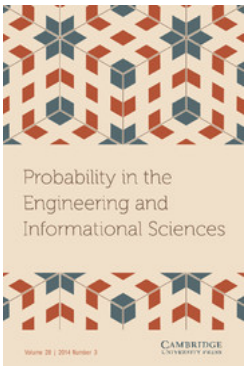

APPLYING VARIANCE REDUCTION IDEAS IN QUEUING SIMULATIONS

Sheldon M. Ross and Kyle Y. Lin

Probability in the Engineering and Informational Sciences / Volume 15 / Issue 04 / October 2001, pp 481 - 494
DOI: null, Published online: 11 January 2002

Link to this article: http://journals.cambridge.org/abstract_S0269964801154045

How to cite this article:
Sheldon M. Ross and Kyle Y. Lin (2001). APPLYING VARIANCE REDUCTION IDEAS IN QUEUING SIMULATIONS. Probability in the Engineering and Informational Sciences, 15, pp 481-494

Probability in the Engineering and Informational Sciences, **15**, 2001, 481–494. Printed in the U.S.A.

APPLYING VARIANCE REDUCTION IDEAS IN QUEUING SIMULATIONS

SHELDON M. ROSS
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA 94720
E-mail: ross@ieor.berkeley.edu

KYLE Y. LIN*
Grado Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061
E-mail: kylin@vt.edu

Variance reduction techniques are often underused in simulation studies. In this article, we indicate how certain ones can be efficiently employed when analyzing queuing models. The first technique considered is that of dynamic stratified sampling; the second is the utilization of multiple control variates; the third concerns the replacement of random variables by their conditional expectations when trying to estimate the expected value of a sum of random variables.

1. INTRODUCTION

There are many standard variance reduction techniques used in simulation (see Ross [11]) and many of them have been used to estimate the performance of queuing systems (see Glynn and Iglehart [3], Lavenberg, Moeller, and Welch [4], and Wilson and Pritsker [13,14]). Moreover, these techniques can often be simultaneously employed to further improve the simulation results; although this idea has appeared in the literature (see Avramidis and Wilson [1] and McGeoch [6]), its use can often be strengthened.

*This work was done when Kyle Lin was at the Department of Industrial Engineering and Operations Research, University of California, Berkeley.

 0269-9648/01 $12.50

In Section 2, we introduce a new idea called *dynamic stratification.* Contrary to the traditional stratified sampling, in which we divide the original problem into strata and run each of them a predetermined number of times, we sequentially allocate each additional run to the stratum yielding the largest estimated variance reduction based on the latest stratum variance estimates.

In Section 3, we illustrate the use of multiple control variables. A computational drawback of this well-known, but underused, technique arises when there are correlations between the control variables (see Fishman [2], Lavenberg and Welch [5], and Nelson [7]). However, in some cases, useful control variables are independent, and in these cases, the computation grows linearly as the number of control variables increases.

If we want to estimate the expected value of the sum of the delays of the first n customers, then a useful technique is to use an estimator that replaces an actual delay by its conditional expectation, given the minimal amount of information that is needed to compute this conditional expectation. Whereas it was shown by Ross in [9] that the sum of these conditional expectations has a smaller variance than does the sum of the actual delays, it is not clear that these replacements should still be made when additional variance reduction techniques are to be implemented. Numerical data relating to this issue are presented in Section 4.

2. DYNAMIC STRATIFICATION

Suppose that we want to use simulation to estimate $E[X]$. Let Y be a discrete random variable with known probabilities $p_i = P\{Y = i\}$, $\sum_{i=1}^{m} p_i = 1$, which is correlated with X, and suppose that it is easy to simulate the value of X conditional on $Y = i$ for each i.

If we plan to do n simulation runs, then it can be shown that rather than simulating n independent copies of X and using the average obtained as an estimator, one obtains a similarly unbiased estimator but with smaller variance by doing np_i of the simulations conditional on $Y = i$, letting θ_i be the average of the runs conditional on $Y = i$, and using

$$\sum_{i=1}^{m} p_i \theta_i$$

as the estimator of $E[X]$ (see Ross [11]). We call the preceding the *prestratification* estimator. However, one can often do much better by a different choice of the number of times to do each of the conditional simulation runs. Indeed, if we simulate n_i runs conditional on $Y = i$, then the variance of the estimator is

$$\sum_{i=1}^{m} \frac{p_i^2 \sigma_i^2}{n_i},$$

where

$$\sigma_i^2 \equiv \mathrm{Var}(X|Y = i).$$

Subject to the constraint $\sum_{i=1}^{m} n_i = n$, the optimal (variance minimizing) choice (see Rubinstein [12]) is to let

$$n_i = n \frac{p_i \sigma_i}{\sum_j p_j \sigma_j}. \tag{2.1}$$

If we let W denote a random variable that is equal to σ_i with probability p_i, $i = 1, \ldots, m$, then the variance of the estimator that uses the values given by Eq. (2.1) is

$$\frac{\left(\sum_i p_i \sigma_i\right)^2}{n} = \frac{E^2[W]}{n},$$

whereas the value of the stratified estimator that uses $n_i = np_i$, $i = 1, \ldots, m$, is

$$\frac{\sum_i p_i \sigma_i^2}{n} = \frac{E[W^2]}{n}.$$

Therefore, $\mathrm{Var}(W)/n$ is the additional variance reduction beyond prestratification that is obtained by letting the number of runs be given by Eq. (2.1). The difficulty with the preceding solution, however, is that the quantities σ_i^2 are unknown. Whereas a common approach is to run a small preliminary simulation to estimate the quantities σ_i^2, $i = 1, \ldots, m$, and then use Eq. (2.1) with the estimates substituting for the conditional variances to determine the n_i, we propose an apparently new approach that we call *dynamic stratification*; it is described as follows:

1. For a given integer k, do k runs conditional on $Y = i$ for each $i = 1, \ldots, m$. Let θ_i and s_i^2 be respectively the sample mean and sample variance of the values of X obtained on the k runs conditional on $Y = i$. Set $n_i = k$ for all i.
2. Repeat Steps 3–6 for a total of $n - mk$ times.
3. Let
$$b_i = p_i^2 s_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i + 1}\right).$$
4. Let j be such that
$$b_j = \max_i b_i.$$
5. Simulate X conditional on $Y = j$. Let the simulated value be x.
6. Reset values
$$s_j^2 = \frac{n_j - 1}{n_j} s_j^2 + \frac{1}{n_j + 1} (\theta_j - x)^2,$$
$$\theta_j = \frac{n_j \theta_j + x}{n_j + 1},$$
$$n_j = n_j + 1.$$

At the end of the simulation, θ_i is the sample mean of X conditional on $Y = i$ (or stratum i), and the estimator is given by $\sum_{i=1}^{m} p_i \theta_i$. Since the number of runs and the sample variance of each stratum are n_i and s_i^2 respectively, the variance of the estimator can be approximated by

$$\sum_{i=1}^{m} p_i^2 \frac{s_i^2}{n_i}.$$

Remark: The rationale for the dynamic stratification approach comes from the fact that the problem of minimizing

$$\sum_{i=1}^{m} \frac{p_i^2 \sigma_i^2}{n_i}$$

can be solved sequentially. Specifically, let $H(n)$ be the minimal value of the preceding when $\sum_{i=1}^{m} n_i = n$, and suppose that $(n_1^*, \ldots, n_m^*)$ is one set of values yielding $H(n)$. Then, because of convexity of the functions

$$f_i(k) \equiv \frac{p_i^2 \sigma_i^2}{k},$$

it can be shown (see the Appendix) that if j is such that

$$f_j(n_j^*) - f_j(n_j^* + 1) = \max_i \{ f_i(n_i^*) - f_i(n_i^* + 1)\},$$

then

$$H(n+1) = f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*).$$

Since the reset values in Step 6 are equal to the new sample mean and sample variance for those runs conditional on $Y = j$, it follows that at each step, the dynamic stratification algorithm chooses the simulation run that yields the largest estimated improvement.

For an illustration of the preceding, consider a queuing system in which customers arrive according to a Poisson process with rate λ, and suppose that, for a specified time t, we are interested in $E[X]$, the expected value of the sum of the delays in the queue of all customers that arrive by time t. Letting $N(t)$ denote the number of arrivals by time t, we can choose integer values n on both sides of $E[N(t)] = \lambda t$ and use dynamic stratification by simulating the arrival process conditional on $N(t) = n$. For instance, if $\lambda t = 100$, we can write

$$E[X] = \sum_{i=65}^{135} E[X|N(t) = i] p_i + E[X|N(t) < 65] P(64) + E[X|N(t) > 135] \bar{P}(135),$$

where

$$p_i = \frac{e^{-\lambda t}(\lambda t)^i}{i!}, \qquad P(j) = \sum_{i=0}^{j} p_i, \qquad \bar{P}(j) = 1 - P(j);$$

that is, in essence, we break up the simulation problem into 73 strata and, at each stage, we use the dynamic stratification algorithm to decide which stratum to simulate. The arrival process conditional on $N(t) = i$ can be simulated by generating i independent uniform (0,1) random numbers and then sorting them (see Ross [10]). The jth sorted value multiplied by t then gives the time of the jth arrival. The sorting step can be eliminated by generating $i + 1$ random numbers $U_1, \ldots, U_{i+1}$ and then taking the natural log of each one. The time of the jth arrival, $j = 1, \ldots, i$, would be given by

$$\frac{t \sum_{k=1}^{j} \log(U_k)}{\sum_{k=1}^{i+1} \log(U_k)} = \frac{t \log\left(\prod_{k=1}^{j} U_k\right)}{\log\left(\prod_{k=1}^{i+1} U_k\right)}.$$

The arrival process conditional on $N(t) < 65$ (or $N(t) > 135$) can be simulated by first using the discrete inverse transform algorithm to generate the value of $N(t)$ conditional on $N(t) < 65$ (or $N(t) > 135$) and then using the preceding to generate the arrival times. Table 1 gives the simulation results for estimating the expected value of the total delay of all customers who arrive by $t = 100$ in an $M/M/1$ queue with $\lambda = \mu = 1$, based on 1 million runs.

The additional work needed for dynamic stratification compared to prestratification is Steps 3 and 4. However, their computational complexities are 1 and $\log m$ (construct an ordered list of b_i and insert the updated b_j each time), respectively. Both of these are negligible when each run involves generating hundreds of random numbers.

TABLE 1. Total Delay of the Customers Arriving by $t = 100$ for an $M/M/1$ Queue, $\lambda = \mu = 1$

	Estimated Variance of Estimator
Raw estimator	0.285
Prestratification	0.161
Dynamic stratification	0.142

The dynamic stratification approach can also be efficiently employed when the arrival process is a nonhomogeneous Poisson process with intensity function $\lambda(t)$. The simulation of the conditional (on the number of arrivals by time t) arrival process can be effected by using the result that the unordered set of arrival times are i.i.d. (independent and identically distributed) according to the distribution function

$$F(x) = \frac{\int_0^x \lambda(y)\,dy}{\int_0^t \lambda(y)\,dy}, \qquad 0 \le x \le t.$$

We can still use dynamic stratification on the arrival process even when we are interested in the expected value of the sum of the delays in the queue of the first r customers (or just the expected delay of customer r), for some specified r. If the arrival process is a homogeneous Poisson process, then we can use that T_r, the time of the rth arrival, is gamma distributed with parameters (r, λ). We can then specify intervals

$$(0, t_1), (t_1, t_2), \ldots, (t_{m-1}, t_m), (t_m, \infty)$$

with $E[T_r] = r/\lambda$ being in one of the intervals near the middle, and let stratum i be conditional on the event that T_r lies in (t_{i-1}, t_i), where $t_0 = 0$ and $t_{m+1} = \infty$. The values of p_i are easily computed and the simulation of the arrival process conditional on T_r lying in a specified interval can be effected by first generating the value of T_r given that it lies in that interval (the rejection method with a uniform distribution can be used). If $T_r = s$, we then generate the arrival times by generating and then ordering $r-1$ uniforms; the jth of these multiplied by s gives the simulated time of the jth arrival.

Another way of using a dynamic stratification approach relating to arrivals when we are interested in the sum of delays of the first r customers is to do the simulation conditional on $N(r/\lambda)$. To accomplish this, first generate the value of this random variable, say its generated value is m. Then, generate the m arrival times conditional on $N(r/\lambda) = m$. If $r \le m$, take the first r of these as the arrival process; if $m < r$, take these as the first m arrival times, and then generate an additional $r - m$ exponentials with rate λ, say $Y_1, \ldots, Y_{r-m}$. The kth, $m < k \le r$, arrival time should be set equal to T_k, where

$$T_{m+1} = \frac{r}{\lambda} + Y_1,$$
$$T_{m+i} = T_{m+i-1} + Y_i, \qquad i = 2, \ldots, r - m.$$

Table 2 gives the simulation results for estimating the expected value of the total delay of the first 100 customers for an $M/M/1$ queue with $\lambda = \mu = 1$ based on 1 million runs.

The use of dynamic stratification breaks up a simulation problem into strata. To efficiently do the simulation, additional variance reduction techniques should be

TABLE 2. Total Delay of the First 100 Customers for an $M/M/1$ Queue, $\lambda = \mu = 1$

	Estimated Variance of Estimator
Raw estimator	0.185
Dynamic stratification on $N(r/\lambda)$	0.130
Dynamic stratification on T_r	0.128

utilized in obtaining the final estimators in the strata. This is indicated in the following section.

3. MULTIPLE CONTROL VARIATES

Consider a queuing system in which the arrival process is a renewal process and the customer service times have known probability distributions. Let D_i, $i = 1,\ldots,n+1$, denote the delay in queue of customer i and suppose that we are interested in using simulation to estimate $E[D]$, where D is a linear combination of the D_i. (For instance, D might be the sum of all the delays, or the delay of customer $n+1$.) Also, let A_i be the interarrival time between customers i and $i+1$, and let S_i be the service time of customer i. Since D is usually an increasing function of the service times and a decreasing function of the interarrival times, the quantity $\sum_{i=1}^{n}(S_i - A_i)$, is often proposed as a control variate (see Ripley [8]). However, since it is not clear that equal weight should be given to the amounts by which $\sum_{i=1}^{n} S_i$ and $\sum_{i=1}^{n} A_i$ exceed their known means, it would be better to use both of these sums as control variates. Indeed, it would be even better to use all $2n$ random variables $A_1, S_1, \ldots, A_n, S_n$ as control variates; that is, we should consider an estimator of the form

$$D + \sum_{i=1}^{n} a_i(A_i - E[A_i]) + \sum_{i=1}^{n} b_i(S_i - E[S_i]),$$

where the best values of the constants a_i and b_i are to be determined by the simulation. It should be noted that when the sequence of random variables $A_1, S_1, \ldots, A_n, S_n$ is independent, the determination of these constants only requires covariance estimations between individual ones of these random variables and D. Therefore, to minimize the variance of the estimator, one should set

$$a_i = -\frac{\mathrm{Cov}(D, A_i)}{\mathrm{Var}(A_i)},$$

$$b_i = -\frac{\mathrm{Cov}(D, S_i)}{\mathrm{Var}(S_i)}.$$

TABLE 3. Total Delay of the First 100 Customers for an $M/M/1$ Queue, $\lambda = \mu = 1$

Control Variate	Estimated Variance of Estimator
Raw estimator	0.185
$\sum_{i=1}^{99} A_i$	0.138
$\sum_{i=1}^{99} S_i$	0.105
$\sum_{i=1}^{99} (S_i - A_i)$	0.060
$\sum_{i=1}^{99} A_i$ and $\sum_{i=1}^{99} S_i$	0.057
$A_1, \ldots, A_{99}$	0.130
$S_1, \ldots, S_{99}$	0.091
$A_1, S_1, \ldots, A_{99}, S_{99}$	0.037

Then, the minimized variance is

$$\mathrm{Var}(D) - \sum_{i=1}^{n} \frac{\mathrm{Cov}^2(D, A_i)}{\mathrm{Var}(A_i)} - \sum_{i=1}^{n} \frac{\mathrm{Cov}^2(D, S_i)}{\mathrm{Var}(S_i)}.$$

In practice, we can estimate $\mathrm{Var}(D)$, $\mathrm{Cov}(D, A_i)$, and $\mathrm{Cov}(D, S_i)$ from the simulation and use them to estimate a_i, b_i, and the variance of the estimator. Table 3 gives simulation results for estimating the expected value of the total delay of the first 100 customers for an $M/M/1$ queue having mean interarrival and mean service times both equal to 1; that is, $D = \sum_{i=1}^{100} D_i$.

One can see that individually using all the A_i and S_i as control variates results in a variance approximately 60% of that obtained by using $\sum(S_i - A_i)$ and one-fifth of that of the raw estimator. Moreover, the only additional work of using each A_i and S_i as a control variable is to keep track of the sample mean of A_i, S_i, and also $\mathrm{Cov}(A_i, D), \mathrm{Cov}(S_i, D)$. This is very little work compared to generating the whole process.

The estimated correlation coefficients of the total delay D and $A_i (S_i)$ are plotted in Figure 1 (Fig. 2). From the figures, we can see that the correlation coefficient between D and S_i is approximately a unimodular function in i, which increases and has its peak around $i = 20$ and then decreases toward 0. Intuitively the correlation between S_i and D decreases in i when i is large, because the larger the value of i, the fewer is the number of customers that will arrive after him. However, consider the first customer; although there are 99 customers behind him, the system is empty when he arrives and S_1 plays a role only if it is greater than A_1. On the other hand, when customer 20 arrives, it is more likely that a line is already there, so S_{20} affects the total delay more than S_1 does.

Remark: Technically speaking, our controlled estimators are not unbiased since they involve estimators of a_i and b_i rather than the actual constants. However, we

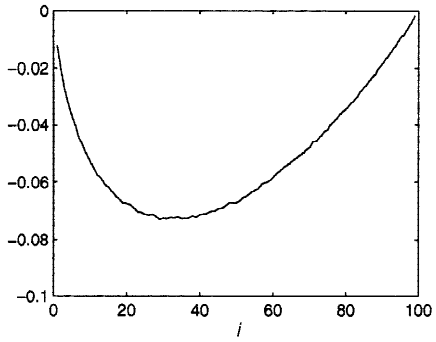

FIGURE 1. The estimated correlation coefficient between A_i and $\sum_{i=1}^{100} D_i$.

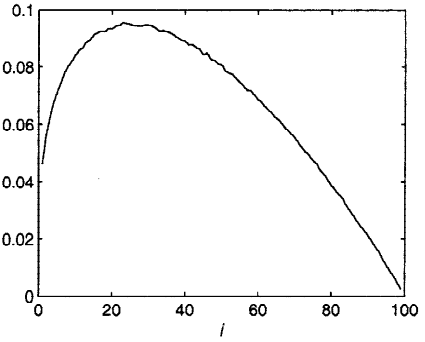

FIGURE 2. The estimated correlation coefficient between S_i and $\sum_{i=1}^{100} D_i$.

should gain almost all of the benefit of using the theoretical best constants. Our variance estimators are estimates of the idealized situation where the best constants rather than estimators of them are used.

Suppose now that the arrival process is a Poisson process and that we have made use of the dynamic stratification approach of Section 2. Assuming that there are enough simulations done conditional on r arrivals, we recommend using each of the r service times as a control variable. If there are not that many runs conditional on r arrivals, then it would not be possible to accurately estimate all of the needed covariances between D and S_i, $i = 1, \ldots, r$, and so we recommend the single control variable $\sum_{i=1}^{r} S_i$ in this case. The simulation results of this strategy are compared in Table 4.

Now let us return to the problem in which D is the sum of the delays of all customers that arrive by a fixed time t. When not stratifying, the natural control

TABLE 4. Total Delay of the first 100 Customers for an $M/M/1$ Queue, $\lambda = \mu = 1$

Dynamic Stratification	Control Variate	Estimated Variance of Estimator
On $N(r/\lambda)$	none	0.130
On $N(r/\lambda)$	$\sum_{i=1}^{99} S_i$	0.051
On $N(r/\lambda)$	$S_1, \ldots, S_{99}$[a]	0.036
On T_r	None	0.128
On T_r	$\sum_{i=1}^{99} S_i$	0.049
On T_r	$S_1, \ldots, S_{99}$[a]	0.034

[a]Use $\sum_{i=1}^{99} S_i$ instead for those strata where the number of runs is less than 1000.

TABLE 5. Total Delay of the Customers by $t = 100$ for an $M/M/1$ Queue, $\lambda = \mu = 1$

Dynamic Stratification	Control Variate	Estimated Variance of Estimator
No	None	0.285
No	$N(t)$	0.169
No	$\sum_{i=1}^{N(t)-1} S_i$	0.078
No	$N(t)$ and $\sum_{i=1}^{N(t)-1} S_i$	0.077
Yes	None	0.142
Yes	$\sum_{i=1}^{r-1} S_i$	0.054
Yes	$S_1, \ldots, S_{r-1}$ [a]	0.034

[a]Use $\sum_{i=1}^{r-1} S_i$ instead for those strata where the number of runs is less than 1000.

variables are $N(t)$, the number of arrivals by time t, and $\sum_{i=1}^{N(t)-1} S_i$, the total service time of the arrivals by time t except the last one.[1] Of course, we can use both of them at the same time, but not much improvement is expected because of the strong correlation between the two of them. When stratifying on $N(t)$, the variance in each stratum lies mainly in the service process, so we suggest using $S_1, \ldots, S_{r-1}$ or $\sum_{i=1}^{r-1} S_i$ as control variables for the stratum $N(t) = r$, depending on whether the number of runs in a stratum is enough to accurately estimate $\mathrm{Cov}(S_i, D)$.

Table 5 gives the simulation results for estimating the expected value of the total delay of all the customers that arrive by $t = 100$ for an $M/M/1$ queue with $\lambda = \mu = 1$ based on 1 million runs.

Although we reduce most of the variance in the arrival process by stratifying $N(t)$, we can still utilize a control variable related to arrivals conditional on $N(t) = r$. Intuitively, D will be smallest when the arrivals are spread out, rather than being clumped in places. For instance, if we are interested in the sum of the delays of all arrivals by time t and the interarrivals A_i have been generated conditional on r arrivals by time t, then the quantity $\sum_{i=1}^{r} A_i^2 + (t - \sum_{i=1}^{r} A_i)^2$, which is minimized when the r arrivals are equally spaced, is a good choice for a control variable. Since each A_i and $t - \sum_{i=1}^{r} A_i$ have, in this case, the same distribution as the smallest order statistic of a set of r i.i.d. uniform $(0, t)$ random variables, it follows that

$$E[A_i^2] = \frac{2t^2}{(r+1)(r+2)}$$

and therefore

$$E\left[\sum_{i=1}^{r} A_i^2 + \left(t - \sum_{i=1}^{r} A_i\right)^2\right] = \frac{2t^2}{r+2}.$$

[1] We can compute the expectation of this control variable since S_i and $N(t) - 1$ are independent.

Another possibility for a control variable for the arrival process is to divide the interval $(0, t)$ into, say, k subintervals, let l_i denote the number of the r arrivals that are in subinterval i, and then use $\sum_{i=1}^{k} l_i^2$ as a control variable.

When applying $\sum_{i=1}^{r} A_i^2 + (t - \sum_{i=1}^{r} A_i)^2$ or $\sum_{i=1}^{k} l_i^2$ along with dynamic stratification in the model described in Table 5, the estimated variance of the estimators are 0.142 and 0.140 (divide the whole interval into $k = 10$ subintervals with equal length), respectively. The improvement over just using dynamic stratification is thus rather minimal. However, since the arrival process and the service process are independent, we can use $S_1, \ldots, S_{r-1}$ and $\sum_{i=1}^{r} A_i^2 + (t - \sum_{i=1}^{r} A_i)^2$ (or $\sum_{i=1}^{k} l_i^2$) at the same time and the reduction in variance is additive.

4. CONDITIONAL ESTIMATOR

To estimate $E[\sum_{i=1}^{n} D_i]$ in a $G/M/k$ (or $G/G/1$) queue, Ross [9] showed that $\sum_{i=1}^{n} E[D_i|H_i]$ is a better estimator than $\sum_{i=1}^{n} D_i$, where H_i is the history of the process up to the moment that customer i arrives and the minimal information needed to compute the conditional expectation. For example, in a $G/G/1$ queue with mean service time $1/\mu$, $H_i = \{R_i \text{ and } s\}$, where R_i is the number of customers in the system and s is the age of the customer in service when customer i arrives. Let S denote the service time of the customer in service when customer i arrives; then,

$$E[D_i|H_i] = \frac{(R_i - 1)^+}{\mu} + E[S - s|S > s]I\{R_i > 0\}.$$

If it is not easy to compute $E[S - s|S > s]$, we can let H_i also include the actual remaining service time of the customer being served.

Now, suppose we want to estimate the expected total delay of the first 100 customers for an $M/M/1$ queue. The conditional estimator of the delay of customer i is given by

$$E[D_i|R_i] = \frac{R_i}{\mu}.$$

In addition to the conditional estimator, we use control variables suggested in Section 3. In Table 6, we compare the estimated variances of the two estimators when we use them together with different choices of control variables, based on 1 million runs.

It is noted in Table 6 that although the variances in column 2 are always smaller than those in column 1, we gain more from the control variables S_i when we use $\sum_{i=1}^{100} D_i$ since the correlation between S_i and $\sum_{i=1}^{100} D_i$ is stronger than that between S_i and $\sum_{i=1}^{100} E[D_i|R_i]$. In fact, when μ gets smaller, it may be the case that it is better to not use the conditional estimator. Table 7 gives an example of when it is the best strategy to use $\sum_{i=1}^{100} D_i$ and all A_i, S_i as control variables. The reason that $\sum_{i=1}^{n} D_i$ is preferable over $\sum_{i=1}^{n} E[D_i|R_i]$ when using control variables involving service time

TABLE 6. Total Delay of the First 100 Customers for an $M/M/1$ Queue, $\lambda = \mu = 1$

	Estimated Variance of Estimator	
Control Variate	$\sum_{i=1}^{100} D_i$	$\sum_{i=1}^{100} E[D_i \mid R_i]$
None	0.185	0.145
$A_1, \ldots, A_{99}$	0.130	0.090
$S_1, \ldots, S_{99}$	0.091	0.084
$A_1, S_1, \ldots, A_{99}, S_{99}$	0.037	0.029

is that $\sum_{i=1}^{n} D_i$ is more correlated with S_i than $\sum_{i=1}^{n} E[D_i \mid R_i]$ is. If we want to take advantage of the conditional expectation, we have to sacrifice the stronger correlation between $\sum_{i=1}^{n} D_i$ and the service time, and sometimes it is not worth it.

Acknowledgment

This research was supported by the National Science Foundation grant DMI-9901053 with the University of California.

References

1. Avramidis, A.N. & Wilson, J.R. (1996). Integrated variance reduction strategies for simulation. *Operations Research* 44: 327–346.
2. Fishman, G.S. (1989). Monte Carlo, control variate, and stochastic ordering. *SIAM Journal on Scientific and Statistical Computing* 10: 187–204.
3. Glynn, P.W. & Iglehart, D.L. (1988). Simulation methods for queues: An overview. *Queueing Systems* 3: 221–256.
4. Lavenberg, S.S., Moeller, T.L., & Welch, P.D. (1982). Statistical results on control variables with application to queueing network simulation. *Operations Research* 30: 182–202.

TABLE 7. Total Delay of the First 100 Customers for an $M/M/1$ Queue, $\lambda = 1$, $\mu = 0.5$

	Estimated Variance of Estimator	
Control Variate	$\sum_{i=1}^{100} D_i$	$\sum_{i=1}^{100} E[D_i \mid R_i]$
None	1.549	0.874
$A_1, \ldots, A_{99}$	1.261	0.586
$S_1, \ldots, S_{99}$	0.297	0.318
$A_1, S_1, \ldots, A_{99}, S_{99}$	0.009	0.030

5. Lavenberg, S.S. & 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.
6. McGeoch, C. (1992). Analyzing algorithms by simulation: Variance reduction techniques and simulation speedups. *ACM Computing Surveys* 24(2): 195–212.
7. Nelson, B.L. (1990). Control variate remedies. *Operations Research* 38: 974–992.
8. Ripley, B.D. (1987). *Stochastic simulation.* New York: Wiley.
9. Ross, S.M. (1988). Simulating average delay—Variance reduction by conditioning. *Probability in the Engineering and Informational Sciences* 2: 309–312.
10. Ross, S.M. (1996). *Stochastic process*, 2nd ed. New York: Wiley.
11. Ross, S.M. (1997). *Simulation*, 2nd ed. Boston: Academic Press.
12. Rubinstein, R.Y. (1996). *Simulation and the Monte Carlo method.* New York: Wiley.
13. Wilson, J.R. & Pritsker, A.A.B. (1984). Experimental evaluation of variance reduction techniques for queueing simulation using generalized concomitant variables. *Management Science* 30: 1459–1472.
14. Wilson, J.R. & Pritsker, A.A.B. (1984). Variance reduction in queueing simulation using generalized concomitant variables. *Journal of Statistical Computation and Simulation* 19: 129–153.

APPENDIX

Let $f_i(n_i)$ be convex functions defined on nonnegative integers, $i = 1, \ldots, m$. For some positive integer n, consider the problem

$$\min \sum_i f_i(n_i)$$

$$\text{s.t. } n_i \text{ are nonnegative integers and } \sum_i n_i = n.$$

Let $H(n)$ denote the minimal value of the preceding and suppose $(n_1^*, \ldots, n_m^*)$ is one set of values such that

$$H(n) = \sum_{i=1}^{m} f_i(n_i^*). \tag{A.1}$$

LEMMA A.1: *If j is such that*

$$f_j(n_j^*) - f_j(n_j^* + 1) = \max_i \{ f_i(n_i^*) - f_i(n_i^* + 1)\}, \tag{A.2}$$

then

$$H(n + 1) = f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*). \tag{A.3}$$

PROOF: We prove by contradiction; that is, if Eq. (A.3) is not true, then we show that Eq. (A.1) would not have been true by presenting $(\hat{n}_1, \ldots, \hat{n}_m)$ such that

$$\sum_i f_i(\hat{n}_i) < \sum_i f_i(n_i^*).$$

Suppose there exists $(\bar{n}_1, \ldots, \bar{n}_m)$ such that

$$\sum_i f_i(\bar{n}_i) < f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*).$$

Consider two cases:

1. $\bar{n}_j > n_j^* + 1$. Let $\hat{n}_i = \bar{n}_i$ for $i \neq j$ and $\hat{n}_j = \bar{n}_j - 1$. Then,

$$\begin{aligned}\sum_i f_i(\hat{n}_i) &= \sum_i f_i(\bar{n}_i) + f_j(\bar{n}_j - 1) - f_j(\bar{n}_j) \\ &< f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*) + f_j(\bar{n}_j - 1) - f_j(\bar{n}_j) \\ &\leq f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*) + f_j(n_j^*) - f_j(n_j^* + 1) \\ &= \sum_i f_i(n_i^*),\end{aligned}$$

   where the first inequality comes from the assumption, and the second comes from the fact that $\bar{n}_j > n_j^* + 1$ and f_j is convex.
2. $\bar{n}_j \leq n_j^* + 1$. In this case, there must exist some k such that $\bar{n}_k \geq n_k^* + 1$. Let $\hat{n}_i = \bar{n}_i$ for $i \neq k$ and $\hat{n}_k = \bar{n}_k - 1$. Then,

$$\begin{aligned}\sum_i f_i(\hat{n}_i) &= \sum_i f_i(\bar{n}_i) + f_k(\bar{n}_k - 1) - f_k(\bar{n}_k) \\ &< f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*) + f_k(n_k^*) - f_k(n_k^* + 1) \\ &\leq f_j(n_j^* + 1) + \sum_{i \neq j} f_i(n_i^*) + f_j(n_j^*) - f_j(n_j^* + 1) \\ &= \sum_i f_i(n_i^*),\end{aligned}$$

   where the second inequality follows because of the way j is selected in Eq. (A.2).

Therefore, the proof is complete. ■