On the Interaction Between Stratification and Control Variates, with Illustrations in a Call Center Simulation

Pierre L'Ecuyer and Eric Buist

GERAD, CIRRELT, and
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

Abstract

Variance reduction techniques (VRTs) are often essential to make simulation quick and accurate enough to be useful. A case in point is simulation-based optimization of complex systems. An obvious idea to push the improvement one step further is to combine several VRTs for a given simulation. But such combinations often give rise to new issues. This paper studies the combination of stratification with control variates. We detail and compare several ways of doing the combination. Nontrivial synergies between the two methods are exhibited. We illustrate this with a telephone call center simulation, where we combine a control variate with stratification with respect to one of the uniform random variates that drive the simulation. It turns out that using more information in the control variate degrades the performance (significantly) in our example. This seemingly paradoxical behavior is not rare and our theoretical analysis explains why.

Keywords: variance reduction, control variates, stratification, call centers

Introduction

The use of simulation to optimize decision parameters in complex stochastic systems is increasingly frequent. This simulation-based optimization typically requires thousands or millions of simulation runs for a complex model, where each run takes a significant amount of time. Consider

for instance a telephone call center for which we want to optimize the number of agents who talk with customers over the phone, and the working schedules of these agents, under constraints on the quality of service and on admissible schedules. Large call centers are complex stochastic systems that can be analyzed realistically only by simulation; tractable queueing models oversimplify reality and are not very reliable. When simulation is combined with an optimization algorithm, simulation speed is a key issue because optimization often requires huge numbers of simulation runs at different parameter settings (Atlason et al., 2004; Ceşik and L'Ecuyer, 2006). In that context, straightforward (or *naive*) Monte Carlo simulation is often too slow to be practical.

Fortunately, proper use of variance reduction techniques (VRTs) such as control variates, stratification, conditional Monte Carlo, common random numbers, importance sampling, etc., can improve simulation efficiency, sometimes by a large factor (Bratley et al., 1987; Fishman, 1996; Glynn, 1994). For larger improvements, an obvious idea is to use two or more VRTs at the same time. However, this often complicates things in an unexpected way. Such combinations are studied in Cheng (1986), Booth and Pederson (1992), Avramidis and Wilson (1996), and Hickernell et al. (2005), for example, in specific settings.

The aim of this paper is to examine some issues that arise when combining two specific VRTs and to show how to handle these issues. We do this via an example of a call center simulation, to make things more concrete for the reader, but our development applies more generally. We study the combination of control variates with stratification with respect to a continuous input variable. In this case, the optimal control variate coefficient turns out to be a function of the input variable on which we stratify. We focus on how to approximate this function in practice.

The next section discusses how stratification with respect to uniform random numbers driving the simulation can be used to reduce the variance. We then study the combination of a control variate with stratification, which is non-standard and requires some care. Then, we give an example of a simple call center on which perform numerical experiments to compare the various ways of making the combination. The simulations were made with *ContactCenters*, a specialized simulation tool for contact centers (Buist and L'Ecuyer, 2005) developed in Java with the SSJ library (L'Ecuyer and Buist, 2005). A preliminary version of this paper was presented at the 2006 *Winter Simulation Conference* (L'Ecuyer and Buist, 2006).

Stratification

Stratified sampling consists in partitioning the set of possible outcomes in a finite number of strata, estimating the quantity of interest separately in each stratum, and computing a weighted average of these estimators, where the weights are the (known) probabilities of the corresponding strata, to obtain the overall estimator. This is easy to implement *if* we can design strata for which we know the exact probabilities and from which we know how to generate samples uniformly. Bratley et al. (1987, page 295) give an example with three strata. For large and complex simulations, it may not be obvious a priori how to achieve this. One way of stratifying a simulation is as follows.

Recall that all the randomness in a simulation typically comes from a sequence of independent $\mathrm{U}(0,1)$ (uniform over the interval (0,1)) random variates. Select d of those uniforms, preferably some whose values are deemed to have a large impact on the result. Partition the d-dimensional unit hypercube $[0,1)^d$ into k rectangular boxes of the same shape and size; these boxes will correspond to the k strata. Each one has probability $1/k$. To generate a sample uniformly from stratum s, we generate a point $\mathbf{U}$ uniformly in box s and take the d coordinates of $\mathbf{U}$ as the values of the d selected uniforms. All other random variates in the simulation are generated as usual, independently of the realizations of the d selected uniforms.

Suppose each simulation run provides an estimator X for $\mu = \mathbb{E}[X]$. Suppose also that we have n_s observations in stratum s for each s, where the n_s's are positive integers such that $n = n_1 + \cdots + n_k$. If $X_{s,1}, \ldots, X_{s,n_s}$ denote the n_s i.i.d. copies of X in stratum s, the (unbiased) *stratified estimator* of μ is (Cochran, 1977):

$$\bar{X}_{\mathrm{s},n} = \frac{1}{k}\sum_{s=1}^{k} \hat{\mu}_s \qquad \text{where} \qquad \hat{\mu}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} X_{s,i} \tag{1}$$

is the sample mean in stratum s. Let $\sigma_s^2 = \mathrm{Var}[X \mid S = s]$, the conditional variance of X given that we are in stratum s. Then,

$$\mathrm{Var}[\bar{X}_{\mathrm{s},n}] = \frac{1}{k^2}\sum_{s=1}^{k} \sigma_s^2/n_s \tag{2}$$

and an unbiased estimator of this variance is

$$S^2_{\mathrm{s},n} = \frac{1}{k^2}\sum_{s=1}^{k} \hat{\sigma}_s^2/n_s, \tag{3}$$

where $\hat{\sigma}_s^2$ is the sample variance of $X_{s,1},\ldots,X_{s,n_s}$, assuming that $n_s \geq 2$.

Stratification with *proportional allocation* takes $n_s = n/k$ for all s. Then, (2) simplifies to

$$\mathrm{Var}[\bar{X}_{\mathrm{sp},n}] = \frac{1}{nk}\sum_{s=1}^{k}\sigma_s^2 \tag{4}$$

where $\bar{X}_{\mathrm{sp},n}$ denotes the corresponding version of (1). The *optimal allocation*, which minimizes the variance (2) with respect to $n_1,\ldots,n_k$ under the constraints that $n_s > 0$ for each s and $n_1+\cdots+n_k = n$ for a given n, is easily found by using a Lagrange multiplier; we must take n_s proportional to σ_s/k: $n_s^* = n\sigma_s/\bar{\sigma}k$ where $\bar{\sigma} = \sum_{s=1}^{k}\sigma_s/k$. (We neglect the rounding of n_s^* to an integer and assume that $n_s \geq 2$.) If $\bar{X}_{\mathrm{so},n}$ denotes the estimator with optimal allocation, we have $\mathrm{Var}[\bar{X}_{\mathrm{so},n}] = \bar{\sigma}^2/n$. Putting these pieces together, the variance can be decomposed as follows (Cochran, 1977):

$$\begin{aligned}
\mathrm{Var}[\bar{X}_n] &= \mathrm{Var}[\bar{X}_{\mathrm{sp},n}] + \frac{1}{nk}\sum_{s=1}^{k}(\mu_s-\mu)^2 &(5)\\
&= \mathrm{Var}[\bar{X}_{\mathrm{so},n}] + \frac{1}{nk}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2 + \frac{1}{nk}\sum_{s=1}^{k}(\mu_s-\mu)^2. &(6)
\end{aligned}$$

The first sum in the last line represents the variability due to the different standard deviations among strata and the second sum represents the variability due to the differences between stratum means. Proportional allocation eliminates the last sum while optimal allocation also eliminates the first. For a given total sample size n, a larger k generally gives more variance reduction, because the strata are smaller so there is less variability within the strata. When $k \to \infty$, we have $\bar{\sigma} \to \int_{[0,1)^d}\sigma(\mathbf{u})d\mathbf{u}$, where $\sigma^2(\mathbf{u}) = \mathrm{Var}[X \mid \mathbf{U} = \mathbf{u}]$. Usually, $\bar{\sigma} > 0$, in which case the marginal variance reduction converges to zero. On the other hand, with a larger value of n/k (a smaller k), we have a more accurate estimator of the variance of the stratified estimator.

Combining with a Control Variate

Control variates (CVs) for simulation are discussed, e.g., by (Lavenberg and Welch, 1981; Glynn and Szechtman, 2002). Here we study how to combine a CV with stratification. To keep the notation simple, we now assume that $d = 1$ and we consider a single control variable, but all our development can be generalized easily to $d > 1$ and to a vector of control variates. The one-

dimensional uniform random vector $\mathbf{U}$ is denoted by U. Any random variable A whose expectation $a = \mathbb{E}[A]$ is known can be used as a CV. Preferably, A should be strongly correlated (positively or negatively) with X. Without the stratification, the CV is used by subtracting from the original estimator X the difference $A - \mathbb{E}[A]$ multiplied by some *constant* coefficient β. The (unbiased) CV estimator is:

$$X_{\mathrm{c}} = X - \beta[A - a].$$

The optimal coefficient β is

$$\beta^* = \mathrm{Cov}[A,X]/\mathrm{Var}[A] \tag{7}$$

and we have $\mathrm{Var}[X_{\mathrm{c}}] = (1-\rho^2[X,A])\mathrm{Var}[X]$ when $\beta = \beta^*$, where $\rho[X,A]$ is the linear correlation between X and A. This β^* can be estimated from preliminary (pilot) simulation runs or from the same runs as X; in the latter case, this gives a slightly biased estimator, but the bias is negligible when the number n of runs is large.

Things become somewhat more complicated if we combine the CV with stratification, because both β^* and the expected value of A generally depend on the strata, or on the value taken by the random variate on which we stratify. We examine and compare various ways of handling this, assuming that we are stratifying on U as in the previous subsection. We can apply the CV on the stratified average $\bar{X}_{\mathrm{s},n}$, or on each stratum average $\hat{\mu}_s$, or on the individual observations $X_{s,i}$. All these methods are equivalent to replacing $X_{s,i}$ by

$$X_{\mathrm{sc},s,i} = X_{s,i} - b_{s,i}(A_{s,i} - e_{s,i}), \tag{8}$$

with different choices of $b_{s,i}$ and $e_{s,i}$, where $A_{s,i}$ is the value of the control variate for the observation $X_{s,i}$, in stratum s. Let $a_s = \mathbb{E}[A_{s,i}]$, the expected value of A given that we are in stratum s, and $a(u) = \mathbb{E}[A \mid U = u]$, the expected value of A conditional on $U = u$. In (8), when $U = u_{s,i}$, we can take $e_{s,i}$ as either a, a_s or $a(u)$. We can also take $b_{s,i}$ as either a common constant β, or a different constant β_s in each stratum s, or a function of u, $\beta(u)$. We examine and compare these possibilities.

If $b_{s,i}$ does not depend on more information that $e_{s,i}$, then (8) is unbiased; otherwise it can be biased. So if $e_{s,i} = a_s$, we cannot take $b_{s,i} = \beta(U)$, whereas if $e_{s,i} = a$, we must have $b_{s,i} = \beta$ (a constant). To show unbiasedness, we take the conditional expectation given the stratum s if $e_{s,i} = a_s$ and given U if $e_{s,i} = a(U)$. For example, if $e_{s,i} = a(U)$ and $b_{s,i} = \beta_s$, then $\mathbb{E}[\beta_s(A_{s,i} - a(U))] =$

$\mathbb{E}[\mathbb{E}[\beta_s(A_{s,i} - a(U)) \mid U]] = \mathbb{E}[\beta_s \mathbb{E}[(A_{s,i} - a(U)) \mid U]] = 0$, but this no longer works if we take a_s together with $\beta(U)$.

Table 1 summarizes the different combinations. Each table entry gives the "correction term" $b_{s,i}(A_{s,i} - e_{s,i})$ used in (8) for the given combination. The dashed entries correspond to biased estimators. On each row, the best estimator is the one on the diagonal. As we shall see later, none of these three diagonal entries always gives a smaller variance than the other two, even if we use the optimal CV coefficient in each case. We will examine each of them in more detail.

Table 1: The different possibilities for $b_{s,i}(A_{s,i} - e_{s,i})$

	β	β_s	$\beta(U)$
a	$\beta(A_{s,i} - a)$	—	—
a_s	$\beta(A_{s,i} - a_s)$	$\beta_s(A_{s,i} - a_s)$	—
$a(U)$	$\beta(A_{s,i} - a(U))$	$\beta_s(A_{s,i} - a(U))$	$\beta(U)(A_{s,i} - a(U))$

A common coefficient β, with $e_{s,i} = a_s$. We define $\bar{A}_{\mathrm{s},n}$ as the weighted average of the n replicates of A, in the same way as $\bar{X}_{\mathrm{s},n}$ in (1):

$$\bar{A}_{\mathrm{s},n} = \frac{1}{k} \sum_{s=1}^{k} \frac{1}{n_s} \sum_{i=1}^{n_s} A_{s,i}.$$

Then, $\mathbb{E}[\bar{A}_{\mathrm{s},n}] = (a_1 + \cdots + a_k)/k = a$. Using $\bar{A}_{\mathrm{s},n}$ as a CV with a single coefficient β and $e_{s,i} = a_s$ gives the estimator

$$\bar{X}_{\mathrm{sc},n} = \bar{X}_{\mathrm{s},n} - \beta(\bar{A}_{\mathrm{s},n} - \mathbb{E}[\bar{A}_{\mathrm{s},n}]) = \bar{X}_{\mathrm{s},n} - \beta(\bar{A}_{\mathrm{s},n} - a). \tag{9}$$

For the choice of β, a first (naive) approach is to use the β^* defined earlier, as if there was no stratification. However, this β^* is no longer optimal, as we now show.

The estimator (9) has variance

$$\mathrm{Var}[\bar{X}_{\mathrm{sc},n}] = \mathrm{Var}[\bar{X}_{\mathrm{s},n}] + \beta^2 \mathrm{Var}[\bar{A}_{\mathrm{s},n}] - 2\beta \mathrm{Cov}[\bar{X}_{\mathrm{s},n}, \bar{A}_{\mathrm{s},n}]. \tag{10}$$

Differentiating with respect to β and equaling the derivative to zero, we find that the variance is

minimized by taking

$$\beta = \beta^*_{\rm sc} = \frac{\mathrm{Cov}[\bar{X}_{{\rm s},n}, \bar{A}_{{\rm s},n}]}{\mathrm{Var}[\bar{A}_{{\rm s},n}]} = \frac{\sum_{s=1}^k \mathrm{Cov}[X_{s,i}, A_{s,i}]/n_s}{\sum_{s=1}^k \mathrm{Var}[A_{s,i}]/n_s}. \tag{11}$$

Here, $\mathrm{Cov}[X_{s,i}, A_{s,i}]$ and $\mathrm{Var}[A_{s,i}]$ are the conditional covariance and variance given that we are in stratum s. Our second combined estimator uses (9) with $\beta = \beta^*_{\rm sc}$. This $\beta^*_{\rm sc}$ generally differs from β^* and it also depends on the allocation used for the stratification. As a result, minimizing $\mathrm{Var}[\bar{X}_{{\rm sc},n}]$ requires finding $\beta^*_{\rm sc}$ and the optimal allocation (the n_s's) *simultaneously* (which is not necessarily easy). If we restrict ourselves to proportional allocation, the n_s's simplify and we obtain

$$\beta^*_{\rm sc} = \beta^*_{\rm scp} = \frac{\sum_{s=1}^k \mathrm{Cov}[X_{s,i}, A_{s,i}]}{\sum_{s=1}^k \mathrm{Var}[A_{s,i}]}.$$

Taking $b_{s,i} = \beta_s$ with $e_{s,i} = a_s$. We now consider a different CV coefficient β_s in each stratum. We replace $X_{s,i}$ by $X_{{\rm sc},s,i} = X_{s,i} - \beta_s(A_{s,i} - a_s)$ for each s and i, so the average $\hat{\mu}_s$ is replaced by

$$\hat{\mu}_{{\rm sc},s} = \hat{\mu}_s - \beta_s(\hat{A}_s - a_s). \tag{12}$$

where $\hat{A}_s$ is the average of the $A_{s,i}$'s in stratum s. We assume that we can compute

$$a_s = k \int_{(s-1)/k}^{s/k} a(u)du \tag{13}$$

with negligible error (e.g., by numerical integration). The variance in stratum s becomes

$$\sigma^2_{{\rm sc},s} = \sigma^2_s + \beta_s^2 \mathrm{Var}[A_{s,i}] - 2\beta_s \mathrm{Cov}[X_{s,i}, A_{s,i}]. \tag{14}$$

and the optimal β_s for stratum s is

$$\beta^*_{{\rm sc},s} = \frac{\mathrm{Cov}[X_{s,i}, A_{s,i}]}{\mathrm{Var}[A_{s,i}]}. \tag{15}$$

With $\beta_s = \beta^*_{{\rm sc},s}$, the variance in stratum s is reduced to

$$\mathrm{Var}[X_{{\rm sc},s,i}] = \left(1 - \rho^2[X_{s,i}, (A_{s,i} - a_s)]\right) \mathrm{Var}[X_{s,i}] = \mathrm{Var}[X_{s,i}] - \rho^2[X_{s,i}, (A_{s,i} - a_s)]\mathrm{Var}[X_{s,i}]. \tag{16}$$

The overall variance here with $\beta^*_{\mathrm{sc},s}$ cannot be larger (and is usually smaller) than if we impose $\beta_s = \beta$ for all s, because we have the flexibility to optimize the constant β_s in each stratum. The difference can be large if the $\beta^*_{\mathrm{sc},s}$ are far from being equal between strata. After estimating the variance with $\beta^*_{\mathrm{sc},s}$ within each stratum, we can find the allocation that minimizes the overall controlled variance. Note that the optimal allocation with the CV, obtained using the $\sigma_{\mathrm{sc},s}$'s, differs from the optimal allocation without CV, obtained with the σ_s's defined earlier.

Taking $e_{s,i} = a(u)$. Once we know $u_{s,i}$, the realization of U for the observation $X_{s,i}$, we can take $e_{s,i} = a(u_{s,i})$ instead of a_s. The CV coefficient $b_{s,i}$ can be any of the three possibilities: β, β_s, or $\beta(u)$. Clearly, the more flexibility we have, the better we can do, so an optimal choice of $\beta(u)$ (a function of u) is always at least as good as an optimal choice of β_s (a function of s), and the latter is always at least as good as an optimal β (a single constant). Note that the optimal values of these coefficients are not the same (in general) with $e_{s,i} = a(u)$ than with $e_{s,i} = a_s$.

Let $C(u) = A - a(u)$. Suppose the CV coefficient can be a function of U, $b_{s,i} = \beta(U)$. Let $\sigma^2(u) = \mathrm{Var}[X \mid U = u]$, $X_{\mathrm{sc}}(u) = X - \beta(u)C(u)$ (the controlled estimator conditional on $U = u$),

$$\sigma^2_{\mathrm{sc}}(u) = \mathrm{Var}[X_{\mathrm{sc}}(u)] = \sigma^2(u) + \beta^2(u)\mathrm{Var}[A \mid U = u] - 2\beta(u)\mathrm{Cov}[X, A \mid U = u]$$

(its conditional variance), $\mu(u) = \mathbb{E}[X \mid U = u]$, $\mu_s = \mathbb{E}[X_{s,i}]$, and let $U_{s,i}$ denote a random variable uniformly distributed over $[(s-1)/k, s/k)$. The variance of the controlled estimator in stratum s is

$$\begin{aligned}
\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] &= \mathbb{E}[\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i}) \mid U_{s,i}]] + \mathrm{Var}[\mathbb{E}[X_{\mathrm{sc}}(U_{s,i}) \mid U_{s,i}]] \\
&= \mathbb{E}[\sigma^2_{\mathrm{sc}}(U_{s,i})] + \mathrm{Var}[\mu(U_{s,i})] \\
&= k\int_{(s-1)/k}^{s/k} \sigma^2_{\mathrm{sc}}(u)du + k\int_{(s-1)/k}^{s/k} (\mu(u) - \mu_s)^2 du. \qquad (17)
\end{aligned}$$

The choice of $\beta(u)$ affects only the first term in (17), i.e., the expectation of the conditional variance. The optimal allocation takes n_s proportional to $(\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})])^{1/2}$. Regardless of the allocation, the variance of the CV estimator is minimized by taking $\beta(u) = \beta^*_{\mathrm{sc}}(u)$, where

$$\beta^*_{\mathrm{sc}}(u) = \frac{\mathrm{Cov}[C(u), X \mid U = u]}{\mathrm{Var}[C(u) \mid U = u]} = \frac{E[C(u) \cdot X \mid U = u]}{E[C^2(u) \mid U = u]}. \qquad (18)$$

With this optimal coefficient, the variance in stratum s is reduced to

$$\begin{aligned}
\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] &= \mathbb{E}[\sigma^2_{\mathrm{sc}}(U_{s,i})] + \mathrm{Var}[\mu(U_{s,i})] \\
&= \mathbb{E}\left[\left(1-\rho^2[X_{s,i},\,(A_{s,i}-a(U_{s,i}))\mid U_{s,i}]\right)\,\sigma^2(U_{s,i})\right] + \mathrm{Var}[\mu(U_{s,i})] \\
&= \mathrm{Var}[X_{s,i}] - \mathbb{E}\left[\rho^2[X_{s,i},\,(A_{s,i}-a(U_{s,i}))\mid U_{s,i}]\,\sigma^2(U_{s,i})\right]. \qquad (19)
\end{aligned}$$

If we impose the additional constraint that $\beta(u)$ must be a constant β_s within each stratum, we have $\sigma^2_{\mathrm{sc}}(u) = \sigma^2_{\mathrm{sc},s}(u) = \mathrm{Var}[X - \beta_s C(u) \mid U = u]$ and the optimal β_s for stratum s is

$$\tilde{\beta}^*_{\mathrm{scu},s} = \frac{\mathrm{Cov}[C(U_{s,i}),X_{s,i}]}{\mathrm{Var}[C(U_{s,i})]} = \frac{E[C(U_{s,i})\cdot X_{s,i}]}{E[C^2(U_{s,i})]}. \qquad (20)$$

Here the CV estimator is unbiased and the last equality holds because $\mathbb{E}[C(U_{s,i})] = 0$. Obviously, with this additional constraint, we cannot get a smaller variance than with $\beta^*_{\mathrm{sc}}(u)$. And by imposing $\beta_s = \beta$ for all s, we can only do worse.

In practice, the function $\beta^*_{\mathrm{sc}}(u)$ can be approximated by approximating the two functions $q_1(u) = E[C(u)\cdot X \mid U = u]$ and $q_2(u) = E[C^2(u) \mid U = u]$. These functions can be estimated from a sample $\{(U_i,C_i,X_i),\, i = 1,\ldots,n\}$ of n realizations of $(U,C(U),X)$, and fitting a curve $\hat{q}_1$ to the points $(U_i,C_i(U_i)X_i)$ and another curve $\hat{q}_2$ to the points $(U_i,C_i^2(U_i))$. For example, we can fit a polynomial by interpolation or by least squares, or use a smoothing spline (de Boor, 1978).

To determine the optimal allocation, we need a good approximation of $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})]$ for each s. This requires approximations of the functions $\sigma^2_{\mathrm{sc}}(u)$, $\mu(u)$, and μ_s. Since $\mu = \int_0^1 \mu(u)du = \sum_{s=1}^k \mu_s/k$, this demands more information than estimating μ. A possible shortcut might be to just use the variance estimates and the optimal allocation for the case where the CV coefficient is constant in each stratum. In practice, this should rarely introduce a significant error, especially when k is large.

Is $e_{s,i} = a(u)$ always better than $e_{s,i} = a_s$? For $e_{s,i} = a(u)$, we have an ordering between β^*, β^*_s and $\beta^*(u)$ in terms of variance reduction; we know that more flexibility in the choice of CV coefficient can only decrease the variance. But is $a(u)$ with $\beta^*(u)$ always better than a_s with β^*_s? At first sight, one might think yes, because $a(u)$ exploits more information than a_s ($C_s = A - a_s$ is the conditional expectation of $C(U) = A - a(U)$ given that we are in stratum s). But on closer

examination, we find that using $a(u)$ might sometimes do worse! The following counterexample, suggested by Roberto Szechtman (private communication), shows that with the optimal CV coefficients, $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})]$ can be either larger or smaller than $\mathrm{Var}[X_{\mathrm{sc},s,i}]$.

Example 1 Suppose $X = A$. With $\beta_s = \beta^*_{\mathrm{sc},s} = 1$, we get $X_{\mathrm{sc},s,i} = a_s$, so $\mathrm{Var}[X_{\mathrm{sc},s,i}] = 0$. On the other hand, $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] = \mathrm{Var}[X - \beta(u)C(u)]$ and we also have $\beta^*_{\mathrm{sc}}(u) = 1$. With this coefficient, we have $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] = \mathrm{Var}[a(U_{s,i})] > 0$ whenever $\mathbb{E}[X \mid U] = \mathbb{E}[A \mid U]$ is not a constant inside stratum s. In this situation, $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] > \mathrm{Var}[X_{\mathrm{sc},s,i}]$. The larger the variation of $\mathbb{E}[X \mid U]$ inside the stratum, the larger the variance of the second CV estimator. So the second estimator has a larger variance when there are fewer strata. When the number of strata increases to infinity, the variance of the second estimator converges to zero, which makes sense because the two estimators are identical in the limit.

For an example where $\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] < \mathrm{Var}[X_{\mathrm{sc},s,i}]$, take $X = A - a(U)$. Then, $\beta^*_{\mathrm{sc}}(u) = 1$ and $X_{\mathrm{sc}}(U_{s,i}) = 0$, which has zero variance, whereas $\mathrm{Var}[X_{\mathrm{sc},s,i}] > 0$. □

To compare (14) with (17) in general, with the stratum-dependent CV and coefficient, the variance in stratum s is

$$\begin{aligned}\sigma^2_{\mathrm{sc},s} &= \mathbb{E}[\mathrm{Var}[X - \beta_s(A - a_s) \mid U_{s,i}]] + \mathrm{Var}[\mathbb{E}[X - \beta_s(A - a_s) \mid U_{s,i}]] \\ &= \mathbb{E}[\mathrm{Var}[X - \beta_s(A - a(U_{s,i})) \mid U_{s,i}]] + \mathrm{Var}[\mu(U_{s,i}) - \beta_s(a(U_{s,i}) - a_s)]. \qquad (21)\end{aligned}$$

With $e_{s,i} = a(u)$ and the optimal coefficient function $\beta^*_{\mathrm{sc}}(u)$, the variance in stratum s is

$$\mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})] = \mathbb{E}[\sigma^2_{\mathrm{sc}}(U_{s,i})] + \mathrm{Var}[\mu(U_{s,i})]. \qquad (22)$$

The estimator with $a(U)$ has a smaller variance than the one with a_s in stratum s if and only if (22) is smaller than (21). Comparing the corresponding terms of (21) and (22), we always have

$$\mathbb{E}[\mathrm{Var}[X - \beta_s(A - a(U_{s,i})) \mid U_{s,i}]] \geq \mathbb{E}[\mathrm{Var}[X - \beta^*_{\mathrm{sc}}(U_{s,i})(A - a(U_{s,i})) \mid U_{s,i}]] = \mathbb{E}[\sigma^2_{\mathrm{sc}}(U_{s,i})],$$

but we *may* have $\mathrm{Var}[\mu(U_{s,i}) - \beta_s(a(U_{s,i}) - a_s)] \leq \mathrm{Var}[\mu(U_{s,i})]$. In our numerical example later in this paper, it turns out that $\sigma^2_{\mathrm{sc},s} < \mathrm{Var}[X_{\mathrm{sc}}(U_{s,i})]$ for this reason. In fact, by comparing (16) and (19),

we see that taking $a(U)$ gives a smaller variance than a_s in stratum s if and only if $A_{s,i} - a(U_{s,i})$ is more strongly correlated with $X_{s,i}$ than $A_{s,i} - a_s$.

From a practical viewpoint, it is easier to estimate the constants $\beta^*_{\mathrm{sc},s}$ for a few strata than fitting a continuous function $\beta^*_{\mathrm{sc}}(u)$. And sometimes, it even gives a smaller variance. This will be illustrated in our numerical examples. On the other hand, when the number of strata is large, fitting the function might be easier than estimating the numerous constants $\beta^*_{\mathrm{sc},s}$. In the limit when the number of strata goes to infinity, the two schemes converge to each other.

Practical Issues

We summarize the required steps to implement the combined methods discussed thus far, focusing on the case where $e_{s,i} = a_s$ and $b_{s,i} = \beta^*_{\mathrm{sc},s}$. The other schemes are obtained via easy adaptations. In each case, there are actually many ways of implementing the procedure; some require pilot runs (e.g., to estimate the optimal allocation in the stratification, and to estimate the optimal CV coefficient independently of the production runs) and there is also more than one way of doing the pilot runs. In the preceding analysis, we took a one-dimensional uniform U and a single CV, but our development extends directly to d-dimensional vectors of uniforms and to vectors of CVs. If $d > 1$, s becomes the index of a d-dimensional box, and the integrals in (13) and (17) are over this box instead of over the interval $[(s-1)/k, s/k]$. If the CV is a vector, then β is also a vector, the covariances become matrices and vectors, and the correlation in (16) and (16) is replaced by a coefficient of determination between $X_{s,i}$ and the CV vector (Glynn and Szechtman, 2002).

The combined variance reduction method can be applied as follows.

1. Select d and define the d-dimensional boxes on which to stratify. Most often, d would not exceed 1 or 2. When $d > 1$, the boxes can be narrower in the dimension(s) deemed more important.

2. (Optional) Perform pilot runs to estimate the optimal allocation and optimal CV coefficients. See the discussion below.

3. Perform the n_s simulation runs in stratum s, for each s. With proportional allocation, $n_s = n/k$ for each s. Estimate each CV coefficient $\beta^*_{\mathrm{sc},s}$ from these runs if this was not done via pilot runs.

4. Compute the combined estimator. The variance within each stratum can be estimated in the standard way for CV estimators (Glynn and Szechtman, 2002). The overall variance is simply a weighted average of the variances within strata, given by (2). This variance estimate can be used to compute a confidence interval for the mean.

If we decide to skip the pilot runs in step 2, we can simply use proportional allocation for the stratification, and estimate the optimal CV coefficients using data from the production runs of step 3. This would introduce a bias, especially if we do this estimation separately for each stratum and if the n_s's are small. In the latter case, the estimators of $\beta^*_{\mathrm{sc},s}$ will also be noisy. When there are many strata, a good idea is to approximate $\mathrm{Cov}[X_{s,i}, A_{s,i} \mid U_{s,i} = u]$ and $\mathrm{Var}[A_{s,i} \mid U_{s,i} = u]$ by smooth functions of u, as discussed earlier with the functions $q_1(u)$ and $q_2(u)$. and then integrate these approximations over each box to obtain estimates of the two terms $\mathrm{Cov}[X_{s,i}, A_{s,i}]$ and $\mathrm{Var}[A_{s,i}]$ in (15). These smooth approximations can be obtained by least-squares fitting, for example.

The advantage of performing pilot runs in step 2 is to give an unbiased estimator. These pilot runs are simulation runs that are independent from those in step 3. They are used only to estimate the variances and covariances that determine the optimal allocation and optimal CV coefficients. This can be achieved via smooth approximating functions of u, as we just discussed. For a given total computing budget, skipping the pilot runs and using the entire budget for step 3 usually provides a smaller mean square error, despite the small bias.

How should we choose the uniforms on which we stratify, in practice? The idea is to pick one or two uniforms that have a large impact on the overall variance. We want to make the last term in (5) as large as possible. Our case study in the next section will give an illustration. As another example, suppose that our estimator is a function of the sample path of a Brownian motion $\{B(t), t \geq 0\}$ over a given time interval $[0, T]$. Then we may use one uniform to directly generate $B(T)$, a second uniform to generate $B(T/2)$ conditionally on $(B(0), B(T))$, and then generate the rest of the path conditionally on these three values. L'Ecuyer and Lemieux (2000) explain how to do that. We can stratify on these two uniforms, perhaps using narrower intervals for the first one. These two uniforms already provide a rough sketch of the sample path, and they typically account for a significant fraction of the variance (see L'Ecuyer and Lemieux (2000) for further details on this).

The choice of A can be guided by the examination of (16) and (19): we want to maximize the

squared correlations in these expressions. Interestingly, one referee suggested that it might be a good idea to stratify on the CV itself (or the uniform used to generate it). But this choice always gives zero variance reduction in (19), because $a(U_{s,i}) = A_{s,i}$ in that case! The correlation in (16) is also likely to be small. What we should look for instead is a CV (scalar or vector) that is highly correlated with X, *conditional on* U. Intuitively, this CV should bring information relevant to the prediction of X in addition to what is already known from U.

A Simple Model of a Call Center

The Model. Telephone call centers, and more generally *contact centers* where mail, fax, e-mail, and Internet contacts are handled in addition to telephone calls, are important components of large organizations (Gans et al., 2003). To illustrate the VRT ideas in this paper, we consider a simple model of a call center where agents answer incoming calls. Real-life call centers often receive different call types and have separate groups of agents with different combinations of *skills* that enable them to handle only a subset of the call types. To simplify the presentation, we assume a single agent type and a single call type, but the model is otherwise inspired by real-life centers. The techniques examined in this paper should behave in a similar way with more complex centers and other similar types of queueing systems.

Each day, the center operates for m hours. The number of (identical) agents answering calls and the arrival rate of calls vary during the day; we assume that they are constant within each hour of operation but depend on the hour. Let n_j be the number of agents in the center during hour j, for $j = 0, \ldots, m-1$. If more than n_{j+1} agents are busy at the end of hour j, calls in progress are completed but new calls are answered only when there are fewer than n_{j+1} agents busy. After the center closes, ongoing calls are completed and calls already in the queue are answered, but no additional incoming call is taken.

The calls arrive according to a Poisson process with piecewise constant rate, equal to $R_j = B\lambda_j$ during hour j, where the λ_j are constants and B is a random variable with mean 1 that represents the *busyness factor* of the day. We suppose that B has the *gamma* distribution with parameters (α_0, α_0), i.e., with mean $\mathbb{E}[B] = 1$ and $\mathrm{Var}[B] = 1/\alpha_0$. This type of arrival process model is motivated and studied by Whitt (1999) and Avramidis et al. (2004).

Incoming calls form a FIFO queue for the agents. A call *abandons* (and is lost) when its

waiting time exceeds its *patience time*. The patience times of calls are assumed to be i.i.d. random variables with the following distribution: with probability p the patience time is 0 (so the person hangs up if no agent is available immediately), and with probability $1-p$ it is exponential with mean $1/\nu$. The *service times* are i.i.d. *gamma* random variables with parameters (α,γ), i.e., with mean α/γ and variance α/γ^2.

For a given time period (an hour, a day, a month, etc.) and a given threshold s_0, the fraction of calls arriving during that period and whose waiting time is less than s_0 seconds (including those who abandoned before s_0 seconds) is called the *service level* for that period, whereas the fraction of calls having abandoned is called the *abandonment ratio*. The service level is widely used as a measure of quality of service in call centers. For certain types of call centers that provide public service, it is regulated by law: The call center operators may be charged a large fine if their service level goes below a given target; for example, 0.80 over each month for $s_0 = 20$ seconds.

Here, we estimate these performance measures over an infinite time-horizon, i.e., on average over an infinite number of days. Let A be the number of arriving calls during the day, $G(s_0)$ the number of those calls waiting less than s_0 seconds (including those who abandoned before s_0 seconds) for a given threshold s_0, and L the number of calls having abandoned. The expected number of arrivals during the day is $a = \mathbb{E}[A] = \sum_{j=0}^{m-1} \lambda_j$. Its variance is $\mathrm{Var}[A] = \mathrm{Var}[\mathbb{E}[A \mid B = b]] + \mathbb{E}[\mathrm{Var}[A \mid B = b]] = a + a^2/\alpha_0$. Define $g(s_0) = \mathbb{E}[G(s_0)]/a$ and $\ell = \mathbb{E}[L]/a$. These two quantities represent the steady-state service level, and abandonment ratio, respectively. Since a is known, here we will estimate only $\mathbb{E}[G(s_0)]$ and $\mathbb{E}[L]$.

We simulate the model for n days. For each day i, let A_i be the number of arrivals, $G_i(s_0)$ the number of calls who waited less than s_0 seconds and L_i the number of calls having abandoned. In what follows, we use X_i to represent either $G_i(s_0)$ or L_i, and $\mu = \mathbb{E}[X]$ to represent any of the two performance measures. A standard (or crude) unbiased Monte Carlo estimator of μ is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

with variance $\mathrm{Var}[\bar{X}_n] = \mathrm{Var}[X_i]/n$. We can estimate $\mathrm{Var}[X_i]$ by the empirical variance and a confidence interval can be computed as usual, using the normal approximation.

For our numerical illustrations, we take the following parameter values, where the time is measured in seconds: $\alpha_0 = 10$, $p = 0.1$, $\nu = 0.001$, $\alpha = 1.0$, $\gamma = 0.01$ (so the mean service time

is 100 seconds), and $s_0 = 20$. The center starts empty and operates for 13 one-hour periods. The number of agents and the arrival rate in each period are given in Table 2.

Table 2: Number of Agents n_j and Arrival Rate λ_j (per hour) for 13 one-hour Periods in the Call Center

j	0	1	2	3	4	5	6	7	8	9	10	11	12
n_j	4	6	8	8	8	7	8	8	6	6	4	4	4
λ_j	100	150	150	180	200	150	150	150	120	100	80	70	60

We stratify on the uniform random variate U used to generate the busyness factor B by inversion: $B = F_{\mathrm{B}}^{-1}(U)$. As a CV, we use the number A of arrivals during the day. The mean and variance of A are $a = \mathbb{E}[A] = 1660$ and $\mathrm{Var}[A] = 1660 + 1660^2/10 = 277\,220$. The expected number of arrivals conditional on $U = u$ is $a(u) = aF_{\mathrm{B}}^{-1}(u)$ and the expected number of arrivals given that we are in stratum s is

$$a_s = \mathbb{E}[A_{s,i}] = k\int_{(s-1)/k}^{s/k} a(u)du = ak\int_{(s-1)/k}^{s/k} F_{\mathrm{B}}^{-1}(u)du = ak\int_{F_{\mathrm{B}}^{-1}((s-1)/k)}^{F_{\mathrm{B}}^{-1}(s/k)} bf_{\mathrm{B}}(b)db$$

where $f_{\mathrm{B}}(b)$ is the density of B.

Variance Estimates for Different Schemes We perform a numerical experiment whose aim is to provide accurate estimates of all the terms in the variance decomposition (6) and other relevant constants, for each scheme. Instead of estimating these terms by the empirical variance of a few pilot runs, as we would normally do in an application, we did the following extensive (and more accurate) computations. We simulated 10^4 replications at $U = u_j = (j + 0.5)/1000$, for $j = 0,\ldots,999$. For each value of u_j, we computed the busyness factor $B_j = F_{\mathrm{B}}^{-1}(u_j)$, performed the runs, and then computed estimates of $\mu(u_j)$, $\sigma^2(u_j)$, and $\mathrm{Cov}[X,A \mid U = u_j]$ based on the 10^4 runs, for each j. We then fitted a cubic smoothing spline to these pointsto obtain accurate approximations of the functions $\mu(u)$, $\sigma^2(u)$, $\mathrm{Cov}[X,A \mid U = u]$, and $\beta^*_{\mathrm{sc}}(u)$. Note that $\mathrm{Var}[A \mid U]$ can be computed exactly, since A has a known Poisson distribution conditional on U.

We integrated these functions numerically to approximate μ_s, σ_s^2, $\mathrm{Var}[A_{s,i}]$, $\mathrm{Cov}[X_{s,i},A_{s,i}]$, $\beta^*_{\mathrm{sc},s}$, and β^*_{scp}, for each relevant pair (s,k), as well as all other relevant constants such as μ, σ^2, $\mathrm{Cov}[X,A]$, and β^*. The value of $\mathrm{Var}[A] = \mathbb{E}[\mathrm{Var}[A \mid U]] + \mathrm{Var}[\mathbb{E}[A \mid U]]$ is known, but $\mathrm{Var}[A \mid S = s]$, the conditional variance of A given that we are in stratum s, must be estimated too. Based on these

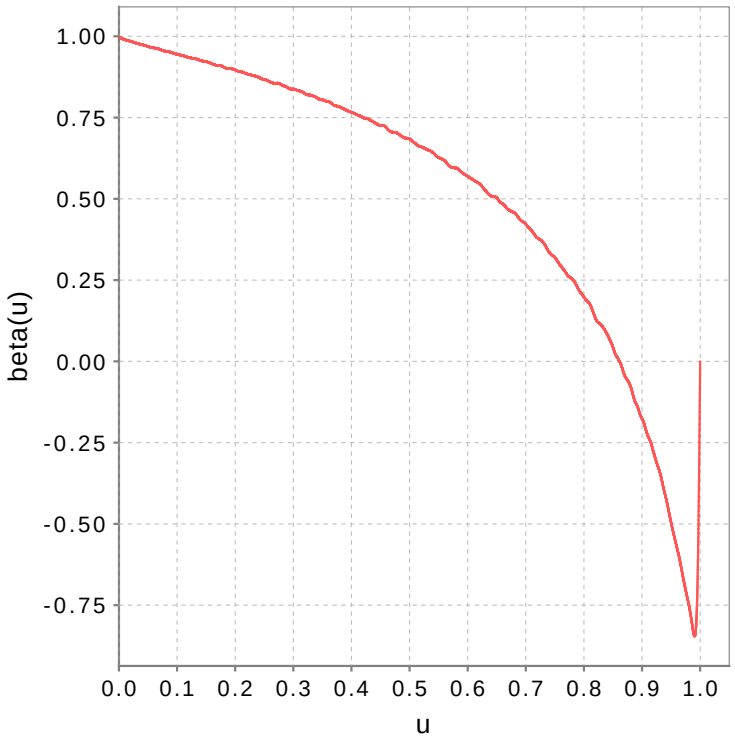

Figure 1: The Function $\beta^*_{\text{sc}}(u)$ for the Number of Calls Waiting less than s_0, Approximated by Smoothing Cubic Splines on 1000 Points

computations, we were able to compute all the numbers reported in Table 3 for the service level $G(s_0)$ and in Table 5 for L, the number of lost calls. We also have $\mu \approx 1418.660$ for $G(s_0)$, and $\mu \approx 60.504$ for L.

Figure 1 shows the behavior of the optimal CV coefficient $\beta^*_{\text{sc}}(u)$ for the estimation of $\mathbb{E}[G(s_0)]$, as a function of u. This coefficient is decreasing in u and has the same sign as $\text{Cov}[C, X \mid U = u]$. It is positive for small u and negative for large u. This can be explained as follows: when u is small, the load on the system is small and the agents are not very busy, so a small increase in the number of arrivals tends to increase $G(s_0)$, which makes the covariance positive. When u is large, on the other hand, the agents are occupied most of the time, so a few more arrivals increases the waiting time of several calls and tends to decrease the number of calls answered within s_0 seconds, whence a negative covariance. Therefore, the CV must correct the estimator in a different direction depending on the value of u. When $u \approx 0.99$, the load on the system is so high that new arrivals do not affect much the number of calls waiting less than s_0; in the limit, this number is zero (a constant) so $\text{Cov}[X, C \mid U = u] \to 0$ while $\text{Var}[C \mid U = u] = F_{\text{B}}^{-1}(u) \to \infty$. Thus, $\beta^*_{\text{sc}}(u)$ converges to 0 when $u \to 1$.

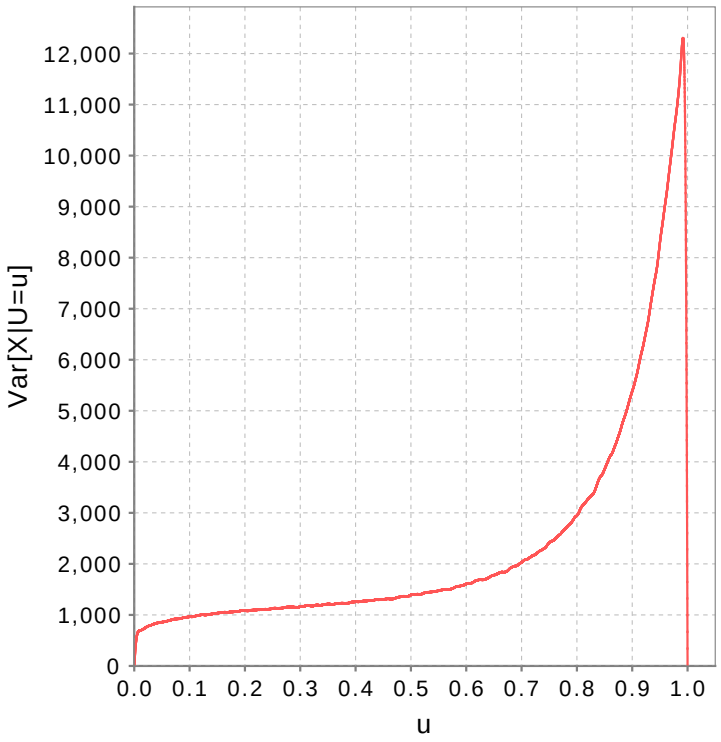

Figure 2: The Function $\sigma^2(u)$ for the Number of Calls Waiting less than s_0, Approximated by Smoothing Cubic Splines on 1000 Points

Figure 2 displays the function $\sigma^2(u) = \mathrm{Var}[X \mid U = u]$ as a function of u, again for the estimation of $X = G(s_0)$. When u increases (i.e., the arrival rate increases), the conditional variance first increases until it hits a sharp peak at $u \approx 0.99$, and then it decreases abruptly to zero. This decrease to zero is again due to the fact that when the arrival rate is too high (when u is too close to 1), practically no call is served within the time limit s_0. The graph of $\mu(u) = \mathbb{E}[G(s_0) \mid U = u]$ as a function of u, in Figure 3, confirms this abrupt convergence of $\mu(u)$ to 0 when $u \to 1$. This function increases for u up to about 0.86, and then it starts to decrease.

We made a similar experiment for the estimation of $\mathbb{E}[L]$, the expected number of lost calls. Figure 4 shows the behavior of $\beta^*_{\mathrm{sc}}(u)$, which is always positive and increasing in this case (the correlation between L and A, conditional on u, is always positive). The functions $\sigma^2(u)$, $\mu(u)$, and $\mathrm{Cov}[X, A \mid U = u]$, have a very similar shape as $\beta^*_{\mathrm{sc}}(u)$.

Tables 3 and 5 report the values of the different terms in the variance decomposition, as a function of the number of strata, k. The following five schemes are considered; they all use the estimator in (8), with some control variate $C = A_{s,i} - e_{s,i}$ and coefficient $b_{s,i}$:

(1) no CV, only stratification ($b_{s,i} = 0$);

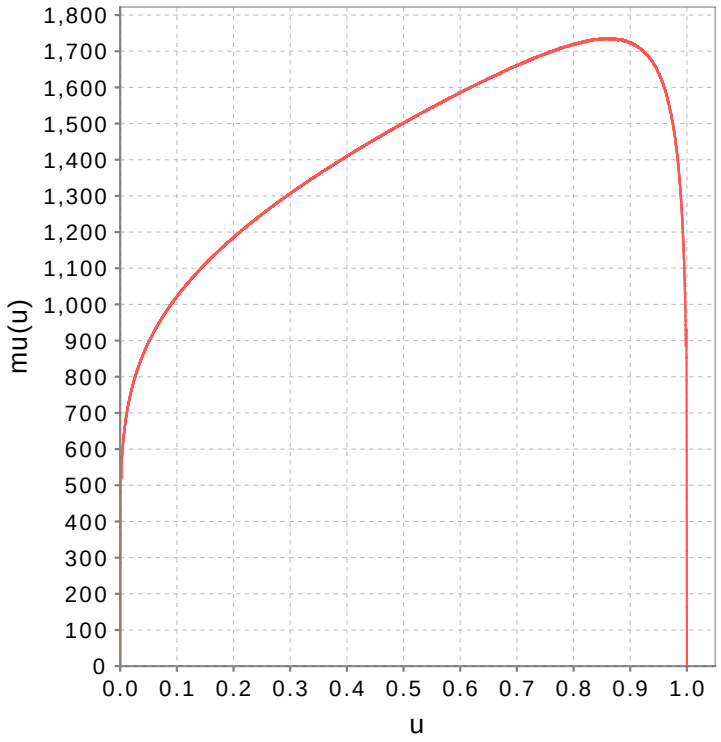

Figure 3: The Function $\mu(u)$ for the Number of Calls Waiting less than s_0, Approximated by Smoothing Cubic Splines on 1000 Points

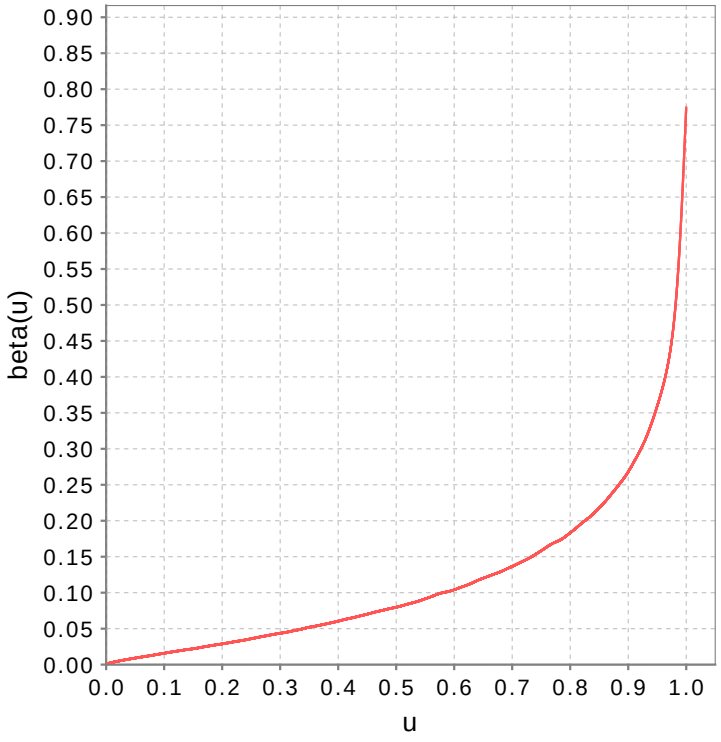

Figure 4: The Function $\beta^*_{sc}(u)$ for the Number of Lost Calls, Approximated by Smoothing Cubic Splines on 1000 Points

Table 3: Terms of the variance decomposition for $G(s_0)$ with k strata, for various estimation schemes

Scheme	k	1	2	3	10	20	50	100	500	1000	∞
	$\frac{1}{k}\sum_{s=1}^{k}(\mu_s-\mu)^2$	0	44610	55448	68600	71803	73868	74514	74928	74986	75096
(1)	$n\mathrm{Var}[\bar{X}_n]$	77444	77444	77444	77444	77444	77444	77444	77444	77444	77444
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	77444	32834	21995	8844	5641	3575	2930	2515	2457	2347
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	30933	20476	5778	3537	2516	2247	2071	2046	2010
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	1901	1519	3066	2103	1060	682	444	412	337
(2)	$n\mathrm{Var}[\bar{X}_n]$	35291	77358	83389	82180	79979	78302	77719	77296	77245	77188
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	35291	32748	27941	13580	8177	4433	3205	2368	2259	2092
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	30280	19580	5297	3103	2042	1749	1546	1517	1480
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	2468	8360	8283	5073	2391	1456	822	742	612
(3)	$n\mathrm{Var}[\bar{X}_n]$	35291	73348	77097	76783	77004	77401	77435	77281	77241	77188
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	35291	28738	21649	8183	5201	3533	2921	2353	2255	2092
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	28734	19971	6460	4102	2722	2150	1633	1562	1479
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	4	1678	1724	1099	811	771	720	694	614
	β^*_{sc}	0.390	0.196	0.074	-0.210	-0.243	-0.108	0.057	0.299	0.342	0.392
(4)	$n\mathrm{Var}[\bar{X}_n]$	35291	53078	59966	70427	73534	75595	76240	76644	76692	76786
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	35291	8468	4517	1827	1732	1727	1726	1715	1706	1690
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	5470	2606	1180	1124	1098	1090	1074	1069	1060
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	2999	1912	647	608	629	636	641	637	630
(5)	$n\mathrm{Var}[\bar{X}_n]$	76786	76786	76786	76786	76786	76786	76786	76786	76786	76786
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	76786	32176	21337	8186	4983	2917	2272	1857	1799	1690
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	30293	19762	4873	2599	1565	1293	1114	1090	1060
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	1883	1576	3313	2384	1352	979	744	709	630

(2) the CV $C = A - a_s$ with constant coefficient $b_{s,i} = \beta^*$;

(3) the CV $C = A - a_s$ with constant coefficient $b_{s,i} = \beta^*_{\mathrm{sc}}$;

(4) the CV $C = A - a_s$ with coefficient $b_{s,i} = \beta^*_{\mathrm{sc},s}$ in each stratum;

(5) the CV $C = A - a(U)$ with coefficient $b_{s,i} = \beta^*_{\mathrm{sc}}(U)$.

The values of k range from 1 to 1000. The extreme case of $k = 1$ corresponds to no stratification. We also consider the limit when $k \to \infty$. The (almost) exact means can be computed via $\mathbb{E}[X] = \int_0^1 \mu(u)du$; they are $\mathbb{E}[G(s_0)] \approx 1419$ and $\mathbb{E}[L] \approx 60.5$ (recall that the mean number of arrivals is 1660).

Note that Scheme (1) with $k = 1$ is the classical Monte Carlo estimator. Schemes (2) to (4) with $k = 1$ are all equivalent and they correspond to using CV only, without stratification.

When $k \to \infty$, the variance with proportional allocation becomes $\mathbb{E}[\mathrm{Var}[X \mid U]]$, while the variance with optimal allocation converges to $\left(\int_0^1 \sigma(u)du\right)^2$, which provides a lower bound on the best that can be achieved by increasing k and using optimal allocation. The variance of the means

and the variance of standard deviations across strata converge to $\mathrm{Var}[\mu(U)]$ and $\mathrm{Var}[\sigma^2(U)]$, respectively. The constant β^*_{scp} for scheme (3) converges to the ratio $\mathbb{E}[\mathrm{Cov}[X,A \mid U]]/\mathbb{E}[\mathrm{Var}[A \mid U]]$. The two CV schemes (4) and (5) are equivalent in the limit, as discussed earlier.

The variation of the means across strata, $(1/k)\sum_{s=1}^{k}(\mu_s - \mu)^2$, depends only on k and not on the CV scheme; it is given in the second line of the table. It increases with k, first very quickly and then slowly. This term represents the variance that is eliminated by doing stratification with proportional allocation, compared with no stratification at all; see Equation (5).

The term $(1/k)\sum_{s=1}^{k}(\sigma_s - \bar{\sigma})^2$, which represents the gain of optimal allocation over proportional allocation, usually first increases with k for k up to 2 to 10 (depending on the scheme), and then decreases with k. One exception to this is Scheme (4). The decrease is important for some of the schemes (e.g., (1), (2), (5)) and less important for others. This decrease could be explained intuitively by the fact that the number of strata increases, the variances within the strata (the σ_s's) tend to get smaller and so their variation decreases.

The variance of the stratified estimators decreases with the number of strata for all the schemes and both types of allocations (proportional and optimal).

With Scheme (1) (no-CV), the stratification with proportional allocation reduces the variance per run from 77246 to 8616 with $k = 10$, and to 2354 when $k \to \infty$ (in the limit). With optimal allocation, it is reduced further to 5701 with $k = 10$ and to 2013 when $k \to \infty$. Thus, we gain by a factor of more than 38.

With schemes (2) and (3), the CV brings practically no additional gain to stratification with proportional allocation. For Scheme (2), it even increases the variance when k is less than about 100. This is explained by the fact that β^* (whose value is 0.390 here) is not really optimal for this scheme. The optimal coefficient β^*_{scp} (given in the table) depends on k and it becomes close to β^* (but not equal) when $k \to \infty$. With the optimal allocation, the CV gives some gain when k is large. But it also increases the variance when k is small in Scheme (3); this is because the coefficient β^*_{scp} that we use is optimal only for the proportional allocation. Without stratification ($k = 1$), these schemes reduce the variance by a factor of 2.

Scheme (4) gives the best results, with both proportional and optimal allocations. The performance is also good even for small values of k, which is quite interesting: there is no need to use a large number of strata (at least for this particular example). For this scheme, each coefficient $\beta^*_{\mathrm{sc},s}$ is optimized to reduce σ_s independently across strata, and the CV works on both components

Table 4: Terms of the Decomposition (21) and (22) for Each Stratum, for $G(s_0)$ with $k = 20$ Strata

		$\mathbb{E}[\mathrm{Var}[X_{\mathrm{sc},s,i} \mid U_{s,i}]]$			$\mathrm{Var}[\mathbb{E}[X_{\mathrm{sc},s,i} \mid U_{s,i}]]$	
s	μ_s	Scheme (1)	Scheme (5)	Scheme (4)	Scheme (5)	Scheme (4)
1	766	743	21	22	12250	369
2	963	915	31	31	1314	0.021
3	1069	998	54	54	670	0.012
4	1149	1060	85	85	449	0.019
5	1218	1101	133	133	339	0.019
6	1278	1144	183	183	276	0.023
7	1334	1186	251	251	234	0.031
8	1385	1228	335	335	207	0.030
9	1433	1281	434	434	185	0.043
10	1479	1339	565	566	167	0.034
11	1523	1429	721	721	155	0.074
12	1565	1530	921	921	140	0.104
13	1605	1684	1175	1176	125	0.099
14	1643	1889	1496	1497	112	0.168
15	1677	2199	1934	1936	87	0.341
16	1707	2653	2515	2517	57	0.517
17	1728	3367	3330	3334	20	0.834
18	1732	4547	4528	4537	10	1.909
19	1692	6647	6373	6393	555	8.679
20	1427	10009	8706	8750	48508	543

of $\sigma_s^2 = \mathbb{E}[\mathrm{Var}[X \mid U_{s,i}]] + \mathrm{Var}[\mathbb{E}[X \mid U_{s,i}]]$. The σ_s's tend to be smaller and their variation is also smaller.

Scheme (5), in which the CV and its coefficient are functions of U, is not doing better than Scheme (4). When $k \to \infty$ the two schemes become equivalent, so there is not much difference between them when k is large. But for small k, Scheme (5) gives a much larger variance with both the proportional and optimal allocations. We found this result rather surprising at first. However, it can be explained by the fact that second term of (22) is much larger than that of (21) in some strata, in this example, when k is small. The last two columns of Table 4 compare these two terms, which represent the values of $\mathrm{Var}[\mathbb{E}[X_{\mathrm{sc},s,i} \mid U_{s,i}]]$ for the two schemes, in each stratum, for $k = 20$. We see that the values are much larger for Scheme (5) than for Scheme (4), especially for the strata where U is close to 0 or 1. The term $\mathbb{E}[\mathrm{Var}[X_{\mathrm{sc},s,i} \mid U_{s,i}]]$ is smaller for Scheme (5) than for Scheme (4), but only by a very small amount.

The results for L, given in Table 5, are similar. In particular, Scheme (4) (clearly) remains the best performer, especially for small or moderate k. Some minor changes are that here, Scheme (2) does not increase the variance compared with Scheme (1), and the variation of the standard deviations across strata is a decreasing function of k for all the schemes.

Table 5: Terms of the variance decomposition for L with k strata, for various estimation schemes

Scheme	k	1	2	3	10	20	50	100	500	1000	∞
	$\frac{1}{k}\sum_{s=1}^{k}(\mu_s-\mu)^2$	0	2458	3743	6552	7538	8330	8662	8965	8996	9007
(1)	$n\mathrm{Var}[\bar{X}_n]$	9192	9192	9192	9192	9192	9192	9192	9192	9192	9192
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	9192	6735	5449	2640	1654	862.5	530.2	227.5	195.9	185.7
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	3880	2337	579.1	307.2	172.2	132.7	103.7	100.7	99.36
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	2855	3111	2061	1347	690.4	397.5	123.8	95.2	86.3
(2)	$n\mathrm{Var}[\bar{X}_n]$	2681	4964	5985	7925	8484	8872	9015	9135	9148	9153
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	2681	2506	2242	1373	945.5	542.1	353.3	170.2	151.6	146.5
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	2231	1474	421.0	242.2	147.9	118.8	96.8	94.6	93.7
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	275.3	767.5	952.1	703.3	394.2	234.4	73.4	57.0	52.8
(3)	$n\mathrm{Var}[\bar{X}_n]$	2681	4667	5483	7331	8074	8709	8958	9134	9148	9153
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	2681	2209	1739	779.1	536.0	379.7	296.3	168.9	151.5	146.5
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	2198	1579	586.4	391.9	256.6	181.4	103.5	96.1	93.8
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	11.5	160.8	192.7	144.2	123.1	114.9	65.4	55.4	52.7
	β^*_{sc}	0.153	0.206	0.242	0.352	0.388	0.365	0.303	0.180	0.160	0.154
(4)	$n\mathrm{Var}[\bar{X}_n]$	2681	3365	4312	6747	7681	8456	8787	9085	9112	9120
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	2681	906.9	568.6	195.0	142.6	126.0	125.1	120.6	115.8	113.4
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	556.4	285.0	93.7	76.2	70.9	70.2	69.0	68.3	68.0
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	350.5	283.6	101.2	66.4	55.1	54.9	51.6	47.4	45.4
(5)	$n\mathrm{Var}[\bar{X}_n]$	9120	9120	9120	9120	9120	9120	9120	9120	9120	9120
	$n\mathrm{Var}[\bar{X}_{\mathrm{sp},n}]$	9120	6662	5377	2568	1582	790.2	457.8	155.1	123.5	113.4
	$n\mathrm{Var}[\bar{X}_{\mathrm{so},n}]$	—	3832	2293	536.7	268.0	137.1	99.6	72.3	69.4	68.0
	$\frac{1}{k}\sum_{s=1}^{k}(\sigma_s-\bar{\sigma})^2$	0	2831	3084	2031	1314	653.1	358.2	82.8	54.1	45.4

Conclusion

We have studied how to combine stratification with respect to a few uniform random numbers that drive the simulation, with one or more CVs. Our variance analysis and empirical results have exhibited some unexpected behavior in the combination. Among the different combination schemes that we have discussed, based on our analysis and experimentation, we recommend Scheme (4), with a moderate value of k. If we prefer a large k, then the optimal CV coefficients should probably be estimated by approximating the variances and covariances by smooth functions of u, via least-squares. In our empirical experiments with other examples, this scheme never performed much worse (and usually better) than the other schemes. Our detailed example provides insight by showing how the different variance and covariance components vary as functions of design parameters such as the number of strata, as functions of the uniform on which we stratify, and with the combination scheme. It also provides ideas on how to implement the method in practice.

Acknowledgments

This research has been supported by Grants OGP-0110050 and CRDPJ-320308 from NSERC-Canada, a Canada Research Chair, and a grant from Bell Canada via the Bell University Laboratories, to the first author. The second author benefited from an Industrial Scholarship from NSERC-Canada and the Bell University Laboratories. The paper was written while the first author was at IRISA, in Rennes, France.

References

Atlason, J., Epelman, M. A., and Henderson, S. G. (2004). Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127:333–358.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

Avramidis, A. N. and Wilson, J. R. (1996). Integrated variance reduction strategies for simulation. *Operations Research*, 44:327–346.

Booth, T. E. and Pederson, S. P. (1992). Unbiased combinations of nonanalog Monte Carlo techniques and fair games. *Nuclear Science and Engineering*, 110:254–261.

Bratley, P., Fox, B. L., and Schrage, L. E. (1987). *A Guide to Simulation*. Springer-Verlag, New York, NY, second edition.

Buist, E. and L'Ecuyer, P. (2005). A Java library for simulating contact centers. In *Proceedings of the 2005 Winter Simulation Conference*, pages 556–565. IEEE Press.

Ceẓik, M. T. and L'Ecuyer, P. (2006). Staffing multiskill call centers via linear programming and simulation. *Management Science*. To appear.

Cheng, R. C. H. (1986). Variance reduction methods. In *Proceedings of the 1986 Winter Simulation Conference*, pages 60–68. IEEE Press.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, New York, NY, second edition.

de Boor, C. (1978). *A Practical Guide to Splines*. Number 27 in Applied Mathematical Sciences Series. Springer-Verlag, New York.

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer Series in Operations Research. Springer-Verlag, New York, NY.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5:79–141.

Glynn, P. W. (1994). Efficiency improvement techniques. *Annals of Operations Research*, 53:175–197.

Glynn, P. W. and Szechtman, R. (2002). Some new perspectives on the method of control variates. In Fang, K.-T., Hickernell, F. J., and Niederreiter, H., editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 27–49, Berlin. Springer-Verlag.

Hickernell, F. J., Lemieux, C., and Owen, A. B. (2005). Control variates for quasi-Monte Carlo. *Statistical Science*, 20(1):1–31.

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

L'Ecuyer, P. and Buist, E. (2005). Simulation in Java with SSJ. In *Proceedings of the 2005 Winter Simulation Conference*, pages 611–620. IEEE Press.

L'Ecuyer, P. and Buist, E. (2006). Variance reduction in the simulation of call centers. In *Proceedings of the 2006 Winter Simulation Conference*, pages 604–613. IEEE Press.

L'Ecuyer, P. and Lemieux, C. (2000). Variance reduction via lattice rules. *Management Science*, 46(9):1214–1235.

Whitt, W. (1999). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212.