MANAGEMENT SCIENCE
Vol. 37, No. 6, June 1991
Printed in U.S.A.

THE EFFICIENCY OF ONE LONG RUN VERSUS INDEPENDENT REPLICATIONS IN STEADY-STATE SIMULATION*

WARD WHITT
AT&T Bell Laboratories, Room 2C-178, Murray Hill, New Jersey 07974-2070

We evaluate the efficiency of one long run versus independent replications in steady-state discrete-event simulation, assuming that an initial portion of each replication will be deleted to allow the process to approach steady state. We provide supporting evidence in favor of one long run, but we also show that multiple replications can be more efficient. The advantage of one long run increases if the amount deleted increases or if the covariance function decreases more quickly (assuming it is nonnegative and decreasing). Thus, assuming that the amount deleted depends on the way the process approaches steady state, one long run tends to be efficient when the covariance function decays rapidly compared to the rate the process approaches steady state. We also discuss ways to determine the initial portion to delete. We consider the case of an exponential covariance function in detail, and use it as a basis for approximations. We also consider the $M/G/\infty$ queueing model and reflected Brownian motion, the latter as an approximation for the $G/G/1$ queueing model. For these models starting at the origin, one long run is efficient, but a moderate number of independent replications is essentially equally efficient. In agreement with Kelton and Law (1984), for such examples our analysis only rules out many replications of very short runs.

(SIMULATION; EXPERIMENTAL DESIGN; SIMULATION EFFICIENCY; INDEPENDENT REPLICATIONS; COVARIANCE FUNCTION; INITIAL BIAS)

1. Introduction

Two curses of steady-state discrete-event simulation are the initial bias and the autocorrelations. *Initial bias* is the difference between the expected value of the estimator, here presumed to be a sample mean, and the quantity it is estimating. Initial bias occurs because we cannot start the process in steady state. The effect of initial bias can be reduced by deleting an initial portion of the run.

Autocorrelations are the correlations between successive values of the process. Autocorrelations occur because new states of the process typically depend strongly on previous ones. The autocorrelation function is typically positive and decreasing (or nearly so), reflecting the fact that new states of the process are similar to previous states, with the similarity decreasing (in a random way) as the process evolves. The effect of autocorrelations can be reduced by using independent replications, but at the cost of having to delete an initial portion from each replication. Thus it is natural to ask: *Is it better to make one long run or many independent replications?*

Of course, this question has been studied before, see Fishman (1972), Cheng (1976), Law (1977) and Kelton and Law (1984), but there evidently is no consensus about what is an appropriate simulation strategy; see p. 80 of Bratley, Fox and Schrage (1987). We contend that this is inevitable, because *there is no simple answer*. We provide examples showing that each strategy can be much more efficient than the other, thus demonstrating that a simple unqualified conclusion is inappropriate.

We are thinking of conventional simulation, but the issue of one long run versus independent replications also arises in parallel simulations. With many processors, we may do a distributed simulation of one run or do independent replications with one

* Accepted by James R. Wilson; received August 15, 1989. This paper has been with the author 1 month for 2 revisions.

0025-1909/91/3706/0645$01.25

replication per processor; see Heidelberger (1986), Glynn and Heidelberger (1989, 1990) and references cited there. Our analysis may provide some useful insights for parallel simulation, but we do not discuss it further.

In this paper we provide support for the conclusion that it usually is more efficient to make one long run than make independent replications, but *we conclude that it usually does not matter much*. If the total simulation run length is long enough to obtain reasonably good estimates, then several replications are usually just as efficient as one longer run. In agreement with Kelton and Law (1984), our analysis only rules out a very large number of independent replications of very short runs.

A main idea here is to evaluate these simulation strategies using *the criterion of estimation efficiency*, as discussed in Glynn and Whitt (1991). In particular, estimation efficiency here is defined to be the reciprocal of the product of the mean squared error and the total simulation run length. We are thus identifying the total simulation run length with the total cost of conducting the experiment, which of course is a simplifying assumption. By using the efficiency criterion, we are able to circumvent the complicated optimization problem obtained when we try to minimize the mean squared error subject to a constraint on the total simulation run length (see §2).

However, it should be noted that there are other relevant criteria that we do not consider. By focusing on efficiency, we consider the quality of the estimate (variance or mean squared error), but not the quality of any estimate of the quality of the estimate, i.e., we do not consider how to construct confidence intervals or whether they have proper coverage. We also do not consider other measures to improve the efficiency, such as initializing the simulation to make the process approach steady state more quickly, e.g., see Kelton (1990). Our goal is not to identify the best simulation strategy, considering the proper balance of all relevant criteria, but to provide additional insight into the trade-off between initial bias and autocorrelations. Thus we consider only the efficiency criterion.

Our analysis indicates that doing fewer runs (e.g., only one) is more efficient when the autocorrelations decrease rapidly compared to the rate the process approaches steady state. This conclusion seems to provide a useful practical guideline for choosing simulation strategies. Even though it is usually not possible to determine precisely how the autocorrelations decrease or precisely how the process approaches steady state, we often have some idea about what should happen (perhaps by doing pilot runs). In fact, the analysis of the examples in this paper draws on previous investigations of the autocorrelations and the way the process approaches steady state; e.g., Abate and Whitt (1987a, b, 1988).

A second main idea here is to *initially assume that the amount to delete has been specified*. We assume that the initial portion to delete has been chosen sufficiently large that the process of interest after the deletion is approximately stationary. By assuming that the initial portion to delete has been specified, we separate the problem of bias from the problem of variance. Moreover, our analysis thus covers cases in which the initial portion to delete is larger than necessary, as well as the case in which it is chosen properly. We also investigate ways to determine an appropriate amount to delete. Our analysis indicates that the appropriate amount to delete should usually be a relatively small portion of the total simulation run length.

The assumptions underlying our analysis seem most appropriate in a large sample context. Indeed, our results seem consistent with the large-sample asymptotics of Glynn (1987, 1988) and Glynn and Heidelberger (1989, 1990), which we were unaware of when we did this work. These asymptotic results describe the limiting behavior of the mean squared error (equivalently, the asymptotic efficiency) as the computational budget or total simulation run length increases, depending on the number of replications and the amount of initial transient deletion. These asymptotic results also show that there can be difficulties if the number of replications is too large. Moreover, they show that once a minimal amount of initial transient deletion is performed (which goes to infinity

slowly as the computational budget increases) the asymptotic efficiency of the estimator is essentially independent of the number of replications (provided that the number of replications does not grow too quickly as the computational budget increases).

Before going further, we give an *intuitive* argument in favor of one run. Suppose that a common fixed starting state is used for each run, an initial portion of length s is deleted from each run, and the total run length is $2(s + t)$. As one candidate, consider two replications, i.e., two runs each of length $s + t$, deleting the initial interval $[0, s]$ from each. First, we contend that it should be better to do one run of length $2(s + t)$ deleting the two intervals $[0, s]$ and $[t + s, t + 2s]$, because we expect that the process at time $t + 2s$ will be in steady state approximately independent of the process before time $t + s$, assuming that this was true for the case $t = 0$. Indeed, we expect this to be more likely to be true, because the process is presumably in steady state at time $t + s$ but not at time 0. (An example in §7.1 shows that this logic can fail, but we usually expect it to be valid.) Second, given that we have one long run in steady state after time s, we expect that we cannot gain efficiency by throwing away data. Thus, we expect it to be better to use one run of length $2(s + t)$ deleting only the one interval $[0, s]$.

Here is how the rest of this paper is organized. In §2 we formulate our problem precisely in terms of the reciprocal of the efficiency, which we call the risk. In §3 we analyze the risk function in detail for the special case of an exponential covariance function. In §4 we discuss exponential approximations to determine the initial portion to delete. In §5 we analyze the M/G/∞ queueing model. For exponential service-time distributions, this model satisfies the exponential assumptions of §§3 and 4, but for other service-time distributions it does not. In §6 we analyze regulated Brownian motion (RBM) and the M/M/1 queue; they serve as useful approximations for other queueing models; see Whitt (1989). In §7 we present two rather extreme examples, which demonstrate that one long run can be either much more efficient or much less efficient than multiple replications. Finally, in §8 we draw some conclusions.

2. The Efficiency Criterion with Specified Deletion per Replication

Consider a real-valued stochastic process $\{Y(t) : t \geq 0\}$ such that $Y(t) \Rightarrow Y(\infty)$ as $t \to \infty$, where $Y(\infty)$ is a proper random variable and $\Rightarrow$ denotes convergence in distribution. Our object is to estimate the steady-state mean $m \equiv E[Y(\infty)]$.

REMARKS. (2.1) This setting is more general than it might appear, because we can have $Y(t) = f(X(t))$, where $\{X(t) : t \geq 0\}$ is a stochastic process with general state space and f is a real-valued function on this state space. For example, if f is the indicator function of a set A, then $m = P(X(\infty) \in A)$.

(2.2) We only consider continuous-time stochastic processes, but a similar analysis can be done for discrete-time stochastic processes.

To begin with, we assume that a decision has been made to delete an initial portion of length s from each run. Moreover, we assume that s is large enough so that the stochastic process $\{Y(t) : t \geq s\}$ can be regarded as a strictly stationary stochastic process with mean $E[Y(t)] = E[Y(\infty)] = m$, variance $\text{Var}[Y(t)] = \text{Var}[Y(\infty)] = \sigma_\infty^2 < \infty$ and (auto) covariance function

$$R(u) = \text{cov}[Y(t), Y(t + u)] = E[Y(t)Y(t + u)] - (E[Y(t)])^2, \qquad u \geq 0, \quad (1)$$

for all $t \geq s$, where the integrals $\int_0^\infty |R(u)|\,du$ and $\int_0^\infty R(u)du$ are well defined and finite.

Of course, only rarely does a stochastic process actually reach steady state in finite time; i.e., only rarely does there exist finite s such that $Y(t) \stackrel{d}{=} Y(\infty)$ for all $t \geq s$, where $\stackrel{d}{=}$ denotes equality in distribution and, more generally, $\{Y(t) : t \geq s\} \stackrel{d}{=} \{Y(t + h) : t \geq s\}$ for all $h > 0$. (Examples appear in §§5.2 and 7.1 here and in Glynn and Iglehart

1988.) We make this assumption, not because we are primarily interested in stochastic processes that actually reach steady state in finite time, but because we believe that it is a reasonable approximation.

REMARK. (2.3) Of course, in practice choosing s is difficult. A suitably large s will often be obtained by choosing s larger than necessary "to be on the safe side." We discuss ways to select s in §4.

Let $\{Y_i(t) : t \geq 0\}$, $1 \leq i \leq n$, be n independent copies of $\{Y(t) : t \geq 0\}$. We propose to estimate the steady-state mean m by the *grand sample mean*

$$\hat{m} \equiv \hat{m}(n, t) = n^{-1} \sum_{i=1}^{n} \bar{Y}_i(t), \tag{2}$$

where $\bar{Y}_i(t)$ is the sample mean for process i over the interval $[s, s + t]$, i.e.,

$$\bar{Y}_i(t) = t^{-1} \int_s^{s+t} Y_i(u)du. \tag{3}$$

In other words, we are going to use n *independent replications* of length t after deleting an interval of length s from each process to reduce the initial bias. Hence, the total simulation run length is $n(s + t)$.

Under our stationarity assumption, $\hat{m}(n, t)$ is an unbiased estimator of m, i.e., $E[\hat{m}(n, t)] = m$, with variance

$$\text{Var}\,[\hat{m}(n, t)] = n^{-1} \text{ Var } [\bar{Y}_1(t)] = \frac{2}{nt^2} \int_0^t (t - u)R(u)du \tag{4}$$

and limiting *time-average variance constant*

$$\lim_{t \to \infty} t \text{ Var } [\hat{m}(n, t)] = n^{-1} \lim_{t \to \infty} t \text{ Var } [\bar{Y}_1(t)] \equiv \frac{\bar{\sigma}_\infty^2}{n} = \frac{2}{n} \int_0^\infty R(u)du, \tag{5}$$

by standard calculations; e.g., Chapter 3 of Parzen (1962).

The standard approach at this point, with s given, is to introduce a budget constraint B on the total run length, and then look for values of n and t that minimize the variance Var $[\hat{m}(n, t)]$ in (4) subject to the budget constraint $n(s + t) \leq B$. However, we propose an alternative approach that eliminates the variable n from consideration. In particular, we do not directly introduce the budget constraint. In order to determine what are good choices for n and t, we use the criterion of *efficiency*; see Glynn and Whitt (1991). The efficiency is the reciprocal of the product of the mean squared error of $\hat{m}$ as an estimator of m and the total run length. Since the processes $\{Y_i(t) : t \geq s\}$ are being regarded as stationary, the mean squared error coincides with the variance of $\hat{m}$, and the efficiency is

$$e(n, t) = \frac{1}{\text{Var}\,(\hat{m}(n, t))n(s + t)} = \frac{1}{(2(t + s)/t^2) \int_0^t (t - u)R(u)du}. \tag{6}$$

REMARK. (2.4) We are using the total run length for the total cost of computation in our efficiency criterion. Of course, this is not always appropriate, especially with small sample sizes, but it seems to be a natural way to proceed in further investigation here.

Note that the variable n cancels out in (6), which is convenient for optimization purposes. Of course, this would also be true with the standard approach involving the budget constraint $n(s + t) \leq B$ if we simply chose n so that $n(s + t)$ is approximately equal to B; then the criterion would also be (6).

For convenience, we actually work with the *risk* $r(t) = 1/e(n, t)$, i.e,

$$r(t) = \frac{2(t + s)}{t^2} \int_0^t (t - u)R(u)du, \qquad t > 0, \tag{7}$$

with $r(0)$ and $r(\infty)$ defined by the limits of $r(t)$ as $t \to 0$ and $t \to \infty$, respectively. Of course, the risk $r(t)$ depends on s too, but s is assumed given.

To interpret the risk $r(t)$ in relation to the question about one long run versus multiple replications with a budget constraint, note that doing one long run ($n = 1$) is optimal (yields minimum variance Var $[\hat{m}(n, t)]$ subject to the budget constraint $n(s + t) \leq B$) for all s and all budgets B if $r(t)$ is nonincreasing in t.

REMARKS. (2.5) For any given s and run length budget B, having $r(t)$ nonincreasing is *sufficient* for $n = 1$ to be optimal with a budget constraint, but obviously *not necessary*. However, we believe that it is easier to investigate when $r(t)$ is nonincreasing than to directly investigate the constrained optimization problem. Moreover, there rarely is a hard budget constraint, so that the combinatorial difficulties it presents do not seem worth addressing.

(2.6) In practice there are other important criteria besides minimizing the risk; e.g., we might want reliable estimates for confidence intervals, but it is nevertheless interesting to see what happens with the risk criterion alone.

In general, the risk $r(t)$ is hard to evaluate because the covariance function $R(t)$ in (7) is hard to evaluate. (See Reynolds 1975, Abate and Whitt 1988, and Glynn 1989 for some positive results.) However, we often can evaluate the extreme cases $t = 0$ and $t = \infty$. Assuming that $\{Y_i(t) : t \geq 0\}$ has right-continuous sample paths, we have $\bar{Y}_i(0) = Y_i(s)$ from (3). With this convention, the case $t = 0$ corresponds to taking only a single observation from each replication after deleting the initial portion of length s. From (7) we obtain

$$r(0) \equiv \lim_{t \to 0} r(t) = sR(0) = s\sigma_\infty^2, \tag{8}$$

where $\sigma_\infty^2 \equiv \text{Var } Y(\infty)$ is the steady-state variance, and

$$r(\infty) \equiv \lim_{t \to \infty} r(t) = 2 \int_0^\infty R(u)du = \bar{\sigma}^2, \tag{9}$$

where $\bar{\sigma}^2 \equiv \lim_{t \to \infty} t \text{ Var } \bar{Y}_1(t)$ is the limiting time-average variance constant as in (5). Expressions for $\bar{\sigma}^2$ for various Markov processes are given in Whitt (1991) and references cited there.

From (8) and (9), we immediately see that there is no simple answer to the question of one long run versus multiple replications applying to all circumstances; the appropriate procedure depends on s and $R(t)$.

THEOREM 1. *$r(\infty) < r(0)$ if and only if $s > \bar{\sigma}^2/\sigma_\infty^2$.*

Hence, assuming that $\bar{\sigma}^2 > 0$ (which implies that $\sigma_\infty^2 > 0$), there are values of s such that $r(\infty) < r(0)$ and other values such that $r(0) < r(\infty)$. Thus, for given $R(t)$, *the answer depends on s*. Indeed, for small s, independent replications are desirable.

It is very significant that $r(t)$ typically converges to the positive limit $r(\infty) = \bar{\sigma}^2$ as $t \to \infty$. This convergence implies that the risk for a run of length t is essentially the same as for an infinitely long run if t is sufficiently large. This property is the basis for concluding that a moderate number of long replications is often just as efficient as one even longer run when one long run is most efficient; see Examples 5.1 and 6.1.

From (7) we can also draw a few other elementary conclusions. First, for each t, the risk $r(t)$ is nondecreasing in s. The risk $r(t)$ is also nondecreasing in $R(t)$; i.e., if $R_1(t) \leq R_2(t)$ for all t, then $r_1(t) \leq r_2(t)$ for all t. By considering the case $R_1(0) = R_2(0)$ and $0 \leq R_1(t) \leq R_2(t)$, we can conclude that one long run becomes more efficient ($r(\infty)$ decreases while $r(0)$ remains unchanged) when a positive covariance function $R(t)$ decays more rapidly.

We can also calculate the derivative of the risk, obtaining

$$r'(t) = 2t^{-2}\int_0^t uR(u)du + 2st^{-3}\int_0^t (2u - t)R(u)du. \tag{10}$$

In general, we need to know the covariance function $R(t)$ in order to determine the form of the risk function $r(t)$. However, we can obtain some general results for a large class of covariance functions.

THEOREM 2. (a) *Suppose that* $R(t) = R(0) + \int_0^t R'(u)du$ *with* $R(t) > 0$ *and* $R'(t) < 0$ *for all* t. *Then*

$$\int_0^t (2u - t)R(u)du < 0 \qquad \text{for all } t, \tag{11}$$

so that there is a function $s(t)$ *with* $0 < s(t) < \infty$ *for* $0 < t < \infty$ *such that* $r'(t) < 0$ *for all* $s > s(t)$ *and* $r'(t) > 0$ *for all* $s < s(t)$.

(b) *In addition,* $s(t)$ *is nonincreasing in* t *if and only if*

$$t^2R(t)\int_0^t (t - u)R(u)du \le 2\left(\int_0^t uR(u)du\right)^2, \tag{12}$$

in which case $r(t)$ *is unimodal with maxima at all* t *such that* $s(t) = s$. *When this holds, the minimum risk is always attained by* $t = 0$ *or* $t = \infty$.

PROOF. For (a), use integration by parts with (11) to obtain

$$\int_0^t (2u - t)\, R(u)du = -\int_0^t (u^2 - tu)R'(u)du < 0.$$

Hence, the first term on the right in (10) is always positive, while the second term is always negative. Hence, the function $s(t)$ exists with the claimed property. For (b), apply (10) to obtain

$$s(t) = \frac{t\int_0^t uR(u)du}{\int_0^t (t - 2u)R(u)du}, \qquad t > 0. \tag{13}$$

Differentiating (13), we obtain (12). We obtain the unimodality from (a) and the nonincreasing property of $s(t)$. □

REMARKS. (2.7) A sufficient condition for $R(t) \ge 0$ for all t is to have the random variables $Y(t)$, $t \ge 0$, be associated; see p. 29 of Barlow and Proschan (1975). A sufficient condition for $R(t) \ge 0$ and $R'(t) \le 0$ is for the underlying stochastic process $\{X(t) : t \ge 0\}$ to be a Markov process with stochastically monotone transition function and for the function f in $Y(t) = f(X(t))$ to be monotone; see Daley (1968) and p. 71 of Stoyan (1983).

(2.8) To see that (12) is not satisfied under the conditions of Theorem 2(a) alone, let $R(t) = Kt^{-5/2}$ for $t \ge t_0$ and $R(t) = R(t_0) + (t_0 - t)$ for $0 \le t < t_0$. Then $t^3R(t)\int_0^t R(u)du \to \infty$ as $t \to \infty$, while the other terms in (12) converge to finite limits. Moreover, §5.4 and Table 1 show that (12) need not be satisfied when the covariance function $R(t)$ is completely monotone, i.e., a mixture of exponentials.

EXAMPLE 2.1. To quickly see that t for which $0 < t < \infty$ can be optimal and that $R(t)$ need not be nonnegative, let $Y(k + t) = m + (-1)^k Z(t)$, $0 \le t < 1$, for all nonnegative integers k, where $Z(t)$ is a stationary process with $0 < E[Z(t)^2] < \infty$. Then, for all $k \ge 0$, $R(2k + 1) = -\text{Var}\,[Z(t)]$ and $\bar{Y}(2k) = \text{Var}\,[\bar{Y}(2k)] = r(2k) = 0$. One run with $s = 0$ and $t = 2$ achieves zero variance. □

Unfortunately, however, it seems difficult to find the minimum risk, even if we assume that $R(t) \geq 0$ and $R'(t) \leq 0$ for all t, neither of which have to hold. In the following sections we consider several examples, all of which satisfy these properties. For these examples, one of the extreme strategies $t = 0$ or $t = \infty$ is good, if not optimal. Hence, it seems that Theorem 1 provides useful guidance in the general case. Lacking detailed information about $r(t)$ for $0 < t < \infty$, we would tend to use one long run if $s > \bar{\sigma}^2/\sigma_\infty^2$. If $s \leq \bar{\sigma}^2/\sigma_\infty^2$, then we would first check to see if s is large enough so that we can regard $Y(t)$ as being approximately in steady state; if it is, then we would tend to use many replications of much shorter runs.

3. The Exponential Covariance Model

In the spirit of Fishman (1972) and Kelton and Law (1984), we consider concrete examples to do further analysis. In particular, *throughout this section we assume that the covariance function decays exponentially*, i.e., we assume that

$$R(t) = \sigma_\infty^2 e^{-\alpha t}, \qquad t \geq 0, \tag{14}$$

for $\alpha > 0$. We focus on (14) to see what happens in one special case, but also to provide an approximate analysis for other cases. This assumption is satisfied exactly by the M/M/∞ queue considered in §5.1 and its heavy-traffic limit, the Ornstein-Uhlenbeck diffusion process. A stationary process $\{Y(t) : t \geq s\}$ has the exponential covariance function in (14) if and only if the process is Markov in the wide sense; see p. 233 of Doob (1953). More generally, (14) is consistent with the asymptotic behavior of many covariance functions; e.g., for finite-state Markov chains (assuming that the eigenvalue having second largest real part has multiplicity one). The theorems then say that $e^{\alpha t}R(t)$ converges to a positive constant as $t \to \infty$.

From (7) and (14), we can express the risk as

$$r(t) = 2\sigma_\infty^2(t + s)\frac{(e^{-\alpha t} - 1 + \alpha t)}{(\alpha t)^2}, \qquad t > 0. \tag{15}$$

Let $t^*(s)$ be the value of t (including ∞) yielding the minimum value of $r(t)$. We will show that $t^*(s)$ is always 0 or ∞ under (14), with $t^*(s) = 0$ if and only if $s \leq 2/\alpha$. Formula (15) quantifies the inefficiency of a suboptimal strategy. We will show that the simulation strategy does not matter much when s is near $2/\alpha$, but that the simulation strategy is very important when s is not near $2/\alpha$.

The shape of $r(t)/\sigma_\infty^2$ as a function of t and $s = r(0)/\sigma_\infty^2$ is depicted in Figure 1. Time t is measured on the horizontal axis, while the normalized risk $r(t)/\sigma_\infty^2$ is measured on the vertical axis. By (8), $s = r(0)/\sigma_\infty^2$, so different curves are obtained by using (15) with different values of s. For each s, $r(t)/\sigma_\infty^2 \to \bar{\sigma}^2/\sigma_\infty^2$ by (9). We are interested in the values of t that yield the minimum for each curve. The actual shapes of the curves are justified below.

REMARK. (3.1) To better understand (14), it is helpful to note that there is an invariance under changes of scale by α; i.e., if $t = t'/\alpha$ and $s = s'/\alpha$, then

$$\frac{\alpha r(t)}{\sigma_\infty^2} = 2(t' + s')\frac{(e^{-t'} - 1 + t')}{(t')^2}.$$

For large t', we have

$$\frac{\alpha r(t)}{\sigma_\infty^2} \approx \frac{2(t' + s')(t' - 1)}{(t')^2} \approx 2 + \frac{(s' - 1)}{t'}.$$

To do further analysis, it is convenient to express $r(t)$ as

$$r(t) = 2\sigma_\infty^2(sf(t) + g(t)), \tag{16}$$

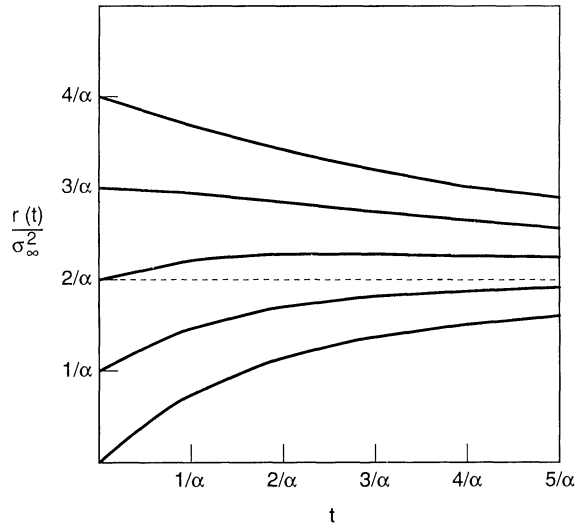

FIGURE 1. The normalized risk $r(t)/\sigma_\infty^2$ as a function of $s = r(0)/\sigma_\infty^2$ and t for the exponential covariance model in (14).

where $f(t) = (e^{-\alpha t} - 1 + \alpha t)/(\alpha t)^2$ and $g(t) = tf(t)$. The first two derivatives of f and g are

$$f'(t) = \frac{(2 - \alpha t) - (2 + \alpha t)e^{-\alpha t}}{\alpha^2 t^3},$$

$$f''(t) = \frac{-6(1 - e^{-\alpha t}) + 2\alpha t(1 + 2e^{-\alpha t}) + (\alpha t)^2 e^{-\alpha t}}{\alpha^2 t^4},$$

$$g'(t) = \frac{1 - e^{-\alpha t} - \alpha t e^{-\alpha t}}{\alpha^2 t^2},$$

$$g''(t) = \frac{-2(1 - e^{-\alpha t}) + 2\alpha t e^{-\alpha t} + (\alpha t)^2 e^{-\alpha t}}{\alpha^2 t^3}. \tag{17}$$

We summarize some of their basic properties in the following theorem.

THEOREM 3. *The function f is nonnegative, nonincreasing, and convex, while the function g is nonnegative, nondecreasing, and concave.*

PROOF. First, $f(t) \geq 0$ for all t is equivalent to $e^{-x} \geq 1 - x$ for all x, which is immediate for $x \geq 1$. For $0 \leq x < 1$, we see that $e^x = \sum_{k=0}^{\infty} x^k/k! < \sum_{k=0}^{\infty} x_k = 1/(1 - x)$ by comparing the coefficients. Second, $f'(t) \leq 0$ for all t is equivalent to $(2 - x) \leq (2 + x)e^{-x}$ for all x, which is immediate for $x \geq 2$. For $x < 2$, note that $e^{2x} < (1 + x)/(1 - x) = 1 + 2\sum_{k=1}^{\infty} x^k$ by comparing the coefficients. Third, $f''(t) \geq 0$ for all t is equivalent to

$$a(x) \equiv -6(1 - e^{-x}) + 2x(1 + 2e^{-x}) + x^2 e^{-x} \geq 0 \qquad \text{for all } x, \tag{18}$$

which holds because $a(0) = 0$, $a'(0) = 0$, and $a''(x) = x^2 e^{-x} > 0$. Fourth, $g(t) \geq 0$ because $g(t) = tf(t)$. Fifth, $g'(t) \geq 0$ for all t is equivalent to $1 - e^{-x} - xe^{-x} \geq 0$ for all x, which holds because $e^x \geq 1 + x$. Finally, $g''(t) \leq 0$ for all t is equivalent to

$$b(x) \equiv -2(1 - e^{-x}) + 2xe^{-x} + x^2 e^{-x} \leq 0 \qquad \text{for all } x, \tag{19}$$

which holds because $b(0) = 0$ and $b'(x) = -x^2 e^{-x} < 0$. □

COROLLARY. (a) *$r(t)$ is increasing in s;*
(b) *for each $t \geq 0$, there is an $s_1(t) \leq \infty$ such that $r''(t) \geq 0$ for $s \leq s_1(t)$ and $r'(t) \leq 0$ for $s \geq s_1(t)$;*
(c) *for each $t \geq 0$, there is an $s_2(t) \leq \infty$ such that $r''(t) \leq 0$ for $s \leq s_2(t)$ and $r''(t) \geq 0$ for $s \geq s_2(t)$.*

The following two theorems show that $t^*(s)$ must be either 0 or ∞. The numerical value 2.68 appearing in the next theorem is an approximate solution to a transcendental equation.

THEOREM 4. *$r(\infty) < r(t)$ for all $t > 0$, so that $t^*(s) = \infty$, if and only if $s \geq 2/\alpha$. For $s = 2/\alpha$, $r(\infty) = r(0)$ and the maximum relative inefficiency occurs at $t = 2.68/\alpha$; $r(2.68/\alpha)/r(\infty) = 1.14$.*

PROOF. If $s < 2/\alpha$, then $r(0) < r(\infty)$ by Theorem 1, because $r(\infty) = \bar{\sigma}^2 = 2\sigma_\infty^2/\alpha$, so that $r(t) < r(\infty)$ for all suitably small t by continuity. Suppose that $s \geq 2/\alpha$. Since $r(t)$ is increasing in s (see the Corollary to Theorem 3), it suffices to let $s = 2/\alpha$. Then

$$\frac{r(t) - r(\infty)}{2\sigma_\infty^2} = \frac{(\alpha t + 2)e^{-\alpha t} + \alpha t - 2}{\alpha^3 t^2} > 0, \tag{20}$$

because $e^{-2x} > (1 - x)/(1 + x)$ for $x > 0$, as shown in the proof of Theorem 3. By differentiating (20), we see that the maximum occurs at the unique positive zero of the transcendental function

$$\phi(t) \equiv e^{-\alpha t}(\alpha^2 t^2 + 3\alpha t + 4) + \alpha t - 4, \tag{21}$$

which is approximately $2.68/\alpha$. □

Theorem 4 says that worst choice of t is 14% worse than the optimal choice of t when $s = 2/\alpha$. (For example, this means that the variance $\mathrm{Var}[\hat{m}(n, t)]$ would be 14% greater with the worst strategy than with the best strategy given a common total simulation run length.) We interpret this result to mean that the simulation strategy does not matter too much when s is near $2/\alpha$. After the next theorem, we will show that the situation is very different when s is not near $2/\alpha$.

THEOREM 5. *$r(0) \leq r(t)$ for all t, so that $t^*(s) = 0$, if and only if $s \leq 2/\alpha$.*

PROOF. By Theorem 4, $r(0) > r(\infty)$ if $s > 2/\alpha$, so that $s \leq 2/\alpha$ is necessary. For $s = c/\alpha$, it suffices to show that

$$r(t) = 2\sigma_\infty^2 \frac{(\alpha t + c)(e^{-\alpha t} - 1 + \alpha t)}{\alpha^3 t^2} \geq r(0) = s\sigma_\infty^2 = \frac{c\sigma_\infty^2}{\alpha} \qquad \text{for } t \geq 0 \tag{22}$$

or, equivalently, after setting $x = \alpha t$,

$$e^{-x} - 1 + x \geq \frac{cx^2}{2(x + c)} \qquad \text{for } x \geq 0. \tag{23}$$

Since the right side of (23) is increasing in c, it suffices to consider $c = 2$. However, for $s = 2/\alpha$, $r(0) = 2\sigma_\infty^2/\alpha = \bar{\sigma}^2 = r(\infty)$, so that Theorem 4 implies that (22) holds. □

Consider the constrained optimization problem: min $r(t)$ subject to $n(s + t) \leq B$. Theorem 4 shows that it is optimal to use one long run ($n = 1$) for all sufficiently large budgets B if $s \geq 2\alpha$. However, one long run is optimal for this constrained problem for *all* budgets B when $r(t)$ is nonincreasing in t.

THEOREM 6. *For $t \geq 0$, $r(t)$ is nonincreasing in t if and only if $s \geq 3/\alpha$. For $s \geq 3/\alpha$, the maximum inefficiency is $r(0)/r(\infty) = \alpha s/2$.*

PROOF. First, it is easy to see that $r'(0) \le 0$ if and only if $s \ge 3/\alpha$. Hence $s \ge 3/\alpha$ is necessary. By the Corollary to Theorem 3, it suffices to consider $s = 3/\alpha$. Then $r'(t) \le 0$ for all t if and only if

$$c(x) \equiv 6 - 2x - 6e^{-x} - 4xe^{-x} - x^2e^{-x} \le 0 \qquad \text{for all } x,$$

which holds because $c(x) = -a(x)$ for $a(x)$ in (18). □

The following result, together with Theorem 6, allows us to characterize the extreme cases of maximum inefficiency.

THEOREM 7. *$r(t)$ is increasing in t for all $t \ge 0$, so that maximum efficiency occurs at $t = \infty$, if and only if $s \le 1/\alpha$. Then the maximum inefficiency is $r(\infty)/r(0) = 2/(\alpha s)$.*

PROOF. Since

$$\lim_{t\to\infty} t^2 r'(t) = \frac{2\sigma_\infty^2(1 - \alpha s)}{\alpha^2},$$

$r'(t)$ is eventually negative for $s > 1/\alpha$. For $s \le 1/\alpha$, it suffices to consider $s = 1/\alpha$ by the Corollary to Theorem 3. Then $\alpha^{-1}f'(t) + g'(t) \ge 0$ for all t if and only if

$$d(x) \equiv 2 - 2e^{-x} - 2xe^{-x} - x^2e^{-x} \ge 0 \qquad \text{for all } x,$$

which holds because $d(x) = -b(x)$ for $b(x)$ in (19). □

Theorems 6 and 7 show that $r(t)/r(t^*) \to \infty$ as $s \to 0$ and as $s \to \infty$.

REMARK. (3.2) From the results above, it follows that $\alpha^{-1} \le r(t)/\sigma_\infty^2 \le 3\alpha^{-1}$ provided that $\alpha^{-1} \le s \le 3\alpha^{-1}$. Moreover, $r'(t) \ge 0$ for all very small t in this region. In the proof of Theorem 7 we showed that $r'(t)$ is negative for all sufficiently large t when $s \ge 1/\alpha$. Indeed, for $\alpha^{-1} \le s \le 2\alpha^{-1}$, it can be shown that $r(t)$ is unimodal rising up from its minimum at $r(0)$ to a value above $r(\infty)$ and then coming down to $r(\infty)$. Also, for $2\alpha^{-1} \le s \le 3\alpha^{-1}$, $r(t)$ is unimodal and $2\alpha^{-1} \le r(t)\sigma_\infty^2 \le 3\alpha^{-1}$. Finally, $r(t)$ is decreasing and convex if and only if $s \ge 4/\alpha$.

4. Determining the Initial Portion to Delete

Provided that we regard the exponential covariance model as a reasonable approximation, §3 provides a basis for determining whether a single replication is more efficient than multiple replications. From Theorem 6, we conclude that a single replication minimizes $r(t)$ subject to $n(s + t) \le B$ for all budgets B in the exponential covariance model if $s \ge 3/\alpha$. We believe that a single replication is usually optimal in this context, because we believe that we usually have $s \ge 3/\alpha$. (We are thinking of queueing models, as in Whitt 1989 and §§5 and 6 here.) We investigate this further by considering what is an appropriate choice of s.

4.1. *A Criterion for Choosing s*

The quantity s to delete from each run is typically determined by bias considerations (as opposed to whether $Y(t) \overset{d}{=} Y(\infty)$ approximately for $t \ge s$), and we shall follow that approach. To consider the bias, we drop our stationarity assumption; i.e., we do not assume that $Y(t) \overset{d}{=} Y(\infty)$ for $t \ge s$. Given that we delete an initial portion of length s and make a run of length t as in (3), *the (absolute) bias* is

$$\beta(s, t) = t^{-1} \int_s^{s+t} |E[Y(u)] - m| \, du. \tag{24}$$

If t is very large, then the bias tends to be negligible, being dominated by $t^{-1}\bar{\beta}_0$ where $\bar{\beta}_s = \lim_{t\to\infty} t\beta(s, t)$, which is usually finite; e.g., see Whitt (1991). However, for t very small, the bias is approximately $|E[Y(s)] - m|$. Thus, for small t, we would want to

be sure that the *relative bias* $\beta(s, t)/m \approx |E[Y(s)] - m|/m$ is suitably small. The bias (24) is typically decreasing in both s and t. Given that $|E[Y(t)] - m|$ usually decreases rapidly at first and then more slowly as t increases, considerable bias reduction is obtained by making s positive but the bias reduction decreases as s increases.

To justify our assumption that $Y(t)$ is approximately stationary for $t \geq s$, we would like to choose s so that $|E[Y(t)] - m|/m$ is suitably small for all $t \geq s$. Note that this covers the relative bias for small t. Thus, it is natural to consider the criterion

$$s = \inf\{t \geq 0: \sup_{u \geq t} |E[Y(u)] - m|/m \leq p\}, \tag{25}$$

where p is a small positive number such as 0.01.

Indeed, (25) is the definition of the time when the stochastic process $Y(t)$ reaches steady state used by Gafarian, Ancker, and Morisaku (1978); see (6) of Wilson and Pritsker (1978). Note, however, that the mean being close to its steady-state limit does not guarantee that the full process is nearly stationary, i.e., that we approximately have $Y(t) \stackrel{d}{=} Y(\infty)$ for $t \geq s$. (Given that $Y(t) \stackrel{d}{=} Y(\infty)$ approximately, we expect that $\{Y(t): t \geq s\}$ is approximately a strictly stationary stochastic process, i.e., that $(Y(t_1), \ldots, Y(t_k)) \stackrel{d}{=} Y(t_1 + h), \ldots, Y(t_k + h))$ for all $h > 0$, integers $k > 0$ and $t_k > t_{k-1} > \cdots > t_1 > s$.) Criterion (25) seems reasonable as a time when bias has been reduced, but it is somewhat questionable as a criterion for $Y(t)$ truly reaching steady state. Nevertheless, we suggest (25) as a criterion for choosing s.

REMARKS. (4.1) It might be fruitful to modify (25) to require that the maximum distance between the probability distributions of $Y(u)$ and $Y(\infty)$ for $u \geq t$ is suitably small, but this would take us away from the relatively simple analysis here. For example, we could use the total variation metric, i.e., $\sup\{|P(Y(u) \in A) - P(Y(\infty) \in A)|\}$ where the supremum is over all measurable sets A.

(4.2) We have assumed that $Y(t) \Rightarrow Y(\infty)$ as $t \to \infty$ in §2, but the estimation problem is relevant even if only $E[\bar{Y}_1(t)] \to m \equiv E[Y(\infty)]$ as $t \to \infty$, for $\bar{Y}_1(t)$ in (3). For example, we could be simulating a periodic process. However, if we have a periodic process, then additional care should be given to the choice of s. Then we expect to obtain significant bias reduction by making the length of the undeleted portion a multiple of the period.

4.2. *The Second Exponential Assumption*

In the spirit of (14), now suppose that the mean approaches steady state exponentially; i.e.,

$$\frac{|m - E[Y(t)]|}{m} = \gamma e^{-\eta t}, \qquad t \geq 0. \tag{26}$$

The justification for (26) is essentially the same as for (14); e.g., it is exact for the M/M/∞ model in §5.1 and a representative approximation more generally.

Our criterion (25) dictates that we set $\gamma e^{-\eta s} = p$ for some suitably small p. Then $\bar{\beta}_s = mp/\eta$.

THEOREM 8. *With criterion* (25) *and both exponential assumptions* (14) *and* (26), $s \geq 3/\alpha$ *and* $r(t)$ *is decreasing if and only if* $p \leq \gamma e^{-3\eta/\alpha}$.

REMARK. (4.3) Perhaps the main idea in this paper is that one long run tends to be more efficient than multiple replications if the covariance function decays faster than the process approaches steady state, which here is measured by (25). Given both exponential assumptions (14) and (26), Theorem 8 expresses this idea by concluding that one long run is optimal for all budget constraints if the ratio of the decay rates, η/α, is suitably small (assuming that $p < \gamma$).

4.3. *The Common Parameter Assumption*

In many circumstances it is reasonable to assume that (14) and (26) hold with a common parameter, i.e., $\alpha = \eta$. Moreover, in (26) it is often reasonable to assume that $\gamma = 1$. For example, this is the case with the M/M/∞ queue in §5.1. If we are using approximations obtained via asymptotic representations as $t \to \infty$ (i.e., relaxation times; see Keilson 1979), then we often have $\alpha = \eta$; e.g., this is true for finite Markov chains. We obtain $\gamma = 1$ by considering the case $t = 0$; we are thinking of the mean $E[Y(t)]$ approaching the steady-state mean m monotonically or approximately so, such as occurs in most queueing systems when started empty. Then we can determine an appropriate strategy without knowing the value of the decay parameter α. Since $e^{-3} \approx 0.05$, we see that $p = 0.05$ is the cutoff point for $s \geq 3/\alpha$.

COROLLARY. *If $\alpha = \eta$ and $\gamma = 1$ with both exponential assumptions* (14) *and* (26), *then $s \geq 3/\alpha$ based on criterion* (25) *and $r(t)$ is decreasing if and only if $p \leq 0.05$.*

Since we believe that it is reasonable to have $p \leq 0.05$, we regard the Corollary to Theorem 8 as positive support for one long run. Indeed, in practice we might choose s even larger as a safety factor.

4.4. *Constructing Exponential Approximations*

In practice, the exponential assumptions (14) and (26) rarely hold exactly, but they may be reasonable approximations. As mentioned above, one way to justify both exponential approximations is via asymptotic expansions related to the relaxation time; i.e., it is often possible to show that (14) and (26) are asymptotically correct as $t \to \infty$. However, experience indicates that the relaxation time limits often do not provide very good approximation for the times t most relevant in practice; e.g., see Abate and Whitt (1987a, b, 1988), Aldous (1987), Aldous and Diaconis (1987), Diaconis and Shahshahani (1987), and Mitra and Weiss (1989). In particular, the stochastic processes often reach steady state in the sense of criterion (25) or in the sense that we approximately have $Y(t) \stackrel{d}{=} Y(\infty)$ much sooner than predicted by these exponential approximations. Hence, here we propose another exponential approximation, which like the relaxation-time method, is exact for true exponentials.

In particular, we suggest approximations for α and η in (14) and (26) based on the asymptotic variance and the asymptotic bias of the sample mean, i.e., $\bar{\sigma}^2$ in (9) and $\bar{\beta} \equiv \bar{\beta}_0$ after (24). As shown in Whitt (1991) and references cited there, $\bar{\sigma}^2$ and $\bar{\beta}$ are relatively easy to calculate for many one-dimensional Markov processes. As in Whitt (1989), we suggest approximating a more complicated process by a basic Markov process and then working with the asymptotic quantities of the approximating Markov process. (Of course, appropriate approximations are often not easy to determine.)

First, we consider the approximate covariance function of the assumed stationary process $\{Y(t) : t \geq s\}$. (It is significant that the limiting time-average variance $\bar{\sigma}^2$ typically depends on the original process only through its stationary version, so we obtain an appropriate match.) The exponential form (14) implies that $\bar{\sigma}^2 = 2\sigma_\infty^2/\alpha$. Hence, the proposed exponential approximation for the covariance function is (14) with decay rate

$$\alpha \approx 2\sigma_\infty^2/\bar{\sigma}^2. \tag{27}$$

Next, we consider the approximate bias function of the process $\{Y(t) : t \geq 0\}$, now no longer assumed to be stationary. We assume that we are given $\bar{\beta} \equiv \bar{\beta}_0$ as defined after (24). If (26) held with $\gamma = 1$, then $\bar{\beta} = m/\eta$. Hence, the proposed exponential approximation is (26) with $\gamma = 1$ and decay rate

$$\eta \approx m/\bar{\beta}. \tag{28}$$

From (24), (26), and (28), the resulting approximate bias is

$$\beta(s, t) = \frac{m(e^{-s} - e^{-(s+t)})}{\eta t}. \tag{29}$$

Applying Theorem 8 with criterion (25), we conclude from (27) and (28) that $r(t)$ is approximately decreasing if and only if $p \le e^{-3\eta/\alpha}$ or

$$p \le \exp\left[\frac{3m\bar{\sigma}^2}{2\bar{\beta}\sigma_\infty^2}\right]. \tag{30}$$

4.5. *Estimating the Initial Portion to Delete*

Our mathematical analysis in this section has primarily been aimed at determining what values of s should be appropriate, e.g., when we should have $s \ge 3/\alpha$ so that one long run is more efficient than multiple replications, using the exponential covariance model in §3. However, our analysis may also contribute to the long-standing problem of empirically determining an appropriate initial portion to delete; e.g., as discussed by Gafarian et al. (1978), Wilson and Pritsker (1978), Law and Carson (1979), Schruben, Singh, and Tierney (1983) and Heidelberger and Welch (1983). Two ways come to mind. First, analytic approximations may be calculated as described here and employed together with existing statistical estimation procedures to determine an appropriate s. For example, the analytical estimates based on (25), (26), and (28) can help avoid selecting s too small based on observed data.

The second approach is based on assuming (26) and possibly (28) with criterion (25), and then using statistical techniques to estimate the unknown parameters γ and η in (26) or m and $\bar{\beta}$ in (28). These ideas remain to be explored.

5. The M/G/∞ Queue

To illustrate the ideas in §2–§4, we now consider a few specific models and some numerical examples. In this section we consider the M/G/∞ queue; in §6 we consider reflected Brownian motion and the M/M/1 queue; and in §7 we consider two rather artificial extreme examples that show that one long run can be either much better or much worse than many independent replications.

Let $Y(t)$ be the number of busy servers at time t in an M/G/∞ queue starting out empty. The M/G/∞ model has infinitely many servers, a Poisson arrival process with rate λ and i.i.d. service times having a general distribution with cdf $G(t)$ and mean μ^{-1}. Suppose that we focus on the steady-state mean $m \equiv E[Y(\infty)]$. It is well known that the steady-state distribution is Poisson with $\sigma_\infty^2 = m = \lambda/\mu$. Moreover,

$$R(t) = \lambda \int_t^\infty [1 - G(u)]du \qquad \text{and} \tag{31}$$

$$\frac{m - E[Y(t)]}{m} = \mu \int_t^\infty [1 - G(u)]du; \tag{32}$$

see Benes (1957), p. 386 of Reynolds (1975), and p. 18 of Ross (1970).

5.1. *The M/M/∞ Model*

In the M/M/∞ special case (exponential service times) the exponential assumptions in (14) *and* (26) *are satisfied with* $\alpha = \eta = \mu$ *and* $\gamma = 1$, so that by the Corollary to Theorem 8, $r(t)$ is decreasing if and only if $p \le 0.05$.

EXAMPLE 5.1. Suppose that we simulate an M/M/∞ model starting out empty with $\lambda = 100$ and $\mu = 1$ to estimate the steady-state mean $m = \lambda/\mu = 100$. Suppose that we

want the absolute width of a 95% confidence interval to be 1.0 (so that the relative width is 0.01). Assuming that we make one long run starting in steady state, the appropriate run length is approximately $t = 4\bar{\sigma}^2(1.96)^2 = 3073$, which corresponds to 307,300 arrivals; see (9) of Whitt (1989).

Let us choose s according to criterion (25) using $p = 0.001$ to be conservative; then $s = 6.91$, which corresponds to 691 arrivals. *Note that s is much less than t*. Since $s \geq 3$, $r(t)$ is decreasing and one long run minimizes the risk $r(t)$ subject to $n(s + t) \leq B$ for all budgets B. From (24) and (32), we see that the bias $\beta(s, t)$ with these values of s and t is

$$\beta(s, t) = \beta(6.91, 3073) = 100(3073)^{-1} \int_{6.91}^{3079.9} e^{-u} du = 3.25 \times 10^{-5},$$

which clearly is negligible. (Even with $s = 0$, the bias would only be $100/3073 = 0.033$.) The optimal risk without budget constraint is $r(\infty) = \bar{\sigma}^2 = 200$. From (15), our risk with run length $t = 3073$ is $r(3073) = 200.38$, which is about 0.2% above optimal.

Now suppose, instead, that we decide to do independent replications with the same total run length $s + t = 3079.9$ and $s = 6.91$. If there are 10 independent replications, then $t = [3079.9 - 10(6.91)]/10 = 301.08$, which yields a bias $\beta(s, t) = 3.3 \times 10^{-4}$ and a risk $r(t) = 203.9$. Note that the bias is again negligible and the risk is only 1.96% higher than the optimal value $r(\infty) = 200$. On the other hand, if there are 100 independent replications, then $t = 23.89$, which yields a bias $\beta(s, t) = 4.19 \times 10^{-3}$ and a risk $r(23.89) = 247$. The bias is still negligible, but the risk is now 23.5% higher than the optimal value without a budget constraint, $r(\infty) = 200$.

Finally, suppose that we do 100 replications with the same total run length 3079.9 and $s = 0$, so that $t = 30.8$. Then the bias is $\beta(s, t) = 1/30.8 = 3.2$, which is larger than the width of the desired 95% confidence interval. Hence, bias does play a role with many replications, without initial deletion. The process clearly is not nearly stationary with $s = 0$, but if it were, then the risk would be 193.5, which is 3.2% *less* than the optimal value $r(\infty) = 200$ for $s \geq 3$.

In conclusion, for this example one long run and 10 independent replications are nearly equivalent with risk nearly equal to the optimal value without a budget constraint, $r(\infty)$. In both cases, we can set $s = 0$ without much harm, because the resulting bias is small. On the other hand, 100 replications is worse. With initial deletion of $s = 6.9$, the bias is negligible, but the risk is 23% higher; without initial deletion ($s = 0$), the bias is 3.2, larger than the width of the desired 95% confidence interval. For this example, we conclude that it does not matter much whether we delete an initial portion and it does not matter much whether we use one long run or multiple replications, unless the number of replications is large; only a very large number of multiple replications of relatively short run leads to difficulties. Then the estimation procedure performs poorly whether or not we delete an initial portion of each run.

5.2. *Stochastic Comparisons*

Formulas (31) and (32) imply that the exponential decay assumptions in (14) and (26) are *not* satisfied exactly if the service-time distribution is not exponential. However, the decay rates in (31) and (32) are identical, providing support for the common parameter assumption in §4.3.

From (31) and (32), we see that the approach to steady state in the M/G/∞ model gets slower, i.e., the functions $R(t)$ and $|m - E[Y(t)]|$ both increase, as the service-time distribution becomes more variable in the convex stochastic order; see p. 8 of Stoyan (1983). Recall that one random variable X_1 is less than or equal to another random variable X_2 in the *convex stochastic order*, denoted by $X_1 \leq_c X_2$, if $E[h(X_1)]$

$\leq E[h(X_2)]$ for all convex real-valued functions h. Since $h(x) = x$ and $h(x) = -x$ are both convex, we must have $E[X_1] = E[X_2]$ when $X_1 \leq_c X_2$. For example, $X_1 \leq_c X_2$ if X_i is normally distributed with mean m and variance σ_i^2 with $\sigma_1^2 \leq \sigma_2^2$. Suppose that X_1 and X_2 are two candidate service times with cdf's $G_1(t)$ and $G_2(t)$. Since $X_1 \leq_c X_2$ if and only if

$$\int_t^\infty [1 - G_1(u)]du \leq \int_t^\infty [1 - G_2(u)]du \qquad \text{for all } t, \tag{33}$$

see Stoyan (1983), it is immediate that $R_1(t) \leq R_2(t)$ and

$$|m - E[Y_1(t)]|/m \leq |m - E[Y_2(t)]|/m$$

for all t.

For fixed s, independent replications thus become more favorable as the distribution becomes more variable. ($R(t)$ decays more slowly, so that there is more to gain from replication.) However, the appropriate value of s when the system starts out empty increases as well, so that the overall effect is unclear. To see what happens more generally, we consider two specific nonexponential service-time distributions. The first is the deterministic distribution, which is less variable in the convex stochastic order than an exponential variable with the same mean, and the second is a hyperexponential (H_k, a mixture of k exponentials), which is more variable in the convex stochastic order than an exponential variable with the same mean.

5.3. *The $M/D/\infty$ Model*

For a deterministic distribution, $G(t) = 0$ for $t \leq \mu^{-1}$ and $G(t) = 1$ otherwise. Hence, $R(t) = (\lambda/\mu)(1 - \mu t)$ for $t \leq \mu^{-1}$ and $R(t) = 0$ for $t > \mu^{-1}$, so that $\bar{\sigma}^2 = \lambda/\mu^2$ and, from (7),

$$r(t) = \begin{cases} \dfrac{\lambda}{\mu}(t+s)\left(1 - \dfrac{\mu t}{3}\right), & 0 \leq t \leq \mu^{-1}, \\ \dfrac{\lambda}{\mu}(t+s)\left(\dfrac{1}{\mu t} - \dfrac{1}{3\mu^2 t^2}\right), & t \geq \mu^{-1}. \end{cases} \tag{34}$$

Hence, $r(\infty) < r(0)$ if and only if $s > \mu^{-1}$. However, by (32), $m = E[Y(t)]$ for all $t \geq \mu^{-1}$. Hence, $s \leq \mu^{-1}$; indeed if p is our cutoff probability, then $s = (1-p)\mu^{-1}$. Moreover, we see that *in this case the full process actually reaches steady state at time μ^{-1}*. Indeed, this is true in the strongest sense that $\{Y(u) : u \geq t\}$ is a stationary process for all $t \geq \mu^{-1}$. It is easy to see that $Y(t) \overset{d}{=} Y(\infty)$ for all $t \geq \mu^{-1}$, because $Y(t)$ has a Poisson distribution for all t and $E[Y(t)] = E[Y(\infty)]$ for $t \geq \mu^{-1}$.

Since the process reaches steady state by time μ^{-1}, s should not be larger than μ^{-1}. In this case, for any given p, $0 < p < 1$, multiple replications with $s = (1-p)\mu^{-1}$ and $t = 0$ is optimal. However, $r(0) = r(\infty)$ for $s = \mu^{-1}$, so the two extreme strategies are essentially equivalent. For $s = \mu^{-1}$, $r(t)$ in (34) is unimodal with maximum at $t = s = \mu^{-1}$; $r(\mu^{-1}) = 4\lambda/3\mu^2 = \frac{4}{3}r(0)$. Hence, for $s = \mu^{-1}$, the two extreme strategies are optimal and the maximum inefficiency compared with one of the extreme strategies is 33%.

The exponential approximations based on (27), (28), (31), and (32) are $\alpha = \eta = 2\mu$ and $\gamma = 1$, i.e., $R(t) \approx (\lambda/\mu)e^{-2\mu t}$ and $[m - EY(t)]/m \approx e^{-2\mu t}$. The approximation indicates that the rate of approach to steady state is as in an $M/M/\infty$ queue with service rate 2μ instead of μ, correctly reflecting the faster approach. The approximation indicates that we should use one long run and have $s > \mu^{-1}$ (assuming $s > 2/\alpha$). Note that one long run is optimal in the actual model for $s > \mu^{-1}$. On the other hand, if we let $s = \mu^{-1}$, then $s = 2/\alpha$, so that the approximation correctly predicts that the extreme strategies t

$= 0$ and $t = \infty$ are equivalent and optimal (Theorem 4). Then the maximum inefficiency is estimated to be 14% instead of the exact value of 33%.

5.4. *The $M/H_k/\infty$ Model*

A service-time distribution that is more variable in the convex stochastic order than an exponential distribution with the same mean is the H_k (hyperexponential) distribution, which is the mixture of k exponential distributions. Suppose then that $G(t) = \sum_{i=1}^{k} p_i(1 - e^{-\alpha_i t})$, where $\sum_{i=1}^{k} p_i = 1$, $\sum_{i=1}^{k} p_i \alpha_i^{-1} = \mu^{-1}$ and $p_i \geq 0$, $1 \leq i \leq k$. From (31),

$$R(t) = \sum_{i=1}^{k} p_i R_i(t) = \sum_{i=1}^{k} p_i \frac{\lambda}{\alpha_i} e^{-\alpha_i t}, \qquad t \geq 0, \tag{35}$$

where $R_i(t)$ is the covariance function for the $M/M/\infty$ queue with exponential service-time distribution having mean α_i^{-1}. As before, $r(0) = sR(0) = s\sigma_\infty^2 = s\lambda/\mu$, but $r(\infty) = \bar{\sigma}^2 = \lambda(c_s^2 + 1)/\mu^2$, where c_s^2 is the squared coefficient of variation (variance divided by the square of the mean) of the H_k service-time distribution. For an H_k distribution, $c_s^2 > 1$ so that $\bar{\sigma}^2$ is greater in the $M/H_k/\infty$ model than in the $M/M/\infty$ model with a common mean service time. (This qualitative property follows from the convex stochastic order comparison.)

Assuming that none of the probabilities p_i are too small, (32) is often well approximated by $p_i e^{-\alpha_i t}$ using the smallest α_i. Then we can apply (25) and (26) to determine an appropriate s and see if $r(0) = s\sigma_\infty^2 < \bar{\sigma}^2 = r(\infty)$. Given this s, we can easily calculate $r(t)$ using (14), (15), and (35).

EXAMPLE 5.2. For a concrete example, consider a balanced H_2 distribution with $\mu = 1$, $p_1 = 0.8873$, $\alpha_1 = 2p_1 = 1.775$, $\alpha_2 = 2p_2 = 0.2254$, and $c_s^2 = 4.0$; see (3.7) of Whitt (1982). Let the criterion be $p = 0.05$, so that s approximately satisfies $0.1127e^{-0.2254s} = 0.05$ or $s = 3.61$. (It is easy to verify that this is an excellent approximation.) With this value of s, $r(0) = 3.61\lambda < 5\lambda = \bar{\sigma}_\infty^2 = r(\infty)$, so we should *not* use one long run. (However, for $p = 0.01$, $s = 10.75$, and $r(0) > r(\infty)$.) Representative values of $r(t)$ for $s = 3.61$, 6.0, and 20.0 appear in Table 1. Unlike the exponential models in §§3 and 4 and the previous $M/G/\infty$ models, here we can have $0 < t^*(s) < \infty$. (This is consistent with the conclusions of Kelton and Law 1984 for a different model and a different criterion.) However, the evidence suggests that the best extreme strategy is pretty good; e.g., the inefficiencies of $t = 0$ when $s = 3.61$ and $t = \infty$ when $s = 6.0$ are evidently less than 5%. □

TABLE 1

The Risk $r(t)$ for Three Values of s for the $M/H_2/\infty$ Queue Example in §5.4

	The Risk $r(t)$		
Time t	$s = 3.61$	$s = 6.00$	$s = 20.00$
0	3.61λ	6.00λ	20.00λ
0.1	3.59λ	5.90λ	19.44λ
0.5	3.54λ	5.60λ	17.66λ
1.0	3.52λ	5.34λ	16.03λ
2.0	3.53λ	5.03λ	13.84λ
4.0	3.79λ	4.98λ	11.95λ
10.0	4.36λ	5.13λ	9.61λ
100.0	4.97λ	5.08λ	5.75λ
∞	5.00λ	5.00λ	5.00λ

In this case, the exponential approximations from §4.4 is equivalent to the M/M/∞ model with individual service rate $2\mu/(c_s^2 + 1)$, i.e., $R(t) = (\lambda/\mu)e^{-2\mu t/(c_s^2+1)}$. This properly reflects the slower approach to steady state of the M/H_k/∞ model compared to the M/M/∞ model with service rate μ.

6. RBM and M/M/1

Suppose that $Y(t)$ is regulated or reflecting Brownian motion (RBM) on the positive half line with drift coefficient -1 and diffusion coefficient 1, which has an exponential stationary distribution with mean $\frac{1}{2}$; see Abate and Whitt (1987a). As before, we focus on the steady state mean; then $m = \frac{1}{2}$ and $\sigma_\infty^2 = \frac{1}{4}$. As indicated in Whitt (1989), RBM is a useful approximation for stochastic processes associated with the standard GI/G/1 queue and related models. Explicit expressions for the covariance function $R(t)$ appear in Ott (1977) and Corollary 1 of Abate and Whitt (1988). From these expressions, we deduce that $\bar{\sigma}^2 = \frac{1}{2}$; see Corollary 9 to Theorem 1 of Abate and Whitt (1988). However, as indicated in Whitt (1989, 1991), the asymptotic variance is easy to derive directly.

From the above, we can immediately calculate the extreme risk values: $r(0) = s\sigma_\infty^2 = s/4$ and $r(\infty) = \bar{\sigma}^2 = \frac{1}{2}$. Hence, $r(\infty) < r(0)$ (without any approximations) if and only if $s \geq 2$. From (27) we obtain the rough exponential approximation with rate $\alpha = 1$, i.e., $R(t) \approx \frac{1}{4}e^{-t}$, $t \geq 0$.

Suppose that we consider RBM starting at zero. An explicit expression for the time-dependent mean $E[Y(t)]$ appears in Corollary 1.1.1 of Abate and Whitt (1987a). The asymptotic bias starting in 0 is just $m = \frac{1}{2}$ times the mean of the first-moment cdf there. Hence, from Corollary 1.3.4 of Abate and Whitt (1987a), $\bar{\beta} = \frac{1}{4}$ and we obtain the approximation $\eta = 1$ and $|m - E[Y(t)]|/m \approx e^{-t}$ from (28). Since $\alpha = \eta$ and $\gamma = 1$ using these approximations, we can apply the Corollary to Theorem 8 in §4.3 to conclude that with these approximations $s \geq 3/\alpha$ if and only if $p \leq 0.05$.

We can also use the exact formulas or more accurate approximations to do an analysis without the simplifying exponential assumptions. From (1.13) of Abate and Whitt (1987a), a good approximation for the distance from steady state for the mean is

$$\frac{m - E[Y(t)]}{m} \approx 0.724e^{-5.236t} + 0.276e^{-0.764t}, \qquad t \geq 0. \tag{36}$$

As shown in Table 4 of Abate and Whitt (1987a), it suffices to use only the second term of (36) provided that $0.001 \leq p \leq 0.1$, so that a better approximation for determining s is (26) with $\gamma = 0.276$ and $\eta = 0.764$. From (25), $s \approx (-\log p - 1.29)/0.764$. With $p = 0.05$, $s \approx 2.24$, as compared to $s \approx 3.0$ using the more elementary exponential approximation. For $p \leq 0.027$, $s \geq 3.0$, so that $p \leq 0.027$ is the condition for $r(t)$ to be decreasing using (25).

Now suppose that we specify $p = 0.05$ with the refined approximation above, and obtain $s = 2.24$. The rough exponential approximation for $R(t)$ still indicates that a single replication is optimal for all sufficiently long total run lengths. However, suppose that we use a better approximation. From Corollary 6 of Abate and Whitt (1988) and Table 5 of Abate and Whitt (1987a), a very good approximation for the covariance function is

$$R(t) \approx \frac{e^{-2t/3} + e^{-2t}}{8}. \tag{37}$$

Note that the dominant terms for larger t in (36) and (37) are $0.276e^{-0.764t}$ and $0.125e^{-0.667t}$, so that the rates of decrease in (36) and (37) are similar.

Table 2 displays the risk function for several values of t and $s = 2.24$, 3.00, and 6.00.

The first value $s = 2.24$ is based on the refined approximation (36) with $p = 0.05$; the second value $s = 3.00$ is based on the basic exponential approximation with $p = 0.05$; and the third value $s = 6.00$ reflects a larger value to have a safety factor. For all three values of s, $r(t)$ in Table 2 is decreasing in t. However, for $s = 2.24$ there is a relatively small advantage for large t of approximately 10%. The best (worst) strategy is always $t = \infty$ ($t = 0$); the risk of these strategies is exact. To put the times t in Table 2 in perspective, note that we need $t = 4\,\bar{\sigma}^2(1.96)^2/(0.01)^2 = 76{,}832$ for a 95% confidence interval to have absolute width 0.01 (relative width 0.02); see (9) of Whitt (1989). For such t, $r(t)$ is obviously very close to $r(\infty)$. The conclusions deduced about the M/M/∞ queue in Example 5.1 evidently can be made here as well.

A similar analysis applies to the queue length process (counting the customer in service, if any) and related descriptive processes in the M/M/1 queue starting out empty, drawing on Abate and Whitt (1987b, 1988). Indeed, the heavy-traffic behavior of the M/M/1 queue is essentially a scaled version of RBM above.

Good simple approximations for the correlation function, and thus the covariance function, of the M/M/1 queue length process appear in (3.7) and (3.8) of Abate and Whitt (1988); good simple approximations for the time-dependent mean appear in (2.1)–(2.7) of Abate and Whitt (1987b). Since the exact values of the steady-state variance and the asymptotic variance of the sample mean (assuming that the service rate is 1 and the arrival rate is ρ) are

$$\sigma_\infty^2 = \frac{\rho}{(1-\rho)^2} \qquad \text{and} \qquad \bar{\sigma}^2 = \frac{2\rho(1+\rho)}{(1-\rho)^4}, \tag{38}$$

$r(\infty) < r(0)$ if and only if $s > 2(1+\rho)/(1-\rho)^2$. The approximation in (2.3) of Abate and Whitt (1987b) yields

$$s \approx -2(1-\rho)^{-2}[c(\rho)\log(p[1+2c(\rho)]) \qquad \text{where} \tag{39}$$

$$c(\rho) = \frac{2+\rho+[5-(1-\rho)(5+\rho)]^{1/2}}{4} \approx \frac{(1+\rho^{1/2})^2}{(2+\rho^{1/4})}, \tag{40}$$

which is always greater than $2(1+\rho)/(1-\rho)^2$ if $p < 0.05$. For example, if $\rho = 0.70$, then $c(\rho) \approx 1.13$ and $r(\infty) < r(0)$ if and only if $p < 0.068$.

EXAMPLE 6.1. Suppose that we simulate an M/M/1 model starting out empty with arrival rate $\rho = 0.8$ and service rate 1 to estimate the steady-state mean queue length $m = \rho/(1-\rho) = 4.0$. From (38), $\sigma_\infty^2 = 20$ and $\bar{\sigma}^2 = 1800$. Suppose that we want the absolute width of a 95% confidence interval to be 0.08 (so that the relative width is 0.02).

TABLE 2

The Risk r(t) for Three Values of s for RBM Based on the Covariance Function Approximation in (37)

	The Risk $r(t)$		
Time t	$s = 2.24$	$s = 3.00$	$s = 6.00$
0.0	0.560	0.750	1.500
1.0	0.558	0.689	1.206
2.0	0.556	0.656	1.049
4.0	0.551	0.618	0.883
6.0	0.546	0.596	0.795
30.0	0.515	0.527	0.575
100.0	0.506	0.510	0.525
∞	0.500	0.500	0.500

If we could assume that we are making one long run starting in steady state, the appropriate run length is approximately $t = 4\bar{\sigma}^2(1.96)^2/(0.08)^2 = 4{,}321{,}800$, which corresponds to about 3.5 million arrivals; see (9) of Whitt (1989). Let us choose s according to criterion (25) with $\rho = 0.001$. Using (39), we obtain $s \approx 338.6$, which corresponds to 271 arrivals. Note that s is again very small compared to t. Also note that

$$r(\infty) = \bar{\sigma}^2 = 1800 < r(0) = s\sigma_\infty^2 = 6772.$$

As in Example 5.1, the risk $r(t)$ for $t = 4{,}321{,}800$ is nearly $r(\infty)$, and would remain so even with a moderate number of independent replications under the budget constraint $n(s + t) \le B$ where $B = 4{,}321{,}800 + 339$. Moreover, the bias is negligible, even with $s = 0$. In this case we use the approximations instead of the exact values to do the calculations; e.g., $R(t) \approx 10(e^{-0.0148t} + e^{-0.0442t})$. Even with 100 replications, $t = 42{,}879$ and $r(t) \approx 1801.5$, which is only 0.08% above $r(\infty) = 1800$. On the other hand, if we let $t = 0$, then we could have $n = 4{,}321{,}800/338.6 = 12{,}764$ replications. The variance of the estimator is then 0.001567 and a 95% confidence interval has width 0.155. This width is about two times what can be obtained with long runs.

7. Extreme Examples

For the models in §§5 and 6, it did not matter much whether we used one long run or multiple replications, provided that the individual runs were long enough. However, in general, either strategy can be much better than the other. To obtain additional insight, we consider two more rather extreme examples.

7.1. *When Many Replications Is Much More Efficient*

To show that many independent replications can be much more efficient than one long run, we give an example in which the process approaches steady state very rapidly, so that s can be very small, while the steady-state covariance function decays very slowly, so that the asymptotic variance of the sample mean is very large. (We aim for simplicity rather than realism, but we also want to avoid extremely pathological processes.) Our example involves a continuous-time Markov chain (CTMC) $\{Y(t) : t \ge 0\}$ with state space $\{0, 1, 2\}$ that begins in state 0. Suppose that we want to estimate the steady-state mean. Let the infinitesimal transition rates be $Q_{01} = Q_{02} = \lambda = -\frac{1}{2}Q_{00}$, $Q_{12} = Q_{21} = \mu$, $Q_{10} = Q_{20} = \lambda^{-1}$, and $Q_{11} = Q_{22} = -(\lambda^{-1} + \mu)$. To achieve the desired effect, let λ be very large and μ be very small. Then the process very quickly reaches state 1 or state 2, each with probability $\frac{1}{2}$. Then the time-dependent mean approaches steady state approximately at rate λ, while the steady-state covariance function $R(t)$ decays approximately at rate 2μ. Note the detailed analysis reduces to the two-state case (we can treat states 1 and 2 as one), for which all relevant formulas are given on p. 104 of Kemeny and Snell (1961). The steady-state probability vector is $(1/(1 + \lambda^2), \lambda^2/2(1 + \lambda^2), \lambda^2/2(1 + \lambda^2))$ and the time-dependent transition probabilities starting in 0 are

$$P_{01}(t) = P_{02}(t) = [\lambda/2(\lambda + \lambda^{-1})](1 - e^{-(\lambda+\lambda^{-1})t}).$$

Hence, $E[Y(t)]/E[Y(\infty)] = 1 - e^{-(\lambda+\lambda^{-1})t}$, so that the time-dependent mean starting in 0 approaches its limit monotonically at rate $\lambda + \lambda^{-1}$ (which is essentially λ since $\lambda \gg 1$). Suppose that we let s be the first time that $E[Y(t)]/E[Y(\infty)] = 1 - p$. Then $s = -(\log p)/(\lambda + \lambda^{-1})$, which is of order λ^{-1} for $\lambda \gg 1$. Indeed, $s \to 0$ as $\lambda \to \infty$.

For the remaining calculations, suppose that $\lambda = \infty$, corresponding to an instantaneous transition from state 0, so that the process is effectively stationary. Then the process behaves like the reduced CTMC with the state space $\{1, 2\}$ and time-dependent probabilities $P_{11}(t) = P_{22}(t) = \frac{1}{2}(1 + e^{-2\mu t})$. Note that the reduced CTMC reflects the original CTMC with large λ because the transition probabilities and the asymptotic variance of

the sample mean in the original model converge to those of the reduced model as $\lambda \to \infty$. For large finite λ, the process is not stationary. For the reduced model, the covariance function is $R(t) = \frac{1}{2}P_{11}(t) - \frac{1}{4} = \frac{1}{4}e^{-2\mu t}$, so that the time-average variance constant is $\bar{\alpha}^2 = \frac{1}{4}\mu$. Since μ is very small, $\bar{\alpha}^2$ is very large.

Finally, note that for any given budget B associated with the budget constraint $n(s + t) \leq B$, we can achieve a variance for the estimate of $\frac{1}{4}n$ by using n independent replications with $s = t = 0$. (Of course, we can do even better by setting $t = B/n$.) Indeed, since $s = 0$, (8) implies that $r(0) = 0 \leq r(t)$ for all t, so that $t = 0$ and infinitely many replications is optimal. In conclusion, in this example we have $r(0) = 0$ and $r(\infty) = \bar{\alpha}^2$ where $\bar{\alpha}^2$ is very large, so indeed many replications is much better than one long run.

7.2. *When Many Replications Is Much Less Efficient*

To show that independent replications can be much less efficient than one long run, we give an example in which the process approaches steady-state very slowly, so that s needs to be very large, while the steady-state covariance function decays very rapidly, so that the asymptotic variance of the sample mean is very small. Our example is a deterministic linear motion on $[0, y]$ starting at 0 plus a CTMC on the two states $y + 1$ and $y + 2$; i.e., let $Y(t) = t$, $0 \leq t < y$, and $Y(t) \in \{y + 1, y + 2\}$, $t \geq y$, with the $Y(t)$ entering states $y + 1$ and $y + 2$ each with probability $\frac{1}{2}$ at time $t = y$. On $\{y + 1, y + 2\}$ let the holding time in each state be exponential with mean μ^{-1} before transitioning to the other state. (A more realistic stochastic process with essentially the same behavior as the deterministic motion component is a diffusion process on $[0, y]$ with a reflecting boundary at 0 and an absorbing boundary at y, diffusion coefficient $\sigma^2(x) = \epsilon$ and drift coefficient $\mu(x) = 1$. For suitably small ϵ, the diffusion process is well approximated by the deterministic motion.)

The idea now is to choose y and μ very large, so that the process remains in the transient deterministic motion component a relatively long time before reaching the CTMC component, but the covariance of the CTMC component decays rapidly. Obviously the process reaches steady state precisely at time $t = y$, so it is appropriate to set $s = y$. Just as for the reduced model in §7.1, the covariance function for the CTMC is $R(t) = \frac{1}{4}e^{-2\mu t}$, but here μ is supposed to be large instead of small. In this example, the risk decreases from $r(0) = sR(0) = y/4$ to $r(\infty) = \bar{\sigma}^2 = \frac{1}{4}\mu$ as t increases.

In conclusion, in this example we have $r(\infty)$ much less than $r(0)$, so indeed one long run is much better than multiple replications.

8. Conclusions

The analysis here seems to justify our common practice with steady-state queueing simulations of performing one long run, deleting an initial portion. We try to be on the safe side and delete more than necessary, but we try to make the rest of the run sufficiently long that the portion of the run deleted is negligible, certainly less than 5%. Indeed, in Examples 5.1 and 6.1 it sufficed to delete 0.2% and 0.1%, respectively.

To estimate confidence intervals, we typically use nonoverlapping batch means with about 20 batches, acting as if they are independent and normally distributed; see Schmeiser (1982). However, if we want to have more confidence in our confidence intervals, and we are not too nervous about the time required to reach steady state, then we perform about 10 independent replications of slightly shorter runs. (This assures independence, but not the normal distribution, so that there still can be coverage problems.) We try to make these shorter runs sufficiently long that the efficiency is nearly the same as for one long run. As illustrated by Examples 5.1 and 6.1, the runs in steady-state queueing simulations often need to be surprisingly long. Thus the efficiency for 10 replications is often

essentially the same as for one long run. Moreover, the bias without initial deletion is often negligible. In this context, the only radically different alternative is a very large number of independent replications of much shorter runs. Our analysis tends to rule out this alternative, as did Kelton and Law's (1984).

Even though our analysis tends to favor long runs, we have seen that many replications of relatively short runs can be efficient when the process $Y(t)$ and its mean $E[Y(t)]$ approach steady state relatively quickly compared to the rate that the covariance function decays. Of course, to achieve this efficiency, we must recognize the rapid approach to steady state and let s be appropriately small. (We have offered some suggestions about how to choose s.) When s is relatively large, either because it needs to be or because we have insufficient information, one long run or a small number of replications tends to be more efficient. From our analysis, we conclude that the key to answering the replication question is the transient behavior of the stochastic process. The issue is whether or not the estimator with its specified initial condition approaches steady state faster than the covariance function decays.

We have suggested three levels of analysis for any particular case: First, given s, compare the exact values $r(0)$ and $r(\infty)$, which are relatively easy to calculate; see (8) and (9). Second, perform the exponential approximation based on the asymptotic variance and the asymptotic bias of the sample mean in §4.4. From §3, we know that $t = 0$ or $t = \infty$ will then be optimal, so that this step adds to the first primarily by providing an estimate for s. Third, do an analysis using (7) and (25) based on more reliable expressions for the covariance function $R(t)$ and the time-dependent mean $E[Y(t)]$. This final more reliable method was illustrated in §5–7.[1]

[1] I am very grateful to Professors Peter W. Glynn, Robert G. Sargent, and James W. Wilson for helpful comments.

References

ABATE, J. AND W. WHITT, "Transient Behavior of Regulated Brownian Motion, I: Starting at the Origin," *Adv. Appl. Probab.*, 19 (1987a), 560–598.

——— AND ———, "Transient Behavior of the M/M/1 Queue Starting at the Origin," *Queueing Systems*, 2 (1987b), 41–65.

——— AND ———, "The Correlation Functions of RBM and M/M/1," *Commun. Statist. Stochastic Models*, 4 (1988), 315–359.

ALDOUS, D., "Random Walks on Finite Groups and Rapidly Mixing Markov Chains," in *Seminaire de Probabilites XVII*, Springer-Verlag, New York, 1987, 243–297.

——— AND P. DIACONIS, "Shuffling Cards and Stopping Times," *Amer. Math. Monthly*, 93 (1987), 333–348.

BARLOW, R. E. AND F. PROSCHAN, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, 1975.

BENEŠ, V. E., "Fluctuations of Telephone Traffic," *Bell System Tech. J.*, 36 (1957), 965–973.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE, *A Guide to Simulation*, Second ed., Springer-Verlag, New York, 1987.

CHENG, R. C. H., "A Note on the Effect of Initial Conditions on a Simulation Run," *Oper. Res. Quart.*, 27 (1976), 467–470.

DALEY, D. J., "Stochastically Monotone Markov Chains," *Z. Wahrscheinlichkeitsheorie verw. Gebiete*, 10 (1968), 305–317.

DIACONIS, P. AND M. SHAHSHAHANI, "Time to Reach Stationarity in the Bernoulli-Laplace Diffusion Model," *SIAM J. Math. Anal.*, 18 (1987), 208–218.

DOOB, J. L., *Stochastic Processes*, John Wiley and Sons, New York, 1953.

FISHMAN, G. S., "Bias Considerations in Simulation Experiments," *Oper. Res.*, 20 (1972), 785–790.

GAFARIAN, A. V., C. J. ANCKER, JR., AND T. MORISAKU, "Evaluation of Commonly Used Rules for Detecting 'Steady-State' in Computer Simulations," *Naval Res. Logist. Quart.*, 25 (1978), 511–529.

GLYNN P. W., "Limit Theorems for the Method of Replication," *Commun. Statist.-Stochastic Models*, 3 (1987), 343–355.

———, "A Non-Rectangular Sampling Plan for Estimating Steady-State Means," *Proc. Sixth Army Conf. Appl. Math. and Computing*, 1988, 965–978.

GLYNN P. W., "The Covariance Function of a Regenerative Process," Department of Operations Research, Stanford University, 1989.

——— AND P. HEIDELBERGER, "Analysis of Initial Transient Deletion for Replicated Steady-State Simulations," Research Report 15259, IBM T. J. Watson Research Center, Yorktown Heights, NY, December 1989.

——— AND ———, "Experiments with Initial Transient Deletion for Parallel, Replicated Steady-State Simulations," Research Report 15770, IBM T. J. Watson Research Center, Yorktown Heights, NY, May 1990.

——— AND D. L. IGLEHART, "Conditions Under Which a Markov Chain Converges in Finite Time," *Prob. Engr. Inf. Sci.*, (1988), 377–382.

——— AND W. WHITT, "The Asymptotic Efficiency of Simulation Estimators," *Oper. Res.*, 39 (1991) to appear.

HEIDELBERGER, P., "Statistical Analysis of Parallel Simulations," 1986 *Winter Simulation Conf. Proc.*, J. Wilson and J. Henriksen (Eds.), IEEE Press, 1986, 290–295.

——— AND P. D. WELCH, "Simulation Run Length Control in the Presence of an Initial Transient," *Oper. Res.*, 31 (1983), 1109–1144.

KEILSON, J., *Markov Chain Models—Rarity and Exponentially*, Springer-Verlag, New York, 1979.

KELTON, W. D., "Random Initialization Methods in Simulation," *IIE Transactions*, 21 (1989), 355–367.

——— AND A. M. LAW, "An Analytical Evaluation of Alternative Strategies in Steady-State Simulation," *Oper. Res.*, 32 (1984), 169–184.

KEMENY, J. G. AND J. L. SNELL, "Finite Continuous Time Markov Chains," *Theor. Probability Appl.*, 6 (1961), 101–105.

LAW, A. M., "Confidence Intervals in Discrete Event Simulation: A Comparison of Replication and Batch Means," *Naval Res. Logist. Quart.*, 24 (1977), 667–678.

——— AND J. S. CARSON, "A Sequential Procedure for Determining the Length of a Steady-State Simulation," *Oper. Res.*, 27 (1979), 1011–1025.

MITRA, D. AND A. WEISS, "The Transient Behavior in Erlang's Model for Large Trunk Groups and Various Traffic Conditions," *Teletraffic Science for New Cost-Effective Systems, Networks and Services—ITC* 12, M. Bonatti (Ed.), North-Holland, Amsterdam, 1989, 1367–1374.

OTT, T. J., "The Stable M/G/1 Queue in Heavy Traffic and Its Covariance Function," *Adv. Appl. Probab.*, 9 (1977), 169–186.

PARZEN, E., *Stochastic Processes*, Holden-Day, San Francisco, 1962.

REYNOLDS, J. F., "The Covariance Structure of Queues and Related Processes—A Survey of Recent Work," *Adv. Appl. Probab.*, 7 (1975), 383–415.

ROSS, S. M., *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

SCHMEISER, B., "Batch Size Effects in the Analysis of Simulation Output," *Oper. Res.*, 30 (1982), 556–568.

SCHRUBEN, L. W., H. SINGH AND L. TIERNEY, "Optimal Tests for Initialization Bias in Simulation Output," *Oper. Res.*, 31 (1983), 1167–1178.

STOYAN, D., *Comparison Methods for Queues and Other Stochastic Models*, John Wiley and Sons, New York, 1983.

WHITT, W., "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," *Oper. Res.*, 30 (1982), 125–147.

———, "Planning Queueing Simulations," *Management Sci.*, 35 (1989), 1341–1366.

———, "Asymptotic Formulas for Markov Processes with Applications to Simulation," *Oper. Res.*, 39 (1991), to appear.

WILSON, J. R. AND A. A. B. PRITSKER, "A Survey of Research on the Simulation Startup Problem," *Simulation*, 31 (1978), 55–58.