q other random variables, presumably correlated with X, with known expectation $E[Y] = \nu = (\nu^{(1)}, \ldots, \nu^{(q)})'$, and called the CVs. Define the controlled estimator

$$X_c = X - \beta'(Y - \nu) = X - \sum_{k=1}^{q} \beta_k (Y^{(k)} - \nu^{(k)}),$$

where $\beta = (\beta_1, \ldots, \beta_q)'$ is a vector of constants. Let $\Sigma_Y = \mathrm{Cov}[Y]$, a matrix whose element (i, j) is the value of $\mathrm{Cov}[Y^{(i)}, Y^{(j)}]$, and $\sigma_{XY} = (\mathrm{Cov}(X, Y^{(1)}), \ldots, \mathrm{Cov}(X, Y^{(q)}))'$. Then, $E[X_c] = E[X] = \mu$ and

$$\mathrm{Var}[X_c] = \mathrm{Var}[X] + \beta' \Sigma_Y \beta - 2\beta' \sigma_{XY}.$$

That variance is minimized with

$$\beta = \beta^* = \Sigma_Y^{-1} \sigma_{XY},$$

in which case

$$\mathrm{Var}[X_c] \ = \ (1 - R^2_{XY}) \mathrm{Var}[X],$$

where

$$R^2_{XY} = \frac{\sigma'_{XY} \Sigma_Y^{-1} \sigma_{XY}}{\mathrm{Var}[X]}$$

is the coefficient of determination (the square of the multiple correlation coefficient) between X and Y. So, the variance could be reduced by either positive or negative correlation, and R^2_{XY} indicates the fraction of the variance that is reduced. In the best possible case, if the multiple correlation is ± 1, the variance is reduced to zero. In the worst case, there is no correlation and the variance is unchanged.

A major difficulty with the CV method is that β^* is typically unknown (sometimes Σ_Y may be known, but practically never Σ_{XY}). Suppose that n independent replications of the simulation are performed. Then, Σ_Y and Σ_{XY} may be estimated by their sample counterparts $\hat{\Sigma}_Y$ and $\hat{\Sigma}_{XY}$, and β^* replaced by $\hat{\beta} = \hat{\Sigma}_Y^{-1} \hat{\sigma}_{XY}$. Let $X_{ce,i}$, $i = 1, \ldots, n$ denote the n replicates of the controlled estimator: $X_{ce,i} = X_i - \hat{\beta}(Y_i - \nu)$, where (X_i, Y_i) is the ith replicate of (X, Y). Let $\bar{X}_{ce}$ and s^2_{ce} be the sample average and sample variance of those $X_{ce,i}$. The CV estimator of μ is then $\bar{X}_{ce}$. Estimating μ and β^* that way turns out to be equivalent to fitting a least-squares regression model of the form $X = \mu + \beta'(Y - \nu) + \epsilon$ to the simulation data.

If we assume that (X, Y) is multinormal, then $\sqrt{n}(\bar{X}_{ce} - \mu)/s_{ce}$ follows the Student t distribution with $n - q - 1$ degrees of freedom (which implies that $\bar{X}_{ce}$ is unbiased), and

$$\frac{\mathrm{Var}[\bar{X}_{ce}]}{\mathrm{Var}[\bar{X}]} = \frac{n-2}{n-q-2}(1 - R^2_{XY}).$$

The latter ratio indicates that the number q of control variables must remain small relative to n.

Unfortunately, the multinormality assumption is not always realistic in practice. Without that assumption, the CV estimator is generally biased and may have a larger variance than the standard one for small n. However, it is generally true that $\sqrt{n}(\bar{X}_{ce} - \mu)/s_{ce} \Rightarrow N(0,1)$ and $s^2_{ce} \stackrel{\text{a.s.}}{\to} (1 - R^2_{XY})\mathrm{Var}[X]$ as $n \to \infty$ (Nelson 1990). Therefore, asymptotically, $\bar{X}_{ce}$ always has a smaller MSE than $\bar{X}$ and there is no loss in having to estimate β^*. Techniques for reducing the bias for small n include jackknifing and splitting; see Avramidis and Wilson (1993), Bratley, Fox, and Schrage (1987) and Nelson (1990).

For more details and further developments on CVs, recommendations, and applications, see also Avramidis, Bauer Jr., and Wilson (1991), Bauer Jr. and Wilson (1992), Fishman (1989), Lavenberg and Welch (1981), Lavenberg, Moeller, and Welch (1982), Porta Nova and Wilson (1993) and Tan and Gleser (1993). The above setup is easy to generalize to the case where μ and the response X are vectors; the variance is then replaced by the generalized variance, i.e., the determinant of the covariance matrix (Rubinstein and Marcus 1985). Nonlinear control variate models could also be considered; however, Glynn (1994a) shows that from the standpoint of asymptotic efficiency (as $n \to \infty$), there is no loss in restricting ourselves to linear schemes as above.

6. IMPORTANCE SAMPLING

Importance sampling (IS) amounts to changing the probability law(s) in order to concentrate the sampling effort in the most important parts of the sample space. It is particularly effective for dealing with *rare events*, by concentrating the sampling in the areas where the rare events are most likely to occur. IS received much renewed attention recently for estimating the probability of certain rare (but expensive) events in two classes of applications: (a) failures in highly dependable systems and (b) buffer overflows and long waiting times in queueing systems. In these application settings, standard estimators are highly inefficient because of the huge amount of simulation time that is typically required to observe a sufficient number of those events.

The idea of IS is to replace the probability measure P by another law Q such that Q dominates P over the region where $h(\omega) \neq 0$; that is, for all $B \in \mathcal{B}$, $\int_B h(\omega)dP(\omega) > 0$ implies $Q(B) > 0$. Then, the *likelihood ratio* $L(P, Q, \omega) = (dP/dQ)(\omega)$ exists and one can write:

$$\begin{aligned} E[h(\omega)] &= \int_\Omega h(\omega)dP(\omega) \\ &= \int_\Omega [h(\omega)(dP/dQ)(\omega)]\, dQ(\omega) \\ &= E_Q\left[h(\omega)L(P,Q,\omega)\right]. \end{aligned}$$

where E_Q is the expectation corresponding to Q. This means that an alternative unbiased estimator for $\mu = E[X]$ is $X_{is} = h(\omega)L(P, Q, \omega)$, where ω is generated from Q.

The optimal Q is given by $Q^*(d\omega) = |h(\omega)| P(d\omega)/\mu^*$, where $\mu^* = \int_\Omega |h(\omega)|dP(\omega)$ is a normalization constant. This Q^* yields the estimator $X^*_{is} = (I[h(\omega) > 0] - I[h(\omega) < 0])\mu^*$, where I is the indicator function. Note that if $P[X \geq 0] = 1$ or if $P[X \leq 0] = 1$, then X^*_{is} is equal to μ with probability one, so the variance is reduced to zero! Unfortunately, finding Q^* is typically much too complicated in practice; it is generally as hard as computing μ itself. This result nevertheless indicates that we should try to construct a Q which is roughly proportional to $|h|P$, and that can often be exploited in practical applications. Is the variance always reduced ? No. Perhaps the worst thing about IS is that the method is often extremely sensitive to the choice of Q. A bad choice may easily increase the variance to infinity!

Example 2 Suppose that we want to estimate $\mu = P[A]$ where $A \in \mathcal{B}$ is a rare event. The standard estimator is $X = I[A]$, whose variance (and MSE) is $\mu(1-\mu)$, and whose absolute error (the square root of the variance) is $\sqrt{\mu(1-\mu)}$. Since μ is very small, both of these quantities are small. However, obtaining a small MSE is trivial here; for example, one might as well just take 0 as an estimator and the MSE would be μ^2. So, it appears more meaningful in this case to consider the *relative MSE*, defined as $\mathrm{MSE}[X]/\mu^2$, or the *relative error*, $\mathrm{RE}[X] = \sqrt{\mathrm{MSE}[X]}/\mu$. For this example, one has $\mathrm{RE}[X] = \sqrt{(1-\mu)/\mu}$, which goes to infinity as μ approaches zero. Of course, the relative error would be divided by $\sqrt{n}$ by making n independent replications of the simulation, but keeping it under control when μ is very small is often much too costly. For instance, if $\mu \approx 10^{-10}$, then we would need $n \approx 10^{12}$ for a 10% relative error.

Here, the optimal Q (which gives zero variance) is $Q^*(\cdot) = I[A]P[\cdot]/P[A] = P[\cdot \mid A]$, the conditional distribution given that A has occured. This Q^* reallocates all of the sampling effort to the area where the rare event A occurs. In practice, one would seek a Q that resembles Q^* and which is easy to sample from. For a general Q, one has $\mathrm{Var}[X_{is}] = E_Q[(I[A] \cdot L(P,Q,\omega))^2] - \mu^2 = E_P[(I[A] \cdot L(P,Q,\omega))] - \mu^2$, so the variance will be reduced if the likelihood ratio tends to be small when A occurs.

Let us parameterize our model by a *rarity parameter* ϵ, so that h, P, μ, X, and X_{is} now depend on ϵ. Suppose that the events or interest get rarer and that $\mathrm{RE}[X(\epsilon)] \to \infty$ as $\epsilon \to 0$. The IS estimator X_{is} (or another alternative estimator) is said to have *bounded relative error* if $\mathrm{RE}[X_{is}(\epsilon)]$ remains bounded as $\epsilon \to 0$. If a probability measure $Q(\epsilon)$ can be found such that $\mathrm{Var}_{Q(\epsilon)}[X_{is}(\epsilon)] \leq K\mu^2(\epsilon)$ for some constant K, then $\mathrm{RE}[X_{is}] \leq \sqrt{K}$ as $\epsilon \to 0$. In the previous example, that will happen if $L(P(\epsilon), Q(\epsilon), \omega) \leq K\mu(\epsilon)$ whenever A occurs. The latter implies $E_{Q(\epsilon)}[I[A]] = E_P[I[A]/L(P(\epsilon), Q(\epsilon), \omega)] \geq 1/K$; that is, A is no longer a rare event under $Q(\epsilon)$. Observe that since the variance is non-negative, $E_{Q(\epsilon)}[X^2_{is}(\epsilon)]$ cannot approach zero faster than $\mu^2(\epsilon)$. When $\log E_{Q(\epsilon)}[X^2_{is}(\epsilon)] \sim \log \mu^2(\epsilon)$, the IS scheme is sometimes called *asymptotically optimal* or *asymptotically efficient*. This means that the relative error grows slower than exponentially fast as $\epsilon \to 0$; it is weaker than having a bounded relative error. Knowing that a given IS estimator has bounded relative error does not mean that it minimizes the variance for any given value of ϵ (and even asymptotically as $\epsilon \to 0$), but it is certainly a large step in the right direction.

Often, h depends on ω only through a sequence of independent random variables $\zeta_0, \zeta_1, \ldots, \zeta_T$ that are generated during the simulation, where $T = T(\omega)$ is a stopping time for $\{\zeta_j,\ j \geq 0\}$, with $P[T < \infty] = 1$, and ζ_j has density f_j. (To be more general, one can also replace $f_j(\zeta_j)$ by $f_j(\zeta_j \mid \zeta_0, \ldots, \zeta_{j-1})$.) One can then replace each f_j by another density g_j with the same support, and the likelihood ratio becomes

$$L(P, Q, \omega) = \frac{f_0(\zeta_0) \cdots f_T(\zeta_T)}{g_0(\zeta_0) \cdots g_T(\zeta_T)}.$$

The formula is similar if the ζ_j's are discrete, with the densities replaced by probability mass functions.

In several rare event contexts, it turns out that the (sometimes only) asymptotically optimal change of measure is the so-called *exponential twisting*: take $g_j(x) = K(\theta)^{-1} \exp(\theta x) f_j(x)$ for some constant θ and with the normalization factor $K(\theta) = E[\exp(\theta \zeta_j)]$. Finding the right value of θ and proving asymptotic optimality can often be done using large deviations theory (Bucklew 1990; Glynn 1994a; Heidelberger 1993).

Example 3 Consider a $GI/GI/1$ queue where A_i is the interarrival time between customers i and $i+1$, while B_i and W_i are the service time and waiting time of customer i, respectively. Suppose that we are interested in estimating $\mu = P[W > \ell]$, where W is the steady-state waiting time and ℓ is a fixed constant. Here, the rarity parameter could be taken as $\epsilon = 1/\ell$. Let $S_k = \sum_{i=1}^{k}(B_i - A_i)$. It is well known that W has the same distribution as $M = \max\{S_k,\ k \geq 0\}$, the maximum of a random walk with negative drift (assuming that the queue is stable). Let T be the smallest k for which $S_k > \ell$. Finding out whether or not $M > \ell$ by simulation normally requires simulating the first T customers, and $T = \infty$ whenever the event $\{M > \ell\}$ does not occur. We would like to change the probability laws to make that event occur with probability one and, ideally, have the system evolve according to the original distribution conditioned on the event $\{M > \ell\}$. Large deviations theory tells us how to approximately achieve that. Let $M(\theta) = E[\exp(\theta(B_i - A_i))]$ and choose $\theta^* > 0$ such that $M(\theta^*) = 1$ (such a θ^* exists if $M(\theta)$ is finite in a neighborhood of 0). Let A_i and B_i have densities f_A and f_B, respectively, and replace those densities by the exponentially twisted densities $g_A(x) = \exp(-\theta^* x) f_A(x) / E[\exp(-\theta^* A_i)]$ and $g_B(x) = \exp(\theta^* x) f_B(x) / E[\exp(\theta^* B_i)]$. Observe that $E[\exp(-\theta^* A_i)] E[\exp(\theta^* B_i)] = M(\theta^*) = 1$, so after T customers, the likelihood ratio becomes:

$$L(P, Q, \omega) = \exp(-\theta^* S_T).$$

That likelihood ratio is the estimator of $P[M > \ell]$ under the probability measure Q which corresponds to the twisted densities. It can be proved not only that this IS scheme is asymptotically optimal, but also that it is the only asymptotically optimal one within a large class of alternative distributions Q.

For more details on the previous example, and for other related examples, see Glynn (1994a), Heidelberger (1993), and the references therein. These references discuss in particular other types of rare events in single server queues, multiple server queues, queues with correlated arrival processes (such as Markov modulated queues), and reliability models. Additional recent references on queueing applications (and analysis) include Anantharam (1992), Chang et al. (1993), Chang, Heidelberger, and Shahabuddin (1993), Devetsikiotis and Townsend (1993), Frater, Lenon, and Anderson (1991), Frater and Anderson (1994), Glasserman and Kou (1994), Kesidis and Walrand (1993), Parekh and Walrand (1989), Sadowsky (1991), and Sadowsky (1993). Ross and Wang (1993) and Ross, Tsang, and Wang (1994) have designed a variant of IS for estimating the normalization constants involved in the solutions of (multiclass) product-form closed queueing networks. IS for reliability models is studied more extensively in Goyal et al. (1992), Heidelberger, Shahabuddin, and Nicola (1994), Nakayama (1994a), Nakayama (1994b), Shahabuddin (1994) and the references therein. Andradóttir, Heyman, and Ott (1993a), Glynn and Iglehart (1989) and Glynn (1994b) analyze IS for Markov chains in general.

7. CONDITIONAL MONTE CARLO

The general idea of CMC (also called the method of *conditional expectation*) is to replace the estimator $X = h(\omega)$ by its conditional expectation given another random variable Z. Roughly, if Z contains much less information than X, then the CMC estimator

$$X_{cm} \stackrel{\text{def}}{=} E[X \mid Z]$$

should have much less variability than X. More specifically, one has (Bratley, Fox, and Schrage 1987) $E[X_{cm}] = E[X]$ and $\mathrm{Var}[X_{cm}] = \mathrm{Var}[X] - E[\mathrm{Var}[X \mid Z]]$, so the variance can only decrease. The variance will be reduced to zero if Z tells us no information about X and will remain the same if X can be expressed as a function of Z alone. More generally, Z can be a vector or even a stochastic process. From a variance point of view, it is best to select a Z that contains as little information as possible, but from a computational point of view, X_{cm} may become too expensive or impossible to compute if Z contains too little information. Therefore, in terms of efficiency, there is a tradeoff to be made.

As an interesting special case, if the system of interest is a continuous-time Markov chain and if the response X can be expressed as the integral over time of a stochastic process whose value at any time depends only on the state of the chain at that time, then one can condition on the sequence of states visited by the chain, i.e., replace the holding times (which are exponential in this case) by their conditional expectations. This can also be generalized to semi-Markov processes and is called *discrete-time conversion* (Fox and Glynn 1986; Fox and Glynn 1990).

In several practical situations, $h(\omega)$ can be written as a sum of the form $h(\omega) = \sum_{i=1}^{t} h_i(\omega_i)$ where t is fixed (say), ω_i represents a "part" of ω that is observable at step i, and $h_i(\omega_i)$ is a "cost" incurred at step i. Instead of conditioning on the same Z for all i, it is often much more convenient to replace each $h_i(\omega_i)$ by $X_{ecm,i} = E[h_i(\omega_i) \mid Z_i(\omega_i)]$; that is, to use a different filter Z_i at each step, based only on the information available at that time. This is called *extended CMC*. It is always true that $\text{Var}[X_{ecm,i}] \leq \text{Var}[h_i(\omega_i)]$ for each i, but not necessarily true that $X_{ecm} = \sum_{i=1}^{t} X_{ecm,i}$ has lower variance than $X = h(\omega)$, because of the possible correlation between the different terms of the sum. Fortunately, in most situations of practical interest, it turns out that $\text{Var}[X_{ecm}] < \text{Var}[X]$. Sufficient conditions for that to happen are given in Glasserman (1993b) and Glasserman (1993a).

8. INDIRECT ESTIMATION

Suppose that the mean μ of interest can be expressed as a (known) function of some other quantity η, say $\mu = f(\eta)$. Then, it may be more efficient to estimate η instead of μ, then apply f to the estimator of η. This is called *indirect estimation*. For example, to estimate the average sojourn time per customer in a single queue (including service), one can estimate the average waiting time in the queue, say w_q, (excluding service) and then add the expected service time, assuming that the latter is known. (This example is also a case of extended CMC.) Suppose now that we want to estimate the (steady-state) average queue size L_q. For the standard estimator, we simulate the system for a long time horizon and take the sample time-average. An alternative indirect estimator is based on Little's law $L_q = \lambda w_q$, where λ is the arrival rate: if λ is known, take the standard estimator of w_q and multiply it by λ. The same can be done with $L = \lambda w$, where L is the average number of customers in the system and w the average sojourn time. Under mild conditions, this reduces the variance asymptotically (Glynn and Whitt 1989). On the other hand, if λ is unknown and must be estimated from the data, then both the indirect and direct estimators (based on Little's law) are equally efficient asymptotically.

9. STRATIFICATION

The general idea of stratification is to partition the sample space into disjoint strata, in such a way that the variance within the individual strata tends to be smaller than the general variance. A nice and convincing illustration of the method is the "bank example" of Bratley, Fox, and Schrage (1987). Suppose that we perform N simulation runs, that there are S strata, and that N_s runs fall into strata s, where $N = \sum_{s=1}^{S} N_s$. Let $X_{s,i}$ denote the ith observation from stata s. If p_s is the probability of falling into strata s under the original distribution, then the stratified estimator is

$$X_s = \sum_{s=1}^{S} p_s \left(\frac{1}{N_s} \sum_{i=1}^{N_s} X_{s,i} \right).$$

If the simulations are performed as usual, then each N_s is a random variable with expectation $E[N_s] = p_s N$. This is called *poststratification*.

In some contexts, it is easy to fix the N_s's a priori; that is, to decide in advance to which stratum each run will belong (for example, if the stratum can be determined easily from a few random variables generated at the beginning of the simulation). Then, one may want to choose the N_s's that minimize the variance of X_s. It turns out that the variance is minimized when $N_s = N p_s \sigma_s / \sum_{j=1}^{S} p_j \sigma_j$, where $\sigma_s^2 = \text{Var}[X_{s,i}]$ is the variance within stratum s. (This solution neglects the fact that N_s must be an integer; but an approximately optimal integer solution can easily be built from it in general). One problem here is that the σ_s's are typically unknown; however they can be estimated from pilot runs. See Bratley, Fox, and Schrage (1987) and Nelson (1985).

10. COMBINED METHODS

To obtain more variance reduction, one may want to use several VRTs simultaneously in the same simulation experiment. For example, to compare two systems, one may perform n pairs of simulation runs for each system, with CRNs across the systems and AVs within each pair for each system. However, even if both CRNs and AVs are individually effective, their combination could conceivably be worse than using only one of them, due to the cross-correlations between the response for the first system and the corresponding antithetic response for the second system (see Kleijnen 1975; Law and Kelton 1991).

Schruben and Margolin (1978) proposed a strategy for combining the CRN and AV methods in an experimental design scheme based on the idea of blocking, for estimating a linear (regression) metamodel of a response expressed as a function of several design variables for the system of interest. They gave conditions under which variance reduction is guaranteed. Several extentions have then been made to that scheme, including the incorporation of control variables, consideration of second-order metamodels, and so on (Donohue, Houck, and Myers 1993; Tew and Wilson 1994).

Avramidis and Wilson (1994) study the pairwise combinations of CV, AV, LHS, and conditional Monte Carlo (CM) for estimating a single response in a finite-horizon model, establish sufficient conditions for the combinations to outbeat each of their constituents alone, and provide asymptotic variance comparisons, which turn out in favor of the combination of LHS with CM. They report large gains in a numerical illustration with a stochastic activity network. Andradóttir, Heyman, and Ott (1993b) and Kwon and Tew (1994) also analyze combined methods.

ACKNOWLEDGMENTS

This work has been supported by NSERC-Canada grant # OGP0110050 and FCAR-Québec grant # 93-ER-1654.

REFERENCES

Anantharam, V. 1992. On fast simulation of the time to saturation of slotted ALOHA. *Journal of Applied Probability*, 29:682–690.

Andradóttir, S., D. P. Heyman., and T. J. Ott. 1993a. Potentially unlimited variance reduction in importance sampling of Markov chains. Technical Report 93-5, Department of Industrial Engineering, University of Wisconsin–Madison.

Andradóttir, S., D. P. Heyman., and T. J. Ott. 1993b. Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Transactions on Modeling and Computer Simulation*, 3(3):167–189.

Avramidis, A. N., K. W. Bauer Jr.., and J. R. Wilson. 1991. Simulation of stochastic activity networks using path control variates. *Journal of Naval Research*, 38:183–201.

Avramidis, A. N., and J. R. Wilson. 1993. A splitting scheme for control variates. *Operations Research Letters*, 14:187–198.

Avramidis, A. N., and J. R. Wilson. 1994. Integrated variance reduction strategies for simulation. *Operations Research*. To appear.

Bauer Jr., K. W., and J. R. Wilson. 1992. Control-variate selection criteria. *Naval Research Logistics*, 39:307–321.

Bratley, P., B. L. Fox., and L. E. Schrage. 1987. *A Guide to Simulation.* second ed. New York: Springer-Verlag.

Bucklew, J. 1990. *Large Deviation Techniques in Decision, Simulation and Estimation.* New York: John Wiley and Sons.

Chang, C. S., P. Heidelberger., S. Juneja., and P. Shahabuddin. 1993. Effective bandwidth and fast simulation of ATM intree networks. In *Proceedings of the Performance'93 Conference*, ed. G. Iazeolla and S. S. Lavenberg, 41–58, Roma, Italy. Elsevier Science.

Chang, C. S., P. Heidelberger., and P. Shahabuddin. 1993. Fast simulation of packet loss rates in a shared buffer communication switch. Technical Report No. 93-79, ICASE, NASA Langley Research Center, Hampton, VA.

Cheng, R. C. H. 1982. The use of antithetic variates in computer simulations. *Journal of the Operational Research Society*, 33:229–237.

Cheng, R. C. H. 1984. Antithetic variate methods for simulation of processes with peaks and troughs. *European Journal of Operational Research*, 15:227–236.

Devetsikiotis, M., and K. R. Townsend. 1993. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Transactions on Networking.* To appear.

Donohue, J. M., E. C. Houck., and R. H. Myers. 1993. A sequential experimental design procedure for the estimation of first- and second-order simulation metamodels. *ACM Transactions on Modeling and Computer Simulation*, 3(3):190–224.

Fishman, G. S. 1989. Monte Carlo, control variates, and stochastic ordering. *SIAM Journal on Scientific and Statistical Computing*, 10:187–204.

Fishman, G. S., and V. G. Kulkarni. 1992. Improving Monte Carlo efficiency by increasing variance. *Management Science*, 38:1432–1444.

Fishman, G. S., and B. D. Wang. 1983. Antithetic variates revisited. *Communications of the ACM*, 26:964–971.

Fox, B. L., and P. W. Glynn. 1986. Discrete-time conversion for simulating semi-Markov processes. *Operations Research Letters*, 5:191–196.

Fox, B. L., and P. W. Glynn. 1989. Simulating discounted costs. *Management Science*, 35(11):1297–1315.

Fox, B. L., and P. W. Glynn. 1990. Discrete-time conversion for simulating finite-horizon Markov processes. *SIAM Journal on Applied Mathematics*, 50:1457–1473.

Frater, M. R., and B. D. O. Anderson. 1994. Fast simulation of buffer overflows in tandem networks of $GI/GI/1$ queues. *Annals of Operations Research*, 49:207–220.

Frater, M. R., T. M. Lenon., and B. D. O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queuing networks. *IEEE Transactions on Automatic Control*, AC-36:1395–1405.

Glasserman, P. 1993a. Filtered monte carlo. *Mathematics of Operations Research*, 18:610–634.

Glasserman, P. 1993b. Stochastic monotonicity and conditional Monte Carlo for likelihood ratios. *Advances in Applied Probability*, 25:103–115.

Glasserman, P., and S.-G. Kou. 1994. Analysis of an importance sampling estimator for tandem queues. Submitted.

Glasserman, P., and D. D. Yao. 1992. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908.

Glynn, P. W. 1994a. Efficiency improvement techniques. *Annals of Operations Research.* To appear.

Glynn, P. W. 1994b. Importance sampling for Markov chains: Asymptotics for the variance. *Stochastic Models.* To appear.

Glynn, P. W., and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science*, 35:1367–1392.

Glynn, P. W., and W. Whitt. 1989. Indirect estimation via $L = \lambda w$. *Operations Research*, 37:82–103.

Glynn, P. W., and W. Whitt. 1992. The asymptotic efficiency of simulation estimators. *Operations Research*, 40:505–520.

Goyal, A., P. Shahabuddin., P. Heidelberger., V. F. Nicola., and P. W. Glynn. 1992. A unified framework for simulating markovian models of highly reliable systems. *IEEE Transactions on Computers*, C-41:36–51.

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte Carlo Methods.* London: Methuen.

Heidelberger, P. 1993. Fast simulation of rare events in queueing and reliability models. In *Performance Evaluation of Computer and Communication Systems*, ed. L. Donatiello and R. Nelson, volume 729 of *Lecture Notes in Computer Science*, 165–202. Springer Verlag.

Heidelberger, P., and D. L. Iglehart. 1979. Comparing stochastic systems using regenerative simulations with common random numbers. *Advances in Applied Probabability*, 11:804–819.

Heidelberger, P., P. Shahabuddin., and V. F. Nicola. 1994. Bounded relative error in estimating transient measures of highly dependable non-markovian systems. *ACM Transactions on Modeling and Computer Simulation*, 4(2):137–164.

Kesidis, G., and J. Walrand. 1993. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. *ACM Transactions on Modeling and Computer Simulation*, 3(3):269–276.

Kleijnen, J. P. C. 1975. Antithetic variates, common random numbers and optimal computer time allocation in simulations. *Management Science*, 21:1176–1185.

Kwon, C., and J. D. Tew. 1994. Combined correlation induction strategies for designed simulation experiments. *Management Science.* To appear.

Lavenberg, S. S., T. L. Moeller., and P. D. Welch. 1982. Statistical results on multiple control variables with application to queueing network simulation. *Operations Research*, 30:182–202.

Lavenberg, S. S., and P. D. Welch. 1981. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science*, 27:322–335.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis.* second ed. New York: McGraw-Hill.

L'Ecuyer, P. 1992. Convergence rates for steady-state derivative estimators. *Annals of Operations Research*, 39:121–136.

L'Ecuyer, P., and S. Côté. 1991. Implementing a random number package with splitting facilities. *ACM Transactions on Mathematical Software*, 17(1):98–111.

L'Ecuyer, P., and G. Perron. 1994. On the convergence rates of IPA and FDC derivative estimators. *Operations Research.* To appear.

L'Ecuyer, P., and F. Vázquez-Abad. 1994. Functional estimation with respect to a threshold parameter. Submitted.

L'Ecuyer, P., and G. Yin. 1994. Convergence rate of stochastic optimization algorithms with budget-dependent bias. Submitted.

Nakayama, M. K. 1994a. A characterization of the simple failure biasing method for simulations of highly reliable markovian systems. *ACM Transactions on Modeling and Computer Simulation.* To

appear.
Nakayama, M. K. 1994b. Fast simulation methods for highly reliable systems. In *Proceedings of the 1994 Winter Simulation Conference*. IEEE Press. (these proceedings).
Nelson, B. L. 1985. An illustration of the sample space definition of simulation and variance reduction. *Transactions of the Society for Computer Simulation*, 2:237–247.
Nelson, B. L. 1986. Decomposition of some well-known variance reduction techniques. *Journal of Statistical and Computer Simulation*, 23:183–209.
Nelson, B. L. 1987a. A perspective on variance reduction in dynamic simulation experiments. *Communications in Statistics—Simulation and Computation*, B16:385–426.
Nelson, B. L. 1987b. Variance reduction for simulation practitioners. In *Proceedings of the 1987 Winter Simulation Conference*, 43–51. IEEE Press.
Nelson, B. L. 1990. Control-variate remedies. *Operations Research*, 38:974–992.
Nelson, B. L. 1993. Robust multiple comparisons under common random numbers. *ACM Transactions on Modeling and Computer Simulation*, 3(3):225–243.
Nelson, B. L., and J. C. Hsu. 1993. Control-variate models of common random numbers for multiple comparisons with the best. *Management Science*, 39(8):989–1001.
Parekh, S., and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, AC-34:54–56.
Porta Nova, A., and J. R. Wilson. 1993. Selecting control variates to estimate multiresponse simulation metamodels. *European Journal of Operational Research*, 71:80–94.
Ross, K. W., D. H. K. Tsang., and J. Wang. 1994. Monte Carlo summation and integration applied to multiclass queueing networks. *Journal of the ACM*. To appear.
Ross, K. W., and J. Wang. 1993. Asymptotically optimal importance sampling for product-form queueing networks. *ACM Transactions on Modeling and Computer Simulation*, 3(3):244–268.
Rubinstein, R. Y., and R. Marcus. 1985. Efficiency of multivariate control variates in Monte Carlo simulation. *Operations Research*, 33:661–667.
Sadowsky, J. S. 1991. Large deviations and efficient simulation of excessive backlogs in a $GI/G/m$ queue. *IEEE Transactions on Automatic Control*, AC-36:1383–1394.
Sadowsky, J. S. 1993. On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Transactions on Information Theory*, IT-39:119–128.
Schruben, L. W., and B. H. Margolin. 1978. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association*, 73:504–525.
Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable markovian systems. *Management Science*, 40(3):333–352.
Tan, M., and L. J. Gleser. 1993. Improved point and confidence interval estimators of mean response in simulation when control variates are used. *Communications in Statistics—Simulation*, 22:1211–1220.
Tew, J. D., and J. R. Wilson. 1994. Estimating simulation metamodels using combined correlation-based variance reduction techniques. *IEE Transactions*. To appear.
Wilson, J. R. 1983. Antithetic sampling with multivariate inputs. *American Journal of Mathematical and Management Sciences*, 3:121–144.
Wilson, J. R. 1984. Variance reduction techniques for digital simulation. *American Journal of Mathematical and Management Sciences*, 4:277–312.
Yang, W., and B. L. Nelson. 1991. Using common random numbers and control variates in multiple comparison procedures. *Operations Research*, 39(4):583–591.

AUTHOR BIOGRAPHY

PIERRE L'ECUYER is a professor in the department of "Informatique et Recherche Opérationnelle" (IRO), at the University of Montreal. He received a Ph.D. in operations research in 1983, from the University of Montreal. From 1983 to 1990, he was with the computer science department, at Laval University, Québec. His research interests are in Markov renewal decision processes, sensitivity analysis and optimization of discrete-event stochastic systems, random number generation, and discrete-event simulation in general. He is the Departmental Editor for the Simulation Department of *Management Science* and an Area Editor for the *ACM Transactions on Modeling and Computer Simulation*.