EFFICIENCY IMPROVEMENT TECHNIQUES

Peter W. GLYNN

Department of Operations Research, Stanford University, Stanford, California 94305-4022

Abstract

This paper provides an overview of the five most commonly used statistical techniques for improving the efficiency of stochastic simulations: control variates, common random numbers, importance sampling, conditional Monte Carlo, and stratification. The paper also describes a mathematical framework for discussion of efficiency issues that quantifies the trade-off between lower variance and higher computational time per observation.

Keywords

variance reduction, simulation, control variates, importance sampling, common random numbers, stratification, conditioning, efficiency

1. Introduction

This paper is intended to give the reader an overview of some of the basic issues that arise in the context of "efficiency improvement techniques (EIT's)", as well as some of the tools that have been developed to deal with those issues. As we shall see in Section 2, efficiency improvement can arise either because of computational enhancements or because of variance reduction. Statistical knowledge of the system under consideration can be used to impact either of these two factors, or both. Our principal concern here will be with using this statistical knowledge to improve computational efficiency. In particular, we will not be concerned with any explicit discussion about how programming practices and choice of data structures can potentially impact efficiency issues. Historically, the body of research concerned with use of statistical methods to im-

prove the efficiency of simulation algorithms has been generally referred to as "variance reduction methodology". However, we choose to use the EIT terminology here, in order to highlight the fact that statistical tools can be used to impact the running time of an algorithm either by choosing a statistically equivalent estimator with substantially lower computational cost or by developing an estimator with reduced variance (or, preferably, both).

This paper is organized as follows. In Section 2, we provide a mathematical framework within which efficiency can be studied precisely. In particular, the framework gives the trade-off between variance reduction and computation time. It is worth noting that Nelson [29, 30] provides a different type of framework for these methods, one that is intended to provide a convenient taxonomy of such techniques.

Sections 3 through 7 discuss the five most commonly used EIT's: control variates, common random numbers, importance sampling, conditional Monte Carlo, and stratification. For a description of some additional (less commonly used) methods, see Hammersley and Handscomb [24] and Wilson [35].

2. Asymptotic Efficiency of Simulation Estimators

Suppose that α is a real-valued quantity that we wish to estimate via simulation. Let t be the available simulation budget, measured in terms of computer time. Let $\alpha_1(t)$ and $\alpha_2(t)$ be two competing estimators, both of which can be constructed within the simulation budget of size t. We wish to develop criteria that will permit us to compare the quality of the two estimators. As is true of much of the statistics and probability literature, we will concentrate our efforts on developing criteria that describe the asymptotic behavior of the estimators as $t \to \infty$, in large part because this is a setting which lends itself to greater mathematical tractability.

Typically, each of the estimators will satisfy a limit theorem of the form

$$t^{\gamma_i}(\alpha_i(t) - \alpha) \Rightarrow L_i \tag{2.1}$$

as $t \to \infty$ (i = 1, 2), where γ_1, γ_2 are positive constants, $\Rightarrow$ denotes "weak convergence", and L_1, L_2 are finite-valued r.v.'s that are not equal to zero with probability one. As we shall see below, the most common form of (2.1) is the situation in which $\gamma_i = 1/2$ and $L_i = N(0, \sigma_i^2)$; this arises as

a consequence of central limit theorem (CLT) type arguments. The relation (2.1) implies that

$$\alpha_i(t) \approx \alpha + t^{-\gamma_i} L_i$$

as $t \to \infty$ ($i = 1, 2$), where $\approx$ means "approximately equal in distribution to". From (2.2), it is clear that, for large t, one ought to prefer the estimator with the larger value of γ_i. However, in many estimator comparisons, $\gamma_1 = \gamma_2 = 1/2$ (the standard CLT normalization), so that a finer criterion is needed. Under fairly mild additional regularity conditions, (2.2) yields

$$(E|\alpha_i(t) - \alpha_i|^p)^{1/p} \sim t^{-\gamma_i} (E|L_i|^p)^{1/p} \tag{2.3}$$

as $t \to \infty$ ($i = 1, 2$), for $p \geq 0$. Relation (2.3) establishes an asymptotic for the so-called L_p error of the competing estimators. The most commonly used values of p are 2 (root mean square error) and 1 (mean absolute deviation). Given that a value of p has been selected and that $\gamma_1 = \gamma_2$, it seems reasonable to prefer that estimator with the lower value of $(E|L_i|^p)^{1/p}$. Thus, in comparing two competing estimators, one should use the following "lexicographic" criterion: select the estimator with the highest value of γ_i. If a tie occurs there, select the estimator with the lowest value of $(E|L_i|^p)^{1/p}$. If a tie is still present, the theory just described asserts that the behavior of the two estimators is identical, at least asymptotically. (Of course, the small-sample behavior may be quite different.)

As has been remarked earlier, it is common that $\gamma_1 = \gamma_2 = 1/2$ with $L_i = N(0, \sigma_i^2)$, ($i = 1, 2$). In this special situation, it turns out that the value of p is irrelevant, since $(E|L_1|^p)^{1/p} \geq (E|L_2|^p)^{1/p}$ then occurs if and only if $\sigma_1^2 \geq \sigma_2^2$ (because the two normal r.v.'s then differ by a scale parameter corresponding to the ratio of the standard deviations). In particular, finding the estimator with the minimal root mean square error simultaneously minimizes the L_p error for all $p>0$.

We will now illustrate the above framework. Perhaps the most important class of estimation problems that arises in simulation can be put into the following form. We are to estimate a parameter α that can be expressed as $\alpha = g(\mu)$, where $g:\Re^d \to \Re$ is assumed continuously differentiable in a neighborhood of μ, and $\mu = EX$ for some $\Re^d$-valued random variable X. Suppose that we can generate i.i.d. copies $X_1, X_2, \ldots$ of the r.v. Set $\overline{X}_n = n^{-1}(X_1 + X_2 + \cdots + X_n)$ and let $\alpha_n = g(\overline{X}_n)$. In order to study the efficiency of this estimator, we need to obtain a limit theorem of the form (2.1), expressed on the time scale of computer time. This requires that we model the interaction

between the computational and statistical characteristics of the problem. Let τ_i be the amount of computer time required to generate X_i. We shall assume (reasonably) that the sequence of pairs $((X_i, \tau_i) : i \geq 1)$ is i.i.d., although X_i typically will be correlated with τ_i for each i. Let $N(t)$ be the number of X_i's generated in t units of computer time, so that $N(t) = \max(n \geq 0: \tau_1 + ...+\tau_n \leq t)$. Then, the estimator available after t time units of computational effort have been expended is given by $\alpha(t) \equiv \alpha_{N(t)}$. Standard arguments establish that if $0 < E\tau_1 < \infty$ and $E||X_1||^2 < \infty$, then

$$t^{1/2}(\alpha(t) - \alpha) \Rightarrow N(0, \sigma^2) \tag{2.4}$$

as $t \rightarrow \infty$, where $\sigma^2 = E\tau_1 \cdot \nabla g(\mu) \cdot \Sigma \cdot \nabla g(\mu)^T$ and Σ is the covariance matrix of X_1; see Glynn and Whitt (1992) for details. When $g(x) = x$ and X_1 is real-valued, then

$$\sigma^2 = E\tau \cdot \text{var}(X_1).$$

In view of our above efficiency discussion, it follows that in comparing two estimators of the type just decribed, one should select that estimator with the smaller value of σ^2. Note that in the case that $g(x) = x$ and X_1 is real-valued, the more efficient estimator may have the higher variance, provided that the average time required to generate an observation is low enough. Thus, variance reduction and efficiency improvement do not necessarily go hand-in-hand. This is perhaps the principal reason why we choose, in this paper, to use the term "efficiency improvement technique" rather than "variance reduction technique". An example of this kind of trade-off is illustrated in Fishman and Kulkarni [8].

We now proceed to discuss limit theorems of the form (2.1) for several other types of estimators that arise in the simulation context.

EXAMPLE (2.1)

Underlying (2.4) was an assumption that one can generate i.i.d. copies $X_1, X_2, ...$ of a r.v. X that is unbiased for μ in the sense that $EX = \mu$. Typically, this type of unbiasedness holds only in the finite-horizon terminating simulation context. However, it should be noted that the regenerative method of steady-state simulation effectively reduces the computation of steady-state means to finite-horizon quantities. In particular, by taking advantage of regenerative structure in the stochastic system under consideration, one can express the steady-state mean of such a process in terms of quantities that involve only a single regenerative cycle. Regenerative steady-state simula-

tion turns out to be a special case of (2.4), in which τ_i corresponds to the amount of computer time required to generate the i'th independent replicate of the cycle and g takes the ratio form $g(x_1, x_2) = x_1/x_2$. It is worth noting that while the resulting ratio estimator is consistent, it can suffer from small-sample bias problems because of the non-linearity of the function g. Of course, a different approach is needed in order to deal with non-regenerative stochastic processes.

Suppose that α is a real-valued parameter that can be expressed as the steady-state mean of a real-valued stochastic process $X = (X(t) : t \geq 0)$. More precisely, assume that

$$\overline{X}(t) \equiv t^{-1} \int_0^t X(s)\, ds \Rightarrow \alpha \tag{2.5}$$

as $t \to \infty$. Under very mild additional conditions on the process, (2.5) can be strengthened to a CLT:

$$t^{1/2}(\overline{X}(t) - \alpha) \Rightarrow N(0, \sigma^2) \tag{2.6}$$

as $t \to \infty$, where σ^2 is a finite constant that is known as the time-average variance constant of X. Note that the time parameter t appearing in (2.6) measures simulated time and not computer time. We therefore need to re-express (2.6) in terms of computer time. Let $\Lambda = (\Lambda(t) : t \geq 0)$ be a non-decreasing process in which $\Lambda(t)$ represents the amount of time simulated with a computer time budget of size t. Then, $\alpha(t) \equiv \overline{X}(\Lambda(t))$ is the estimator for α available after t units of computer time have been expended. In virtually all steady-state simulation settings, it is reasonable to expect that $\Lambda(t)/t \Rightarrow \lambda$ as $t \to \infty$, where λ is a finite positive constant. The constant λ can be interpreted as the rate at which simulated time is generated relative to a unit of computer time. By applying "random time-change" arguments (such results establish conditions under which random variables like $\Lambda(t)$ can be substituted as time parameters in central limit theorems) to (2.6), one can then obtain the limit theorem

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

as $t \to \infty$. Thus, in comparing the efficiency of two such steady-state estimators for a parameter α, one needs to compare the magnitudes of the corresponding variance quantities that here take the form $\lambda_i^{-1}\sigma_i^2$ (i =1, 2).

EXAMPLE (2.2)

We now give an example of a simulation algorithm in which the parameter γ appearing in (2.1) is not equal to 1/2. Suppose that we wish to calculate a parameter $\alpha = \theta^*$ that minimizes a smooth function $\beta(\theta)$. Assume that $\beta(\theta)$ can be represented as $\beta(\theta) = EZ(\theta)$ for some family of r.v.'s $(Z(\theta) : \theta \in \Re)$. Successive estimates of θ^* are obtained from the recursion $\theta_{n+1} = \theta_n - (c/n)X_{n+1}$. In this recursion, c is a positive constant and X_{n+1} is generated from the conditional distribution

$$P(X_{n+1} \in \cdot \mid \theta_0, X_0, \ldots, \theta_n, X_n) = P((Z(\theta_n + hn^{-1/3}) - Z(\theta_n - hn^{-1/3})) / (2hn^{-1/3}) \in \cdot)$$

where $Z(\theta_n + hn^{-1/3})$ and $Z(\theta_n - hn^{-1/3})$ are independently simulated and h is a positive constant. This type of optimization algorithm is known as the Kiefer-Wolfowitz stochastic approximation procedure. Suppose that $b = c\beta''(\theta^*)$, $A = b - 5/6$, $\kappa^2 = 2\text{var}Z(\theta^*)$, and $\sigma^2 = c^2\kappa^2/(2A+1)(4h^2)$. Let $\lambda^{-1}(\theta)$ be the mean amount of computer time required to generate $Z(\theta)$ and assume that $\lambda^{-1}(\cdot)$ is continuous in a neighborhood of θ^*. Let $\alpha(t)$ be the iterate of the sequence $(\theta_n : n \geq 1)$ available after t units of computer time have been expended. Then, under mild additional regularity hypotheses (see Ruppert [31]), it follows that

$$t^{1/3}(\alpha(t) - \alpha) \Rightarrow 2^{1/3}\lambda(\theta^*)^{-1/3}\sigma N(0,1) \qquad (2.7)$$

as $t \to \infty$. The key point in (2.7) is that the variance constant appearing in the limiting normal depends on the family of r.v.'s $(Z(\theta) : \theta \in \Re)$ only through the quantity $\lambda(\theta^*)^{-2/3} \cdot \text{var}Z(\theta^*)$. This implies, for example, that one family of r.v.'s, having half the variance of another, is preferable so long as the time required to generate an observation is less than or equal to $2^{3/2}$ times that of the higher variance estimator. Thus, relative to the examples described above, there is a comparatively higher premium in this setting to reduce variance rather than computer time. This occurs because of the "subcanonical" rate of convergence that appears in (2.7) (i.e. $\gamma < 1/2$).

EXAMPLE (2.3)

We now give an example of an estimation algorithm in which "supercanonical" rates of convergence are achieved, namely $\gamma > 1/2$ in (2.1). Suppose that we wish to evaluate the definite integral $\alpha \equiv \int_0^1 f(x)\,dx$, for some given function $f : [0, 1] \to \Re$. In Fishman and Huang [7], a

"rotation" estimator for this problem was proposed. An easily updated recursive version of this estimator was studied in Glynn and Whitt [22]. Let $\alpha(t)$ be the recursive estimator available after t units of computer time have been expended. In Glynn and Whitt [22], it is shown that

$$t^{3/4}(\alpha(t) - \alpha) \Rightarrow N(0, \sigma^2)$$

as $t \to \infty$, for some finite constant σ. It can be shown, in a manner similar to that used in Example 2.2, that in this supercanonical setting, there is comparatively a greater premium to reduce the mean time required to generate observations, as compared to reducing their variance.

EXAMPLE (2.4)

Here, our intent is to give an example in which the limiting r.v. L is not normally distributed with mean zero. As mentioned earlier, the choice of the quantity p can then have an impact on the determination of optimally efficient estimators. Suppose that our goal is to estimate $\alpha \equiv \beta'(\theta_0)$, where $\beta(\theta)$ is a smooth function of θ which can be represented in the form $\beta(\cdot) = EZ(\cdot)$ for some family of r.v.'s $(Z(\theta) : \theta \in \Re)$. Let X_k be the forward difference defined by $X_k \equiv (Z_k(\theta_0 + hk^{-1/4}) - (Z_k(\theta_0))/(hk^{-1/4})$, where $Z_k(\theta_0 + hk^{-1/4})$ and $Z_k(\theta_0)$ are independently generated and h is a positive constant. Set $\overline{X}_n \equiv n^{-1}(X_1 + \cdots + X_n)$, and let $\alpha(t) \equiv \overline{X}_{N(t)}$, where $N(t)$ is the number of X_i's available within t units of computer time. This recursive derivative estimator is studied in Glynn and Whitt [22]; it is shown there that under reasonable assumptions,

$$t^{1/4}(\alpha(t) - \alpha) \Rightarrow \lambda^{-1/4} N(\eta, \kappa^2)$$

where $\kappa^2 = 4\text{var}Z(\theta_0)/3h^2$, $\eta = 2\beta''(\theta_0)h/3$, and λ^{-1} is the mean time required to generate $Z(\theta_0)$. In this derivative estimation setting, the appropriate choice of estimation strategy may therefore depend on whether one uses mean square error or mean absolute deviation as one's optimality criterion.

For a more extensive and complete discussion of these efficiency issues, see Glynn and Whitt [22]. Throughout the rest of this paper, we shall focus on estimation algorithms in which (2.1) holds with $\gamma = 1/2$ and $L = N(0, \sigma^2)$.

3. Control Variates

The method of control variates is one of the most widely applied efficiency improvement techniques. It owes its popularity in part to the ease with which it can be interfaced with commonly available high-level simulation languages and in part to the number of practical settings in which it can easily be applied. We start by describing the method in its most general form.

Suppose that we wish to estimate a real-valued parameter α. We assume that α can be put in the form $\alpha = g(\mu)$, where $g : \Re^d \to \Re$ is continuously differentiable in a neighborhood of μ. The naive estimator for such problem typically requires existence of an $\Re^d$-valued process $X = (X(t) : t \geq 0)$ such that

$$\mu(t) \equiv t^{-1} \int_0^t X(s)\, ds \Rightarrow \mu$$

as $t \to \infty$. The naive estimator for α is then given by $\alpha(t) \equiv g(\mu(t))$. Many different estimation problems can be put in the above form, including estimation of a mean and ratio estimation.

The method of control variates assumes the existence of an associated $\Re^l$-valued process $Y = (Y(t) : t \geq 0)$ and a constant $c \in \Re^l$, known to the simulator, such that

$$C(t) \equiv t^{-1} \int_0^t Y(s)\, ds \Rightarrow c$$

as $t \to \infty$. We can now take advantage of the fact that the time-average of the process Y is known, as follows. Let $f : \Re^{k+l} \to \Re$ be a function that is chosen so that $f(\cdot, c) = g(\cdot)$ and such that f is continuously differentiable in a neigborhood of (μ, c). Then, it follows from the above laws of large numbers for $C(t)$ and $\mu(t)$ that the estimator $\alpha_c(t) \equiv f(\mu(t), C(t))$ is consistent for α. The question now comes down to choosing the function f so as to minimize the variance of the estimator just constructed. We note that the time-average C is typically referred to as a control in the literature.

We will now follow the analysis of Glynn and Whitt [21], which will establish that we may restrict the choice of f to functions that are essentially linear in the control. This will require that we strengthen the above laws of large numbers for $\mu(t)$ and $C(t)$ to a joint CLT. Specifically, we shall now assume that there exists a $(d + l) \times (d + l)$ positive definite matrix Σ such that

$$t^{1/2}(\mu(t) - \mu, C(t) - c) \Rightarrow N(0, \Sigma) \tag{3.1}$$

as $t \to \infty$, where

$$\Sigma = \begin{pmatrix} \Sigma_{\mu\mu} & \Sigma_{\mu c}^{t} \\ \Sigma_{\mu c} & \Sigma_{cc} \end{pmatrix}$$

and $\Sigma_{\mu\mu}$, $\Sigma_{\mu c}$, and Σ_{cc} are $d \times d$, $l \times d$, and $l \times l$ matrices, respectively.

We observe that (3.1) in fact implies the validity of the laws of large numbers stated earlier in this section. Consequently, $\mu(t)$ and $C(t)$ are therefore close to their respective limits μ and c for t large. A Taylor expansion of $f(\mu(t), C(t))$ about (μ, c) therefore yields

$$f(\mu(t), C(t)) = f(\mu, c) + \nabla_{\mu} f(\xi(t), \eta(t))\,(\mu(t) - \mu) + \nabla_{c} f(\xi(t), \eta(t))(C(t) - c) \qquad (3.2)$$

where $(\xi(t), \eta(t))$ lies on the line segment between (μ, c) and $(\mu(t), C(t))$. It follows from (3.1) and (3.2) that

$$\alpha_c(t) = \alpha + \nabla_{\mu} f(\mu, c)\,(\mu(t) - \mu) + \nabla_{c} f(\mu, c)(C(t) - c) + o_p(t^{-1/2}) \qquad (3.3)$$

and

$$t^{1/2}(\alpha_c(t) - \alpha) \Rightarrow N(0, \sigma_c^2) \qquad (3.4)$$

as $t \to \infty$, where $\sigma_c^2 = \nabla f(\mu, c) \cdot \Sigma \cdot \nabla f(\mu, c)^t$ and $o_p(t^{-1/2})$ represents a stochastic process having the property that $t^{1/2} o_p(t^{-1/2}) \Rightarrow 0$ as $t \to \infty$. A similar argument applies to $\alpha(t)$, yielding

$$\alpha(t) = \alpha + \nabla g(\mu)\,(\mu(t) - \mu) + o_p(t^{-1/2}) \qquad (3.5)$$

and

$$t^{1/2}(\alpha(t) - \alpha) \Rightarrow N(0, \sigma^2) \qquad (3.6)$$

as $t \to \infty$, where $\sigma^2 = \nabla g(\mu) \cdot \Sigma_{\mu\mu} \cdot \nabla g(\mu)^t$. But our choice of f guarantees that

$$\nabla_{\mu} f(\cdot, c) = \nabla g(\cdot)$$

and hence it follows from (3.3) and (3.5) that

$$\alpha_c(t) = \alpha(t) + \nabla_{c} f(\mu, c)(C(t) - c) + o_p(t^{-1/2}) \qquad (3.7)$$

as $t \to \infty$. Note that (3.4) and (3.6) imply that $\alpha_c(t)$ and $\alpha(t)$ have errors, as estimators of α, of order $t^{-1/2}$, while (3.7) shows that the two estimators differ from one another only through a linear combination of the control and an asymptotically negligible term of the order $o_p(t^{-1/2})$. Hence, it follows, from the standpoint of asymptotic efficiency, that we may as well restrict our search for good functions f to functions that differ from $\alpha(t)$ only through a linear combination of the control. (Of course, it may well be possible that in certain settings, the small-sample properties of the "non-linear" control described above are superior to those associated with the "linear" version.)

We henceforth restrict our attention to linear control schemes of the form

$$\alpha_l(t;\lambda) \equiv \alpha(t) - \lambda(C(t) - c).$$

The question now reduces to how to choose the multiplier λ appropriately, so as to minimize the variance of the corresponding estimator. Our joint CLT for $(\mu(t), C(t))$, together with (3.6), shows that

$$t^{1/2}(\alpha_l(t;\lambda) - \alpha) \Rightarrow N(0, \sigma^2(\lambda))$$

where $\sigma^2(\lambda) = \sigma^2 - 2\lambda \sum_{\mu c} \nabla g(\mu)^t + \lambda \sum_{cc} \lambda^t$. To minimize $\sigma^2(\lambda)$ is easy. The minimizing λ is given by

$$\lambda^* = \nabla g(\mu) \sum_{\mu c}^t \sum_{cc}^{-1}.$$

Of course, in practice, the minimizing λ^* is unknown, and must be estimated from the simulated data. Suppose that $\sum_{\mu c}(t)$ and $\sum_{cc}(t)$ are consistent estimators for $\sum_{\mu c}$ and $\sum_{cc}$, respectively, so that $\sum_{\mu c}(t) \Rightarrow \sum_{\mu c}$ and $\sum_{cc}(t) \Rightarrow \sum_{cc}$ as $t \to \infty$. Then, $\lambda(t) \equiv \nabla g(\mu(t)) \sum_{\mu c}(t)^T \sum_{cc}(t)^{-1}$ consistently estimates λ^* and it is easily verified that

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

so that no loss of asymptotic efficiency is incurred by having to estimate λ^*. Note that as the number of components in the control C expands, the minimizing value of the variance decreases. This suggests that it is optimal to use as many components in the control vector as possible, and this assertion is valid in a large-sample sense. However, the empirical evidence is that the method begins to degrade, both in terms of small-sample bias and variance, when too many components are added to the control. This is because the optimal λ^* becomes successively harder to estimate in finite-length samples. This "small-sample" effect has been extensively studied in the literature. One noteworthy result, in this vein, was obtained by Lavenberg, Moeller, and Welch [27]. They studied the situation in which g is linear and $(\mu(t), C(t))$ are multivariate normally distributed. In this setting, they obtain a precise small-sample characterization of the loss of efficiency engendered by having to estimate λ^*.

It may, at first, be unclear as to how controls, with known asymptotic mean, can be constructed in practice. However, controls are easy to construct in almost any simulation context. In particular, simulations are driven by i.i.d. sequences of random variables with known distributions. If one takes a sample mean of all the random variables that were generated from a given

specific distribution, the law of large numbers for i.i.d. sequences guarantees that this quantity will converge with probability one to the known mean of the underlying distribution. This type of sample mean can therefore be used as a component of a control vector *C(t)*. Note that the additional computational burden associated with using the method of control variates largely derives from the additional time required to collect statistics for the control vector. It seems reasonable to expect that this additional time is typically fairly modest, so that the gain in computational efficiency is roughly equal to the variance reductions described above.

It is also worth observing that in simulating an open network of queues, the exogenous arrival rates are known quantities. Thus, the empirical exogenous arrival rates may be used to form a vector of controls. This particular choice of control is closely related to another variance reduction technique called indirect estimation. See Glynn and Whitt [21] and Law [28] for additional details on the connection.

4. Common Random Numbers

The method of common random numbers is an efficiency improvement technique that is widely used in comparison of stochastic systems. To illustrate the method, suppose that we wish to compare the output of two different systems. Assume that the performance measure for system i (i =1, 2) is given by $f_i(X)$, where X is a real-valued r.v. and f_1, f_2 are two given real-valued functions. A reasonable way to compare two such systems is to examine the parameter $\alpha \equiv \mu_1 - \mu_2$, where $\mu_i = Ef_i(X)$ (i =1, 2). The naive approach to estimating α is to generate two independent streams $X_{11}, X_{12}, \ldots$ and $X_{21}, X_{22}, \ldots$, both consisting of i.i.d. replicates of the r.v. X. The naive estimator is then given by

$$\alpha_n \equiv n^{-1}((f_1(X_{11}) - f_2(X_{22})) + \cdots + (f_1(X_{1n}) - f_2(X_{2n}))).$$

Let τ_{ij} be the amount of computer time required to generate $f_i(X_{ij})$ and suppose that the τ_{ij}'s are also i.i.d. Let $N(t)$ be the number of (X_{1i}, X_{2i}) pairs generated in t units of computer time; then $\alpha(t) \equiv \alpha_{N(t)}$ is the estimator available after t units of computer time have been expended. Section 2's discussion implies that, under mild regularity hypotheses,

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

as $t \to \infty$, where $\sigma^2 = (E\tau_{11} + E\tau_{21}) \cdot (\text{var } f_1(X) + \text{var } f_2(X))$.

The above estimator subjects the two systems to independent random "shocks". Intuitively, it seems fairer to correlate the pattern of shocks, so that if one system receives an "unlucky" sequence of random inputs, then so does the other. Suppose, in particular, that in estimating μ_1 and μ_2 above, we use a common sequence of X_i's to drive both systems. Specifically, let X_1, X_2, ... be a sequence of i.i.d. copies of X, and set

$$\alpha_n^c \equiv n^{-1}((f_1(X_1) - f_2(X_1)) + \text{K} + (f_1(X_n) - f_2(X_n))).$$

Suppose that τ_i is the amount of time required to generate and calculate $f_1(X_i)$ - $f_2(X_i)$ and let $N(t)$ = max $(n \geq 0: \tau_1 + ... + \tau_n \leq t)$. Then, $\alpha^c(t) \equiv \alpha_{N(t)}^c$ is the estimator available after t units of computer time have been expended. Its CLT takes the form

$$t^{1/2}(\alpha^c(t) - \alpha) \Rightarrow N(0, \sigma_c^2)$$

as $t \to \infty$, where $\sigma_c^2 = E\tau_1 \cdot \text{var}(f_1(X_1) - f_2(X_1))$. This "common random numbers" (CRN) estimator is therefore more efficient than the naive estimator if $\sigma_c^2 \leq \sigma^2$. It is clear that the amount of time required to compute $f_1(X_1)$ is equal to that required to compute $f_1(X_{11})$; the additional time required to compute $f_2(X_1)$ may be substantially less than that required to calculate $f_2(X_{21})$, depending on how much of the effort goes into the function evaluation as compared to the generation of the additional replicate X_{21}. We conclude that $E\tau_1 \leq E(\tau_{11} + \tau_{21})$. Thus, we obtain a guaranteed efficiency improvement if $\text{var}(f_1(X_1) - f_2(X_1)) \leq \text{var}(f_1(X_{11}) - f_2(X_{21}))$. Since $\text{var } f_i(X_{i1}) = \text{var}(f_i(X_i))$ $(i = 1, 2)$, it follows that an efficiency improvement occurs if $\text{cov}(f_1(X_1), f_2(X_2)) \geq 0$. In order that the r.v.'s $f_1(X_1), f_2(X_1)$ be positively correlated, it seems intuitively reasonable to require that the two systems respond in a similar fashion to the random input X_1, in the sense that $f_1(X_1)$ ought to be large (small) when $f_2(X_1)$ is large (small). There are two types of commonly encountered settings in which this occurs:

i) f_1, f_2 are both increasing (or both decreasing)

ii) f_1 is close to f_2.

Specifically, it is well known that if f_1 and f_2 are both increasing (or both decreasing), then $\text{cov}(f_1(X_1), f_2(X_1)) \geq 0$, providing mathematical justification for i). Also, if $f_1 \to f_2$ pointwise, then it typically follows that $\text{cov}(f_1(X_1), f_2(X_1)) \to \text{var } f_1(X_1) > 0$. This gives (asymptotic) sup-

port for ii).

Of course, most stochastic system comparisons involve (much) more complicated r.v.'s than those described above. However, the principles i) and ii) enunciated above remain valid more generally. A number of theoretical results are available that support these general conclusions. The assumptions typically involve some kind of stochastic monotonicity; see Heidelberger and Iglehart [26], Glynn and Iglehart [17], Glasserman [11], and Glasserman and Yao [15] for such results.

There are some significant practical difficulties that are commonly encountered in trying to apply the method of CRN's. The most fundamental is that it is non-trivial to set up the various random inputs to the alternative systems in such a way that the two systems will be guaranteed to respond to the inputs in a similar fashion. A typical recommendation is to generate all the random variables used in the simulation via inversion, since inversion preserves monotonicity in the underlying uniform r.v.'s. We note, however, that inversion is not a necessary ingredient to the method of CRN's, as our above discussion illustrates (X_i need not be generated by inversion). A second recommendation that is frequently made is to assign different uniform random number generators to each of the sources of random variation within the models, so as to "synchronize" the random inputs to the maximum extent possible; see Bratley, Fox, and Schrage [2] and Glasserman and Yao [15] for further details.

A more subtle problem can arise in applying the method of CRN's to systems comparisons involving steady-state performance measures associated with regenerative systems. As discussed in Glynn [16], applying CRN's to a pair of discrete state space Markov chains can lead to a situation in which the joint process has multiple closed communicating classes, as well as transient states. As a consequence, for a given initial state, there is no guarantee that the system will return to this state infinitely often. In fact, it may be difficult to identify any state as a regenerative state to which the joint process will return infinitely often. This complicates the application of the regenerative method to such simulations.

As indicated above, the method of CRN's leads to computational improvements when the two systems being compared are "close" to one another (see ii) above). One important practical setting in which this occurs is that of derivative estimation, in which the derivative is computed via

a finite difference. If common random numbers are applied to each of the simulations used to form the finite differences, substantial variance reductions can result. In fact, use of CRN's can even impact the rate of convergence as measured in the quantity γ that appears in (2.1); see Glynn [18] for details.

We conclude this section by noting that the efficiency improvements described above for estimating $\alpha \equiv \mu_1 - \mu_2$ also hold for "ratio comparisons" of the form $\alpha \equiv \mu_1/\mu_2$. The asymptotic variances of the ratio estimators that arise in this latter setting involve the same type of covariance as arises in the analysis of CRN's in the non-ratio context considered above.

A related efficiency improvement technique, called antithetics, is designed to take advantage of possible negative correlations that can be induced in certain settings while estimating performance measures associated with a single stochastic system; see Cheng [5] for a discussion.

5. Importance Sampling

The calculation of quantities associated with "rare events" is a computational challenge from a simulation viewpoint. It turns out that importance sampling is a particularly effective efficiency improvement technique for dealing with such problems.

In order to get a sense of the difficulties engendered by "rare event" simulation, consider the problem of estimating the probability of an improbable event. Let A be the event and let $\alpha \equiv P(A)$. The naive approach to estimating such a probability is to generate i.i.d. replicates $I_1, I_2, \ldots$ of the indicator r.v. $I(A)$ associated with the event. Let α_n be the proportion of the replicates for which the event A occurred, namely $\alpha_n \equiv n^{-1}(I_1+\ldots+I_n)$. The CLT asserts that

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

as $n \to \infty$. This suggests the approximation

$$\alpha_n \approx \alpha + \sqrt{\alpha(1-\alpha)/n}\, N(0,1) \qquad (5.1)$$

A rare event is one for which $\alpha = P(A)$ is small. The normally distributed error term in (5.1) goes to zero as the event A becomes successively rarer, so that the absolute error associated with this estimator improves as the event becomes rarer. However, in many rare event simulation settings, one is more interested in estimating the probability to a high level of precision from a relative view-

point. For example, one may need to know the order of magnitude of the probability. (Is it of the order of 10^{-6} or 10^{-9}?) Dividing both sides of (5.1) by α yields the approximation

$$\alpha_n / \alpha \approx 1 + \sqrt{(1-\alpha)/\alpha n}\, N(0,1) \tag{5.2}$$

so that the relative error of this naive estimator degrades significantly as the event gets rarer. This puts a premium on finding more efficient estimators to deal with this applications venue.

A principal difficulty with the above approach is that a very high proportion of the simulation time is spent sampling the uninteresting part of the sample space on which the rare event does not occur. One would ideally like to devote most of one's computational effort to that part of the sample space on which the rare event does occur. Importance sampling is ideally suited to accomplishing this.

Suppose that one's goal is to calculate $\alpha \equiv EX$, where X is a real-valued r.v. defined on a probability space (Ω, F, P). Then, α can be expressed as the integral

$$\alpha = \int_{\Omega} X(\omega) P(d\omega). \tag{5.3}$$

The idea behind importance sampling is that one may use an alternative distribution Q, say, rather than P to do the sampling, and thereby sample more in those regions of Ω which are computationally more important. To relate the outputs from the new simulations to those obtained under the original distribution P, suppose that we can find a r.v. L satisfying

$$X(\omega)L(\omega)Q(d\omega) = X(\omega)P(d\omega) \tag{5.4}$$

$\forall \omega \in \Omega$. One way to guarantee this is to require that $L(\omega)Q(d\omega) = P(d\omega)$, $\forall \omega \in \Omega$. In the language of measure-theoretic probability, this latter requirement amounts to demanding that P be absolutely continuous with respect to Q; the r.v. L is then called the likelihood ratio of P with respect to Q. In the case that $Q(d\omega) = q(\omega)Q^*(d\omega)$ for some density q, absolute continuity forces P to have a density $p(\cdot)$ such that p vanishes everywhere that q vanishes. The likelihood ratio L is then given by $L(\omega) = p(\omega)/q(\omega)$.

Of course, in general, all that is required is that there exist a r.v. L satisfying (5.4). The parameter α can now be re-expressed as

$$\alpha = E_Q(X \cdot L) \tag{5.5}$$

where $E_Q(\cdot)$ denotes the expectation operator taken relative to the distribution Q. The idea underly-

ing importance sampling is to now sample according to the probability distribution Q, and to average the r.v.'s $X_1 \cdot L_1$, $X_2 \cdot L_2$, ... in order to obtain the new estimator. Note that one possible choice for Q is

$$Q^*(d\omega) = |X(\omega)| \, P(d\omega)/\alpha^* \tag{5.6}$$

where α^* is the normalization factor given by $\alpha^* = \int_\Omega |X(\omega)| P(d\omega)$. With this choice, the r.v. L takes the form

$$L(\omega) = \alpha^* / |X(\omega)| \tag{5.7}$$

in which case $X(\omega) \cdot L(\omega) = \alpha^* I(X(\omega) > 0) - \alpha^* I(X(\omega) < 0)$. Observe that if X is a non-negative r.v., $X \cdot L$ is equal to α and is hence deterministic. It follows that, in this setting, Q^* reduces the variance to zero. Of course, this particular choice of Q is not typically implementable, since Q^* depends on α^*, which is effectively the unknown quantity that one is trying to simulate. Nevertheless, it suggests that, in practice, one should try to choose a Q that is roughly proportional to $|X(\omega)| \, p(\omega)$ (assuming that P has a density p). This approach has been used successfully, for example, to numerically compute the normalization constants that appear in the steady-state distributions of product-form networks; see Ross and Wang [32]. Furthermore, in the case that one is trying to calculate the probability of a rare event A, the zero-variance choice of Q is

$$Q^*(d\omega) = P(d\omega \mid A)$$

where $P(\cdot \mid A)$ is the conditional distribution of P given the occurrence of the rare event A. Thus, in rare event simulation, a key step in applying importance sampling is determining at least the approximate behavior of the system on that part of the sample space on which the rare event occurs. Fortunately, that part of probability theory known as "large deviations theory" is largely concerned with exactly this type of calculation. As we shall see later, large deviations plays a key role in rare event simulation for random walks and queues.

One important applications setting in which importance sampling has been successfully applied is that of highly reliable (or highly dependable) systems. This is a problem context in which it is fairly easy to discern intuitively what the conditional distribution looks like, at least approximately. The basic type of model that has been most widely studied is one in which the system under consideration consists of d components. Since these models are frequently used to analyze systems that are designed to be highly reliable, the failure rates for each of the components are

typically very small, particularly in comparison to the repair rates associated with any maintenance policy used to operate the system. The most important types of performance measures for such systems are the mean time to failure (MTTF), conditional on all components being initially operational, and the steady-state unavailability (the long-run proportion of time that the system is down). For both of these measures, conventional simulation can be extremely expensive, because of the substantial amounts of simulated time necessary to observe a sufficient number of failures so as to obtain a reasonable level of precision in the estimates.

Before proceeding any further, we wish to give a more concrete explanation of how importance sampling is typically implemented in the simulation of complex stochastic systems. Suppose that the objective is to estimate α on the basis of the simulation of an S-valued stochastic sequence $X = (X_n : n \geq 0)$. Here, we can take Ω as $\Omega = S^\infty$, the space of sequences with S-valued coordinates. Suppose that $\alpha = E(f_T(X_0, X_1, ..., X_T)\, I\,(T < \infty))$ where, for each $n \geq 0$, $f_n : S^n \rightarrow \Re$ is a given function specified by the simulator and T is a stopping time. (By a stopping time, we mean that for each $n \geq 0$, $I(T = n)$ is a function of $X_0, ..., X_n$ alone and, in particular, does not depend on any X_i with $i > n$.) Let P be the distribution associated with the expectation operator $E(\cdot)$ and let Q be our alternative importance sampling distribution, selected with the property that P is suitably absolutely continuous with respect to Q. In this context, it means that for each $n \geq 0$, there exists a r.v. L_n (necessarily a function of $X_0, ..., X_n$ alone) such that

$$L_n\,(x_0, ..., x_n)\; Q(X_0 \in dx_0, ..., X_n \in dx_n) = P(X_0 \in dx_0, ..., X_n \in dx_n).$$

Let $E_Q(\cdot)$ be the expectation operator corresponding to Q. It is straightforward to verify that

$$\alpha = E_Q(f_T(X_0, ..., X_T)\; I(T<\infty) L_T). \tag{5.8}$$

In most simulation schemes, X_{n+1} is generated from the conditional distribution associated with $X_0, ..., X_n$. Specifically, if the distribution underlying the X_i's is P (Q), then X_{n+1} is typically generated from $P_{n+1}(\cdot \mid X_0, ..., X_n\ )(Q_{n+1}(\cdot \mid X_0, ..., X_n\ ))$, where $P_{n+1}(\cdot \mid x_0, ..., x_n\ ) \equiv P(X_{n+1} \in\ \cdot \mid X_0 = x_0, ..., X_n\ = x_n)$ and $Q_{n+1}(\cdot \mid x_0, ..., x_n\ ) \equiv Q(X_{n+1} \in\ \cdot \mid X_0 = x_0, ..., X_n\ = x_n)$. Under the assumption that P is suitably absolutely continuous with respect to Q, it follows that

$$L_n = l_0(X_0) \cdot \prod_1^n l_i(X_i; X_0, \ldots, X_{i-1}) \tag{5.9}$$

where l_0 satisfies $l_0(x_0)\; Q(X_0 \in dx_0) = P(X_0 \in dx_0)$ and l_n satisfies $l_n\,(x_n; x_0, ..., x_{n-1})\; Q_n(dx_n \mid x_0,$

..., x_{n-1}) = $P_n(dx_n \mid x_0, ..., x_{n-1})$ for $n \geq 1$. This form of the likelihood ratio is particularly convenient to use for simulations of systems that are naturally specified in terms of their conditional distributions. Perhaps the nicest such example is that of a time-homogeneous discrete time Markov chain X taking values in a finite state space S and possessing transition matrix $P = (P(x, y) : x, y \in S)$. Assume that the distribution Q is also Markovian, so that $Q_n(\cdot \mid x_0, ..., x_{n-1}) = Q_n(\cdot \mid x_{n-1})$. In order to guarantee the appropriate absolute continuity, it is required that $Q_n(y \mid x)$ be positive whenever $P(x, y)$ is positive. In any case, L_n then takes the form

$$L_n = l_0(X_0) \cdot \prod_1^n P(X_{i-1}, X_i) / Q_i(X_i \mid X_{i-1}) \tag{5.10}$$

so that the likelihood ratio can then be easily updated recursively. When Q corresponds to the distribution of a time-homogeneous Markov chain having transition matrix $Q' = (Q'(x, y) : x, y \in S)$, the likelihood ratio simplifies even further, since then $Q_n(y \mid x) = Q'(x, y)$. In fact, this type of choice for Q is the most common form of importance sampling distribution for dealing with discrete-time Markov chains.

Returning now to the reliability context, most of the models simulated are either Markovian or can easily be made so by adding appropriate supplementary state variables. By applying regenerative arguments, it turns out that the critical quantity to calculate in such models is the probability that the system fails before it returns to the fully operational state, given that it starts with all components fully operational; see, for example, Goyal et al [23]. With this quantity in hand, both of the performance measures mentioned above can be easily calculated, with a modest amount of additonal simulation. As mentioned earlier, the ideal zero-variance importance sampling distribution for simulating this probability is to use the conditional distribution.

Consider an irreducible Markov chain $X = (X_n : n \geq 0)$ taking values in a finite state space S. Let A, B be two disjoint non-empty subsets of S and view B as representing the fully operational state and A as representing those states in which the system is failed. We wish to calculate the conditional distribution starting from a state $x \in S$, given that A is hit by the chain before B is hit. It is easily verified that the conditional distribution Q^* corresponds to a time-homogeneous Markov chain having transition probabilities given by $Q'(x, y) = P(X_1 = y \mid X_0 = x, A$ is hit before $B)$. Consequently, the types of sampling distributions Q that are typically used in the reliability

context are those corresponding to time-homogeneous Markov chains. Of course, more can be said about the conditional distribution for reliability models. For example, when the failure rates are very small relative to the repair rates, it is quite likely that the system will reach the failed state prior to a return to the fully operational state by following the path to failure that involves the smallest possible number of component failures. (Any additional component failure makes the likelihood of such a path that much smaller.) Hence, the conditional distribution accentuates the likelihood of component failures, with a particular emphasis on those components for which their failure leaves the system most vulnerable to going down. This sort of "failure biasing" approach has been used with success in dealing with a large class of such reliability simulations. For additional details, see Goyal et al [23], Shahabuddin [34], and Heidelberger [25].

We turn now to importance sampling for queueing systems. Since queues are effectively random walks with boundaries, a discussion of the relevant theory for random walks is necessary. This is a setting in which large deviations theory is a powerful guide for choosing a sampling distribution. Let $(S_n : n \geq 0)$ be a random walk, in which $S_0 = 0$ and $S_n = X_1 + X_2 + ... + X_n$, where the X_i's are i.i.d. real-valued r.v.'s. Let μ be the mean of X_1 and note that the law of large numbers implies that, for n large, $S_n \approx n\mu$. A rare event that is of some interest in the random walk setting is the calculation of the tail probability $\alpha \equiv P(S_n > na)$ where $a > \mu$. For n large, it seems reasonable to expect that, conditional on $S_n > na$, $S_n \approx na$. Large deviations theory can be used to justify this assertion. (This idea is implicit in the proof of, for example, the Gartner-Ellis theorem; see p.14-19 of Bucklew [3].) Furthermore, the theory can be used to identify the approximate form of the conditional distribution. Consider the family of distributions P_θ, $\theta \in \Re$ under which the sequence $(S_n : n \geq 0)$ continues to form a random walk, with common increment distribution given by

$$P_\theta(x_i \in dx) = \exp(\theta x - \psi(\theta))P(X_i \in dx) \qquad (5.11)$$

where $\psi(\cdot)$ is the cumulant generating function of the X_i's under P, namely $\psi(\theta) = \log E\exp(\theta X_1)$. It is easy to verify that $E_\theta X_i = \psi'(\theta)$. It follows that one can alter the drift of the random walk via a judicious choice of θ. In particular, one should give the random walk a drift a in order that $S_n \approx na$. This discussion suggests that one should simulate the random walk under distribution P_{θ_a}, where θ_a satisfies $\psi'(\theta_a) = a$, in order to estimate the rare event probability $\alpha =$

$P(S_n > na)$. The type of change of distribution specified by (5.11) is termed an "exponential twist". Large deviations theory proves that the exponential twist P_{θ_a} is in fact the "asymptotic" conditional distribution of the random walk, given that $S_n \approx na$; see Bucklew [3] for an accessible introduction to the general subject of large deviations. (Here, the asymptotic takes hold as $n \to \infty$.)

This theory holds for random walks with much more complicated increment structure than that described above. In particular, consider a Markov random walk in which the increments are derived from a Markov chain. Here, S_n continues to take the form $S_n = X_1 + ... + X_n$. However, we now assume that $X_i = f(Z_i)$ where $Z = (Z_n : n \geq 0)$ is an irreducible Markov chain, with finite state space Γ and transition matrix $K = (K(x, y) : x, y \in \Gamma)$. The key is to describe the analog to the "exponential twist" described above.

Let $K_\theta = (K_\theta(x, y) : x, y \in \Gamma)$ be the non-negative matrix with elements defined by $K_\theta(x, y) = \exp(\theta f(x))K(x, y)$. Perron-Frobenius theory for such matrices guarantees the existence of a positive eigenvalue $\lambda(\theta)$ and corresponding positive column eigenvector $h_\theta = (h_\theta(x) : x \in \Gamma)$ such that $K_\theta h_\theta = \lambda(\theta)h_\theta$. Then, the matrix $G_\theta = (G_\theta(x, y) : x, y \in \Gamma)$ with elements defined by $G_\theta(x, y) = K_\theta(x, y)\, h_\theta(y) / \lambda(\theta)h_\theta(x)$ is stochastic. Let P_θ be the distribution under which Z evolves as a Markov chain with transition matrix G_θ, and $(S_n : n \geq 0)$ continues to be defined as $S_n = fX_1 + ... + fX_n$. This class of distributions corresponds to the "exponentially twisted" distributions described earlier in the i.i.d. setting. In fact, if one sets $\psi(\theta) = \log(\lambda(\theta))$, it turns out that $\psi'(\theta)$ is the drift of $(S_n : n \geq 0)$ under P_θ, in the sense that $\psi'(\theta)$ is the steady-state mean of $f(Z_n : n \geq 0)$ computed under transition matrix G_θ.

Suppose that $a > \mu$ where μ is the steady-state mean of $f(Z_n : n \geq 0)$ computed under K. Then, $P(S_n > na)$ corresponds to the probability of a rare event when n is large. As in the i.i.d. setting, the asymptotic conditional distribution is given by P_{θ_a}, where θ_a satisfies $\psi'(\theta_a) = a$.

We now turn to the analysis of queues. The focus will be on single-station queues, largely because networks are not yet well understood. It is well known that much of the steady-state theory for single-station queues amounts to computing the maximum of an associated random walk. For example, the steady-state waiting time in the standard $GI/G/1$ queue has the same distribution as the maximum $M = \max(S_n : n \geq 0)$, where $S_n = X_1 + ... + X_n$ and $X_i = V_{i-1} - U_i$; here, the V_i's

are the successive service times of customers, and the U_i's are the successive interarrival times. In order that the queue have a well-defined steady-state, the "traffic intensity" of the queue must be strictly less than one. This translates into the associated random walk having negative drift. As a consequence, the computation of the tail probability $P(M > x)$ is a rare event calculation when x is large. Let $T(x)$ be the time at which the random walk first jumps above level x. The random walk theory already descibed asserts that if the random walk first attains level x at time t, then the conditional distribution looks approximately like P_{θ_t}, where θ_t satisfies $\psi(\theta_t) = x/t$. However, large deviations theory also provides a rough approximation of the probability that $S_t \approx x$. The most likely path along which the random walk's maximum is greater than x is then found by maximizing these probability approximations over t. It turns out that the optimizing t takes the form $t \approx \psi'(\theta^*)x$, where θ^* is the positive root of $\psi(\cdot)$. This shows that, in order to calculate the probability that the steady-state waiting time is greater than x (x large), one should generate the associated random walk from P_{θ^*}. A similar conclusion holds for Markov modulated queues, in which either the arrival epochs or the sum of the service times (or both) are Markov random walks. For additional details on importance sampling for random walks and queues, see Chang et al [4], Asmussen [1], Sadowsky [33], Glasserman and Kao [14], and Heidelberger [25]. For some discussion of how to compute the relevant likelihood ratios for general discrete-event simulations, see Glynn and Iglehart [20] and Glynn [19].

Before concluding, we note that one difficulty with importance sampling is that the method is invasive, in the sense that the entire probability dynamics of the simulation must be altered in order to apply the method. This is particularly problematic, since the appropriate sampling distribution can be quite sensitive to the choice of performance measure. Hence, if multiple performance measures are to be estimated from a single simulation, importance sampling may be inappropriate. In addition, the invasive character of the approach is undesirable in any setting in which visualization is deemed important, since the dynamics under the new sampling distribution may be quite different from those associated with the original system. Observing the dynamics of the altered system may then lead to incorrect conclusions about the original system.

6. Conditional Monte Carlo

An intuitively reasonable approach to producing estimators with better statistical characteristics is to try to eliminate as much randomness as possible by replacing random variables by their expectations. Of course, in general, naive replacement of input r.v.'s by their expected values leads to simulation algorithms that are invalid. However, conditional Monte Carlo provides a theoretical environment in which this replacement is not only valid but often desirable.

Suppose that one wishes to estimate $\alpha \equiv EZ$, where Z is a real-valued r.v. Assume that there exists a r.v. Y such that $g(y) \equiv E(Z \mid Y = y)$ can be calculated, either analytically or numerically. Clearly, $\alpha = Eg(Y)$ so there is now a choice between basing the estimation algorithm for α on Z or on $g(Y)$. This is a situation in which the framework of Section 2 can be helpful. As one might expect, the conditional expectation $g(Y)$ is less variable than Z, so $\mathrm{var}g(Y) \leq \mathrm{var}Z$. (This is a setting in which replacing a r.v. by its (conditional) expectation reduces variance.) On the other hand, the time required to compute $g(Y)$ may be substantially larger than that necessary to calculate X. So, despite the fact that this replacement is guaranteed to reduce variance, some care must be taken in order to guarantee an efficiency improvement.

The above idea works more generally. Suppose that there exists a process Y for which $E(Z \mid Y)$ can be calculated. Conditional Monte Carlo involves basing one's estimation algorithm on replications of the r.v. $g(Y) \equiv E(Z \mid Y)$; again, a variance reduction is guaranteed. There is one particularly important class of stochastic models in which a good conditioning process exists. Suppose that Z is a performance measure associated with the simulation of a semi-Markov process $X = (X(t) : t \geq 0)$. One of the key structural features of such a process is that the conditional distribution of the successive state transition epochs is very simple, given the sequence of states visited. In particular, given the sequence of states visited, the sequence of holding times $(\eta_i : i \geq 0)$ (the times spent in each of the states visited) form an independent sequence of r.v.'s, each r.v. having a distribution known explicitly to the simulator. To illustrate how this property can be exploited, let Y_i be the i'th state visited by X and T_i be the time of the transition from the i'th state. Then for a real-valued function f, the conditional expectation of a "cumulative cost" can easily be computed:

$$E(\int_0^{t_n} f(X(s))\, ds | Y_0, Y_1, \ldots) = \sum_0^n f(Y_i)\mu(Y_i, Y_{i+1})$$

where $\mu(x, y)$ is the mean amount of time spent in state x, given that the next state visited by X is y. Since the means $\mu(x, y)$ are easily computed from the basic "building block" data of the process, it is clear that this conditional expectation is typically cheaper to calculate than the cumulative cost itself, since (in particular) the holding time variates need not be generated. Thus, the conditional expectation is a two-way winner here, because both the variance and mean time to compute observations is reduced. Consequently, conditional Monte Carlo is guaranteed to provide an efficiency improvement in the simulation of cumulative costs for semi-Markov processes. Since cumulative costs arise naturally in steady-state estimation, this idea also has implications for the computation of steady-state quantities. Because only the discrete-time sequence $(Y_i : i \geq 0)$ need be simulated in order to calculate the conditional expectation, this method is referred to as "discrete-time" conversion in the literature; see Fox and Glynn [9] for details.

Continuous-time Markov chains form a special subclass of the class of semi-Markov processes, with some additional structure. Because of the fact that the state holding times are exponentially distributed, the principle of conditional Monte Carlo can then be applied to a number of performance measures that are significantly more complex than the cumulative cost decsribed above. See Fox and Glynn [9] for applications to the simulation of a variety of performance measures associated with finite-horizon continuous-time Markov chains.

The above discussion suggests that a great deal of structure must be present in the system being simulated, in order that conditional Monte Carlo be applicable. However, a related idea, called "extended conditional Monte Carlo", is applicable to a much greater class of processes. It is specially designed for performance measures that take the form of a cumulative cost and is easiest to understand when the underlying process is a discrete-time Markov chain $X = (X_n : n \geq 0)$. Given a real-valued function f defined on the state space of the chain, the cumulative cost then takes the form $\sum_1^n f(X_i)$. Because of the Markov structure, the conditional expectation $E(f(X_{i+1}) | X_0, ..., X_i) = E(f(X_{i+1}) | X_i)$ is typically easy to calculate. This suggests replacing the cumulative cost by

$$\sum_{1}^{n} E(f(X_i)|X_{i-1}).$$

Unfortunately, because of the fact that each summand in the above r.v. is conditioned on a different quantity, there is no longer any universal guarantee that one will obtain a variance reduction by basing one's estimation strategy on this. Instead, recent efforts have focussed on establishing sufficient conditions for a variance reduction to result from "extended conditional Monte Carlo"; see Glasserman [12, 13] and Glynn and Iglehart [17] for details.

7. Stratification

Stratification is an approach that permits one to take advantage of the fact that certain probability distributions are known. A good example of a problem setting in which this method can be used effectively is given in Bratley, Fox, and Schrage [2]. In their example, they consider a bank with multiple tellers. On any given day, the number of tellers Y that report for work is random, with known distribution. Let X be a given performance measure and note that $\alpha \equiv EX$ can be written in the form

$$\alpha = \sum_i E(X|Y = i)\ p_i$$

where p_i is the probability that i tellers report for work on a typical day. In a conventional simulation, one samples the r.v. Y according to the mass function $(p_i : i \geq 0)$. In the simulation of n days, let $k_i(n)$ be the number of days on which precisely i tellers report, and let $X_1, X_2, ..., X_n$ be the replicates of X observed over the n days. The conventional estimator for α is, of course, the sample mean of the X_i's given by $\alpha_n = n^{-1}(X_1 + ... + X_n)$. This estimator can be re-written in the form

$$\alpha_n = \sum_i \overline{X}_i(k_i(n))k_i(n) / n$$

where $\overline{X}_i(k)$ is the sample mean of k i.i.d. replicates of the r.v. X generated from its distribution conditional on $Y = i$. When n is large, $k_i(n)/n \approx p_i$ and this suggests that we replace the above estimator by

$$\alpha_p(n) = \sum_i \overline{X}_i(k_i(n))p_i.$$

The estimator $\alpha_p n$ is called a "post-stratified" estimator. This approach provides a guaranteed variance reduction relative to α_n; see Bratley, Fox, and Schrage [2] and Cochran [6] for details. Note that this method is non-invasive, in the sense that one can simulate the process in the conventional fashion and just adjust the final estimator by taking advantage of one's knowledge of the p_i's. (By invasive, we refer to methods that require that one modify the code that generates the simulation of sample paths of the stochastic system under study.)

On the other hand, the method known as stratification is invasive. Here, one pre-assigns the amount of sampling to each "stratum"; in the banking example, stratum i corresponds to that part of the sample space on which $Y = i$. In particular, one assigns an amount of sampling n_i proportional to the stratum's probability, so that n_i is given by the greatest integer less than or equal to $p_i \cdot n$. For each stratum i, one then generates n_i i.i.d. replicates of from its distribution conditional on $Y = i$., yielding a sample mean $\overline{X}_i(n_i)$. An implicit assumption here is that one can generate samples directly from the conditioned sample spaces associated with each stratum. This does not present a difficulty in the banking example, but can be problematic in general. Furthermore, because stratification modifies the natural dynamics of the simulation, it typically would not be suitable for simulations in which visualization of the sample paths is important. In any case, the stratified estimator then takes the form

$$\alpha_s(n) = \sum_i \overline{X}_i(n_i) p_i.$$

The asymptotic variance of this estimator, for n large, is identical to that of the post-stratified estimator. However, the stratified estimator $\alpha_s(n)$ has preferable small-sample behavior relative to $\alpha_p(n)$. On the other hand, post-stratification is non-invasive and hence there are applications settings in which it is preferable.

We note that importance sampling and stratification can often easily be combined. One first changes the sampling weights on each stratum via importance sampling, and then applies stratification (or post-stratification) to the modified probabilities. This idea implicitly underlies the derivation of the optimal sampling weights that have been developed for use in stratified sampling schemes; see Cochran [6].

Acknowledgement

This research was supported by the Army Research Office under Contract Number DAAL-03-91-G-0101.

References

[1] Asmussen, S. Busy period analysis, rare events, and transient behaviour in fluid models. To appear in J. of Appl. Math. and Stoch. Anal. (1993).

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

[3] Bucklew, J. *Large Deviation Techniques in Decision, Simulation, and Estimation.* J. Wiley and Sons, New York (1990).

[4] Chang, C.S., Heidelberger, P., Juneja, S., and Shahabuddin, P. Effective bandwidth and fast simulation of ATM intree networks. IBM Research Report RC 18586, Yorktown Heights, New York (1992). To appear in *Proc. of the Performance 1993 Conference*.

[5] Cheng, R.C.H. The use of antithetic variates in computer simulations. J. Oper. Res. Soc., 21 (1982) 229-237.

[6] Cochran, W.G. *Sampling Techniques*. John Wiley and Sons, New York (1977).

[7] Fishman, G.S. and Huang, B.D.. Antithetic variates revisited. Commun. ACM, 26 (1983) 964-971.

[8] Fishman, G.S. and Kulkarni, V.G. Improving Monte Carlo efficiency by increasing variance. Mgmt. Sci., 38 (1992) 1432-44.

[9] Fox, B.L. and Glynn, P.W. Discrete-time conversion for simulating semi-Markov processes. Oper. Res. Letters, 4 (1986) 49-53.

[10] Fox, B.L. and Glynn, P.W. Discrete-time conversion for simulating finite-horizon Markov processes. SIAM J. Appl. Math., 50 (1990) 1457-1473.

[11] Glasserman, P. Processes with associated increments. J. of Appl. Prob., 29 (1992) 313-333.

[12] Glasserman, P. Stochastic monotonicity and conditional Monte Carlo for likelihood ratios. Adv. in App. Prob., 25 (1993) 103-115.

[13] Glasserman, P. Filtered Monte Carlo. Math. of Oper. Res., 18 (1993) 610-634.

[14] Glasserman, P. and Kao, S.-G. Overflow probabilities in Jackson networks. *Proc. of the 32nd IEEE Conference on Decision and Control*, (1993) 3178-3182.

[15] Glasserman, P. and Yao, D. Some guidelines and guarantees for common random numbers. Mgmt. Sci., 38 (1992) 884-908.

[16] Glynn, P.W. Regenerative structure of Markov chains simulated via common random numbers. Oper. Res. Letters, 4 (1985) 49-53.

[17] Glynn, P.W. and Iglehart, D.L.. Simulation methods for queues: An overview. QUESTA, 3 (1988) 221-256.

[18] Glynn, P.W. Optimization of stochastic systems via simulation. *Proc. of the 1989 Winter Simulation Conference*, (1989) 90-105.

[19] Glynn, P.W.. A GSMP formalism for discrete-event systems. *Proc. of the IEEE*, 77 (1989) 14-23.

[20] Glynn, P.W. and Iglehart, D.L. Importance sampling for stochastic simulation. Mgmt. Sci., 35 (1989) 1297-1325.

[21] Glynn, P.W. and Whitt, W. Indirect estimation via $L=W$. Oper. Res., 37 (1989) 82-103.

[22] Glynn, P.W. and Whitt, W. Asymptotic efficiency of simulation estimators. Oper. Res., 40 (1992) 505-520.

[23] Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V.F., and Glynn, P.W. A unified framework for simulating Markovian models of highly reliable systems. IEEE Trans. on Computers, C-41 (1992) 36-51.

[24] Hammersley, J.M. and Handscomb, D.C. *Monte Carlo Methods.* Methuen, London (1964).

[25] Heidelberger, P. Fast simulation of rare events in queueing and reliability models. IBM Research Report RC 19028, Yorktown Heights, NY (1993).

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

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

[28] Law, A.M. Efficient estimators for simulated queueing systems. Mgmt. Sci., 22 (1975) 30-41.

[29] Nelson, B. An illustration of the sample space definition of simulation and variance reduction. Trans. of the Soc. for Comput. Simul., 2 (1985) 237-247.

[30] Nelson, B. Decomposition of some well-known variance reduction techniques. J. Statist. Comput. Simul., 23 (1986) 183-209.

[31] Ruppert, D. Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz procedures with dependent noise. Ann. Prob. 10 (1982) 178-187.

[32] Ross, K.W. and Wang, J. Monte Carlo summation applied to multichain queueing networks. *Proc. of the 30th IEEE Conference on Decision and Control* (1991) 883-4.

[33] Sadowsky, J.S. Large deviations theory and efficient simulation of excessive backlogs in a GI/G/m Queue. IEEE Trans. Automatic Control, 36(1991) 1383-1394.

[34] Shahabuddin, P. Importance sampling for the simulation of highly reliable Markovian systems. IBM Research Report RC 16729, Yorktown Heights, New York (1993). To appear in Mgmt. Sci.

[35] Wilson, J.R. Variance reduction techniques for digital simulation. Amer. J. Math. Mgmt. Sci. 4 (1984) 277-312.