EXPLOITING REGENERATIVE STRUCTURE TO ESTIMATE FINITE TIME AVERAGES VIA SIMULATION

Wanmo Kang, Perwez Shahabuddin* and Ward Whitt
Columbia University

We propose nonstandard simulation estimators of expected time averages over finite intervals $[0, t]$, seeking to enhance estimation efficiency. We make three key assumptions: (i) the underlying stochastic process has regenerative structure, (ii) the time average approaches a known limit as time t increases and (iii) time 0 is a regeneration time. To exploit those properties, we propose a *residual-cycle estimator*, based on data from the regenerative cycle in progress at time t, using only the data *after* time t. We prove that the residual-cycle estimator is unbiased and more efficient than the standard estimator for all sufficiently large t. Since the relative efficiency increases in t, the method is ideally suited to use when applying simulation to study the rate of convergence to the known limit. We also consider two other simulation techniques to be used with the residual-cycle estimator. The first involves *overlapping cycles*, paralleling the technique of overlapping batch means in steady-state estimation; multiple observations are taken from each replication, starting a new observation each time the initial regenerative state is revisited. The other technique is *splitting*, which involves independent replications of the terminal period after time t, for each simulation up to time t. We demonstrate that these alternative estimators provide efficiency improvement by conducting simulations of queueing models.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics—*renewal theory*; I.6.6 [**Computing Methodologies**]: Simulation and Modelling—*simulation output analysis*

General Terms: Algorithms, Experimentation, Performance, Theory

Additional Key Words and Phrases: efficiency improvement, variance reduction, regenerative processes, time averages

1. INTRODUCTION

Suppose that $X \equiv \{X(t) : t \geq 0\}$ is a continuous-time stochastic process with values in a general state space, and we want to use simulation to estimate the

*Wanmo Kang and Ward Whitt dedicate this paper to the memory of their friend, colleague and co-author - Perwez Shahabuddin - who passed away on November 17, 2005. This work was supported by National Science Foundation Grants DMS-02-2340 and DMI-0457095. Authors' address: Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699; email: {wk2003, ww2040}@columbia.edu

expected value of the associated *time average*

$$Y(t) \equiv t^{-1} \int_0^t f(X(u))\, du, \quad t \geq 0 \ , \tag{1}$$

for some real-valued function f. For example, $f(s)$ may be a reward rate earned in state s; then $E[Y(t)]$ is the expected average reward earned over the interval $[0, t]$. The *standard estimator* of the *expected time average*

$$\alpha(t) \equiv E[Y(t)]$$

is a sample mean from n independent replications, i.e.,

$$\hat{\alpha}_n(t) \equiv n^{-1} \sum_{j=1}^{n} Y_j(t) \ , \tag{2}$$

where $Y_j(t)$ is the random observation of $Y(t)$ in (1) from the j^{th} replication and a hat designates an estimator.

1.1 Alternative Estimators

We consider alternative estimators exploiting *regenerative structure* (see Smith [1955] and Chapter VI of Asmussen [2003] for background) and prove that these estimators are substantially more efficient than the standard estimator for sufficiently large t. What we do applies equally well to discrete-time processes, where the integral in (1) is replaced by a sum and t is a positive integer, but here we restrict attention to continuous-time processes.

In addition to assuming that X is a regenerative process, we assume that (i) the initial time 0 is a regeneration time and (ii) the expected time average converges to a known limit

$$\alpha \equiv \lim_{t \to \infty} \alpha(t) \ .$$

The initial condition can be relaxed, as we indicate in Section 12, but we strongly exploit knowledge of the limit α. It is common for α to be known, while $\alpha(t)$ is not. For example, $X(t)$ may be an irreducible finite-state semi-Markov process, with non-exponential transition-time distributions. The limiting steady-state distribution has an elementary form, given in (7.24) of Ross [2003], from which α is easy to compute, but the expected time average $\alpha(t)$ is not so easy to compute. Another large class of examples are product-form Markovian queueing networks. For them too, the steady-state distribution is readily available - through the product form - while the time-dependent behavior is relatively complicated.

We introduce two new estimation techniques to exploit the regenerative structure: (i) a residual-cycle estimator, defined in Section 3, and (ii) overlapping cycles, defined in Section 6. The *residual-cycle estimator* uses data from the regenerative cycle in progress at time t, using only the data *after* time t. The technique of *overlapping cycles*, paralleling the technique of overlapping batch means in steady-state estimation [Meketon and Schmeiser 1984, Pawlikowski 1990], uses multiple observations from each replication, starting a new observation each time the initial regenerative state is revisited. We also consider using *splitting* together with those two new techniques.

We prove that the residual-cycle estimator is unbiased and asymptotically more efficient than the standard estimator by a factor of t as $t \to \infty$. Through experiments, we show that all three techniques are effective, provided that t is sufficiently large, with the combination of all three techniques together performing best.

1.2 Estimating the Rate of Convergence

Our theorems and experiments show that our proposed methods are especially promising when t is relatively large, but for large enough t, the known limit α itself will be a good approximation for the expected time average $\alpha(t)$. However, we rarely know in advance whether any specific candidate t is large enough. Thus it is natural to rely on simulation if we want to estimate $\alpha(t)$ to within some specified statistical precision.

In fact, since our methods work so well for large t, they are especially well suited to study how fast $\alpha(t)$ converges to α as t increases. Accordingly, we show - under an extra moment condition in the regenerative setting - that

$$t(\alpha(t) - \alpha) \to \beta \quad \text{as} \quad t \to \infty \; ; \tag{3}$$

see Theorem 5.1. When the limit (3) holds, the constant β (or $-\beta$) is called the *asymptotic bias*; see Section 5. In Section 11 we show that our estimators of $\alpha(t)$ can be used together with regression to efficiently estimate β. That is helpful, because the exact form of β - given in (40) - is complicated in general.

For a large class of Markov processes, both α and β can be effectively computed; see Whitt [1992], so that we have the natural approximation

$$\alpha(t) \approx \alpha + \frac{\beta}{t} \; . \tag{4}$$

In those cases we can use simulation to investigate how well approximation (4) performs. When we are genuinely interested in the large-time behavior of the process, it is obviously very helpful to have estimators of $\alpha(t)$ that are highly efficient for large t. When t is large, samples of $Y(t)$ are more expensive to generate.

On the other hand, we recognize that many applications will have a fixed t. At the outset, we admit that the methods here require t not to be too small in order for the extra complexity required to implement these alternative estimators to be worth the effort. But how large must t be? We do not have a definitive answer, but our experiments give some insight. The relevant reference is the random length of a regenerative cycle, here denoted by τ_1. There is hope for efficiency gain if $t > E[\tau_1]$. Our limited experimental work indicates that a more precise reference may be the mean of the stationary-excess (or equilibrium residual lifetime) distribution of the cycle-time distribution, $E[\tau_1^2]/2E[\tau_1]$, which is larger than $E[\tau_1]$ when the distribution is more variable than an exponential distribution, and less otherwise. (See (27).) Our mathematical analysis and simulation experiments provide more information.

1.3 Notions of Efficiency

We introduce these alternative estimators in order to achieve *variance reduction*, but of course variance is not the only issue. We also need to take account of *computational effort*. Thus we seek to increase the *efficiency*, where efficiency is taken as being inversely proportional to the product of the sampling variance and the amount of labor expended in obtaining the estimate. The appropriateness of that definition may not be entirely obvious. See Glynn and Whitt [1992] and references therein for justification.

In our experiments we treat a small class of models, but we try to probe carefully into the efficiency issue. Since the different estimators require different simulation run lengths, we pay attention to *simulation run length*. In contrast to the standard estimator, the residual-cycle estimator uses data after time t, with the amount of data used being random, depending on the length of the remaining cycle.

But, simulation run length does not capture all of the computational effort. Thus, in addition to simulation run length, we also consider *average observed CPU time*. For these observed-CPU-time efficiency measures, our efficiency thus depends on our implementation. Allowing for the possibility of improved implementation, our reported efficiency gains involving average observed CPU time should be regarded as lower bounds on what possibly can be achieved. Of course, the care we have given to the meaning of efficiency is not the end of the story. For example, one might also consider the human programming effort.

1.4 Related Research

Nonstandard estimators of $\alpha(t)$ when α and the steady-state distribution of X are known were recently proposed by Glynn and Wong [1996], Wong, Glynn and Iglehart [1999] and Henderson and Glynn [2002] for the case in which the underlying stochastic process X in (1) is a Markov process. The first two papers exploit coupling, while the last exploits martingales. Our approach exploiting regenerative structure is quite different, and has the advantage that it applies to non-Markov processes as well as Markov processes. Our approach is relatively easy to apply as well. However, the other methods apply to a wider range of transient performance measures, going beyond $\alpha(t)$. These methods remain to be compared.

The idea of exploiting the remaining part of the last regeneration cycle to achieve efficiency improvement in simulations was originally due to Meketon and Heidelberger [1982], but they considered a different problem: They were concerned with estimating α instead of $\alpha(t)$. They showed that the last part of the regenerative cycle in progress at time t can be used to reduce the bias of an estimator of the limiting time average α. Other ways to exploit regenerative structure for simulation efficiency have been proposed by Calvin and Nakayama [1998, 2000].

Of course, the now classical way to exploit regenerative structure in simulation, aiming to estimate the limit α and related steady-state quantities, is through *regenerative simulation*, as reviewed in Section VI.2.d of Asmussen [2003] and Crane and Iglehart [1975]. That is not our concern here; we are not performing regenerative simulation in that way.

1.5 Organization of the Paper

We start in Section 2 by specifying what we mean by regenerative structure and reviewing the large-t asymptotic behavior of the standard estimator. In Section 3 we introduce the alternative residual-cycle estimator, and show that the residual-cycle estimator for $\alpha(t)$ is unbiased and asymptotically more efficient than the standard estimator by a factor of t as $t \to \infty$. In Section 4 we observe that the residual-cycle estimator can be regarded as a control-variable estimator, and show that the control is asymptotically optimal as $t \to \infty$. We show that the difference between the residual-cycle control weight (which turns out to be 1) and the optimal control weight can be bounded above by a quantity that decays as $1/\sqrt{t}$ as $t \to \infty$. In Section 5 we prove the asymptotic-bias limit in (3) in a regenerative setting, and obtain an explicit (but not necessarily tractable) expression for β.

In Section 6 we discuss the two additional simulation techniques to use together with the residual-cycle estimator in order to further improve its efficiency: overlapping cycles and splitting. Our experiments show that each method produces an order-of-magnitude improvement in efficiency when t is large.

In Section 7 we discuss ways to compare the alternative estimators, emphasizing the notion of efficiency, which includes computational effort as well as the variance. In Sections 8 and 9 we investigate how the different estimators perform for a relatively simple queueing model: $M/G/1/0$, which has one server and no extra waiting space. In Section 8 we obtain analytic results for the Markovian $M/M/1/0$ special case; in Section 9 we describe the results of extensive simulation experiments for $M/G/1/0$ models with three different service-time distributions: exponential (M), deterministic (D) and hyperexponential (a mixture of two exponentials, H_2).

In Section 10 we present a few results for a different model: the $M/G/5/0$ queue, which has five servers instead of only one. In Section 11 we discuss ways to estimate the asymptotic bias β in (3) using estimates of $\alpha(t)$. By using linear regression, we can verify that

$$\frac{1}{\alpha(t) - \alpha} \approx \frac{t}{\beta}$$

and estimate both β and $\alpha(t)$.

Finally, we make concluding remarks in Section 12. Additional material appears in an Internet supplement, Kang et al. [2005]. There we provide additional detail about our $M/G/1/0$ and $M/G/5/0$ experiments.

2. REGENERATIVE STRUCTURE

We assume that X is a *regenerative process* with respect to *regeneration times* $0 = T(0) \leq T(1) \leq \cdots$; e.g., see Chapters V and VI of Asmussen [2003]. The i^{th} cycle occurs between $T(i-1)$ and $T(i)$. The main idea is that, for each $i \geq 0$, the shifted pair of stochastic processes $(\{X(T(i)+u) : u \geq 0\}, \{T(i+k) - T(i) : k \geq 1\})$ is independent of the history before time $T(i)$ and has a probability distribution that is independent of i; i.e., everything indeed "starts over" at the regeneration times $T(i)$. By assuming that $T(0) = 0$, *we assume that a first full cycle starts at time* 0. This initial condition can be relaxed, but at the expense of some additional complexity, as we indicate in Section 12. The limit theorems extend, but there are more issues to consider in experiments. In this paper we restrict attention to this special initial condition.

Let $N(t)$ be the number of cycles completed by time t, i.e.,

$$N(t) \equiv \max\{i \geq 0 : T(i) \leq t\}, \quad t \geq 0 \; . \tag{5}$$

The regenerative structure implies that $\{N(t) : t \geq 0\}$ is a renewal counting process.

The key random variables associated with the regenerative cycles are

$$\begin{aligned}
\tau_i &\equiv T(i) - T(i-1),\\
U_i &\equiv \int_{T(i-1)}^{T(i)} f(X(u))\,du = \int_0^{\tau_i} f(X(T(i-1)+u))\,du,\\
\tilde{U}_i &\equiv \int_{T(i-1)}^{T(i)} |f(X(u))|\,du = \int_0^{\tau_i} |f(X(T(i-1)+u))|\,du,\\
W_i &\equiv \sup_{0\le t\le \tau_i} \left\{ \left| \int_0^t f(X(T(i-1)+u))\,du \right| \right\} \ . \qquad (6)
\end{aligned}$$

We assume that $P(\tau_1 > 0) > 0$ and $E[\tau_1] < \infty$.

As a consequence of our assumptions above, the four-tuples $(\tau_i, U_i, \tilde{U}_i, W_i)$, $i \geq 1$, are independent and identically distributed (IID). The random variables $\tilde{U}_i$ and W_i are included to provide regularity conditions; note that $0 \leq |U_i| \leq W_i \leq \tilde{U}_i$ almost surely. The variable $\tilde{U}_i$ is a cruder bound than W_i, but easier to work with.

2.1 Examples

A familiar example of a regenerative process is the queue-length process in a stable $GI/GI/s/r$ queue, having s servers, $r \leq \infty$ extra waiting spaces, and interarrival times and service times coming from independent sequences of independent and identically distributed (IID) random variables with general distributions; then times when arrivals come to an empty system are regeneration times. The intervals between these regeneration times are the familiar *busy cycles* in queueing. We assume that the expected busy cycle is finite. If s is large, then busy cycles can be very long, and that emptiness regeneration may occur only very rarely; indeed, without regularity conditions (allowing interarrival times and service times sharply bounded from above and below), the system might never empty, even for a stable system; see Whitt [1972].

If X is a positive-recurrent continuous-time Markov chain (CTMC), then transitions to any designated state can serve as a regeneration time for the process. Sometimes regeneration times need to be somewhat complicated. For example, for reflected Brownian motion on the positive half line with negative drift, with reflection at the origin, successive visits to the origin after first visiting state 1 constitute regeneration times; we cannot just work with successive visits to the origin, because there is almost surely no next visit a positive distance away, starting at the origin: Starting at the origin at time 0, the process almost surely visits the origin again infinitely often in an interval $(0, \epsilon)$ for any $\epsilon > 0$.

2.2 Classical Limit Theorems

As additional regularity conditions for estimating $\alpha(t)$ in the general regenerative case, we want $Y(t)$ to satisfy a *strong law of large numbers* (SLLN) and a *central limit theorem* (CLT) as $t \to \infty$, i.e.,

$$Y(t) \to \alpha \text{ w.p.1} \quad \text{as} \quad t \to \infty \qquad (7)$$

and

$$\sqrt{t}[Y(t) - \alpha] \Rightarrow \sigma N(0,1) \quad \text{as} \quad t \to \infty \ , \qquad (8)$$

where $N(0,1)$ is a standard (mean 0, variance 1) normal random variable, σ is a positive constant, and $\Rightarrow$ denotes convergence in distribution. From Glynn and Whitt [1993, 2002], we know that a necessary and sufficient condition for the SLLN in (7) is

$$E[W_1] < \infty \ , \tag{9}$$

while necessary and sufficient conditions for the CLT in (8) are

$$E[|U_1|] < \infty \quad \text{and} \quad \mathsf{Var}(U_1 - \alpha\tau_1) < \infty \ , \tag{10}$$

where Var is the variance, in which case

$$E[U_1] = \alpha E[\tau_1] \quad \text{and} \quad \mathsf{Var}(U_1 - \alpha\tau_1) = \sigma^2 E[\tau_1] \ . \tag{11}$$

We assume that these assumptions hold. For (10), it suffices to have $\mathsf{Var}(\tau_1) < \infty$ and $\mathsf{Var}(\tilde{U}_1) < \infty$ (or, equivalently, $E[\tau_1^2] < \infty$ and $E[\tilde{U}_1^2] < \infty$), and we assume that to be the case. We also assume that $\sigma > 0$ to avoid the degenerate case.

We also assume that additional regularity conditions hold, so that the regenerative stochastic processes X and $f(X)$ have steady-state limiting distributions, i.e.,

$$X(t) \Rightarrow X(\infty) \quad \text{and} \quad f(X(t)) \Rightarrow f(X(\infty)) \quad \text{as} \quad t \to \infty \ ,$$

and the limit α arises as the steady-state mean, satisfying the classical *ratio formula*,

$$\alpha = E[f(X(\infty))] = \frac{E[U_1]}{E[\tau_1]} \ .$$

The key extra condition is that the distribution of τ_1 be *nonlattice*, which we assume; see Theorem 1.2 on p. 170 of Asmussen [2003]. *Throughout the rest of this paper we assume that all the assumptions above are in force.* We will also make additional moment assumptions as needed. Corresponding results hold in the lattice case, after obvious adjustments.

2.3 Asymptotics for the Standard Estimator

The regenerative structure enables us to describe the asymptotic behavior of the standard estimator as $t \to \infty$ for any n. Given the SLLN in (7) and the CLT in (8), all that remains is to establish appropriate *uniform integrability.* For background on uniform integrability, see Sections 3.2 and 4.5 of Chung [1974], the Appendix and Chapters I and II of Gut [1988], Section 3, especially Theorems 3.4-3.6, in Billingsley [1999] and the unpublished paper by Glynn and Iglehart [1987]. The following result is known, being contained in Glynn and Iglehart [1987], but we are unaware of a published reference, so we will provide the details. We will be using the same line of reasoning in later proofs.

The main consequence is the variance limit in (15), which implies that the variance of the standard estimator decays as $1/t$ as $t \to \infty$ for any n. We will show that the residual-cycle estimator does better by a factor of t.

Theorem 2.1. (*reference case*) *As $t \to \infty$,*

$$\sqrt{t}[\hat{\alpha}_n(t) - \alpha] \Rightarrow \frac{\sigma}{\sqrt{n}} N(0,1) \ , \tag{12}$$

$$\sqrt{t}[\alpha(t) - \alpha] \to \sigma E[N(0,1)] = 0 \ , \tag{13}$$

$$\sqrt{t}[\hat{\alpha}_n(t) - \alpha(t)] \Rightarrow \frac{\sigma}{\sqrt{n}} N(0,1) \ , \tag{14}$$

$$t\mathit{Var}(\hat{\alpha}_n(t)) = \frac{t\mathit{Var}(Y(t))}{n} \to \frac{\sigma^2}{n} \tag{15}$$

for each n, where σ is the positive constant from (8) *and* (11).

Proof. The first limit (12) is an immediate consequence of the CLT (8). The CLT (8) and the *uniform integrability* (UI) of $\sqrt{t}(Y(t) - \alpha)$ imply (13). We use the assumed moment conditions $E[\tilde{U}_1^2] < \infty$ and $E[\tau_1^2] < \infty$ to show that $t(Y(t) - \alpha)^2$ is UI. Glynn and Iglehart [1987] focused directly on this issue. In particular, they establish the UI of $t^{p/2}(Y(t) - \alpha)^p$, and we will follow their argument, letting $p = 2$. We will make frequent reference to Gut [1988], which of course was not available to Glynn and Iglehart. The approach is to bound the quantity by the sum of more tractable quantities. We repeatedly use Lemmas 1.2 and 1.3 on p. 166 of the Appendix in Gut [1988]. Let Z_i be the centered version of U_i in (6), i.e., $Z_i \equiv U_i - \alpha\tau_i$; let $V_i \equiv \tilde{U}_i + |\alpha|\tau_i$ for $\tilde{U}_i$ in (6). Recall that $E[Z_i^2] = \sigma^2 E[\tau_1] < \infty$ by (11) and $E[V_i^2] < \infty$ by assumption. Then

$$\begin{aligned}\left(\int_0^t \{f(X(s)) - \alpha\} ds\right)^2 &\le \left(\left|\sum_{i=1}^{N(t)+1} Z_i\right| + \left|\int_t^{T(N(t)+1)} \{f(X(s)) - \alpha\} ds\right|\right)^2 \\ &\le 2\left|\sum_{i=1}^{N(t)+1} Z_i\right|^2 + 2V_{N(t)+1}^2 \ . \end{aligned} \tag{16}$$

The first term in (16) is a *random sum* of centered (mean 0) random variables with finite variance, where the random number of terms is governed by the renewal process $N \equiv \{N(t) : t \ge 0\}$. The UI for this random sum, divided by t, is covered by Theorems I.6.2 and I.6.3 on pp. 29-33 and Theorem II.5.1 plus (5.5) on p. 54 of Gut [1988], drawing on Chow et al. [1979].

It remains to treat the second term in (16). That is aided by the fact that V_i is nonnegative; see Theorem 3.6 of Billingsley [1999]. Incorporating the factor $1/t$, we can write

$$\frac{1}{t} V_{N(t)+1}^2 \le \frac{1}{t} \sum_{i=1}^{N(t)+1} V_i^2 \ . \tag{17}$$

Next we apply the SLLN, Wald's equation and the elementary renewal theorem to deduce, first, that

$$\frac{1}{t} \sum_{i=1}^{N(t)+1} V_i^2 \to \frac{E[V_1^2]}{E[\tau_1]} \quad w.p.1 \quad \text{as} \quad t \to \infty \tag{18}$$

and, second, that

$$\frac{1}{t} E\left[\sum_{i=1}^{N(t)+1} V_i^2\right] = \frac{1}{t} E[N(t) + 1] E[V_1^2] \to \frac{E[V_1^2]}{E[\tau_1]} \quad \text{as} \quad t \to \infty \ . \tag{19}$$

However, by Theorem 3.6 of Billingsley [1999], (18) and (19) imply that $\frac{1}{t}\sum_{i=1}^{N(t)+1} V_i^2$ is UI, which then implies $\frac{1}{t}V_{N(t)+1}^2$ is UI by (17). We have thus shown that $t(Y(t)-\alpha)^2$ is UI. That implies that its square root $\sqrt{t}(Y(t)-\alpha)$ is UI, from which we get (13).

Combining (12) and (13) with the converging-together theorem, Theorem 3.1 of Billingsley [1999], yields (14). Next note that $\mathsf{Var}(Y(t)) = E[(Y(t)-\alpha(t))^2] = E[(Y(t)-\alpha)^2] - (\alpha-\alpha(t))^2$. By the CLT in (8) and the continuous mapping theorem, Theorem 2.7 of Billingsley [1999], $t(Y(t)-\alpha)^2 \Rightarrow \sigma^2 N(0,1)^2$, where again σ is from (8) and (11). That, with the UI of $t(Y(t)-\alpha)^2$ established above implies that $E[t(Y(t)-\alpha)^2] \to \sigma^2$ as $t\to\infty$. The limit in (13) implies that $t(\alpha-\alpha(t))^2 \to 0$ as $t\to\infty$. These together imply (15). □

3. THE RESIDUAL-CYCLE ESTIMATOR

We now introduce the alternative estimator based on regenerative structure and knowledge of the limit α. We still use sample means from n independent replications, but we consider different statistics from each replication.

Our alternative estimator for $\alpha(t)$ is based on data from the last cycle at time t, including the portion of that cycle after time t, exploiting knowledge of the limit α. Let the *residual-cycle time* at t be

$$T_+(t) \equiv T(N(t)+1) - t \tag{20}$$

and let the *residual-cycle integral* at t be

$$I_+(t) \equiv \int_t^{T(N(t)+1)} f(X(u))\,du \ . \tag{21}$$

Let the *residual-cycle statistic* be

$$R(t) = \alpha + \frac{\alpha T_+(t)}{t} - \frac{I_+(t)}{t} \ . \tag{22}$$

Then the *residual-cycle estimator* of $\alpha(t)$ based on n replications is

$$\hat{\alpha}_n^r(t) \equiv n^{-1}\sum_{j=1}^{n} R_j(t) \ , \tag{23}$$

where $R_j(t)$ is the random observation of the residual-cycle statistic from the j^{th} replication and we again suppress n in the notation.

We first show that the residual-cycle estimator is unbiased.

THEOREM 3.1. (*unbiased*) *For each n and t, $E[\hat{\alpha}_n^r(t)] = \alpha(t)$.*

PROOF. Let $C(t) \equiv tY(t)$ be the associated *cumulative process.* The residual-cycle statistic naturally arises by looking at the cumulative process up to the end of the last cycle in progress at time t, in particular,

$$\begin{aligned} C(t) \equiv tY(t) &= \int_0^t f(X(u))\,du \\ &= \int_0^{T(N(t)+1)} f(X(u))\,du - \int_t^{T(N(t)+1)} f(X(u))\,du \end{aligned}$$

$$= \sum_{i=1}^{N(t)+1} U_i - \int_t^{T(N(t)+1)} f(X(u))du \tag{24}$$

$$= \sum_{i=1}^{N(t)+1} U_i - \alpha \sum_{i=1}^{N(t)+1} \tau_i + \alpha T(N(t)+1) - \int_t^{T(N(t)+1)} f(X(u))\, du \ .$$

Since $\alpha = E[U_1]/E[\tau_1]$, we can apply Wald's equation, p. 105 of Ross [1996], to obtain

$$E\left[\sum_{i=1}^{N(t)+1} U_i - \alpha \sum_{i=1}^{N(t)+1} \tau_i\right] = 0 \ . \tag{25}$$

Hence, taking expectations on both sides of equation (24), we see that

$$\begin{aligned} E[C(t)] &= tE[Y(t)] = t\alpha(t) = \alpha E[T(N(t)+1)] - E\left[\int_t^{T(N(t)+1)} f(X(u))\, du\right] \\ &= t\alpha + \alpha E[T(N(t)+1) - t] - E\left[\int_t^{T(N(t)+1)} f(X(u))\, du\right] = E[tR(t)] \ , \end{aligned}$$

where $R(t)$ is the residual-cycle statistic in (22). Dividing through by t in the last equation, completes the proof. □

When we consider efficiency, we must account for the fact that the residual-cycle estimator requires extra work, because it uses data after time t. We now show that the excess time and content after time t is asymptotically of order O(1), and thus asymptotically negligible compared to t. Indeed, as $t \to \infty$, the length of the residual cycle is distributed as the stationary-excess distribution of the cycle-length distribution. Let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution.

Theorem 3.2. (*excess*) *As* $t \to \infty$,

$$(T_+(t), I_+(t), t(R(t)-\alpha)) \Rightarrow (T_+(\infty), I_+(\infty), \alpha T_+(\infty) - I_+(\infty)) \quad in \quad \mathbf{R}^3 \ . \tag{26}$$

where the limit is a proper random vector, with $T_+(\infty)$ *having the familiar equilibrium residual lifetime (or stationary-excess) distribution associated with the distribution of* τ_1*; i.e.,*

$$P(T_+(\infty) \le x) = \frac{1}{E[\tau_1]} \int_0^x P(\tau_1 > y)\, dy \ . \tag{27}$$

As a consequence, for each n,

$$t(\hat{\alpha}_n^r(t) - \alpha) \Rightarrow n^{-1} \sum_{i=1}^{n} L_i \quad as \quad t \to \infty \ , \tag{28}$$

where L_i *are IID random variables with* $L_1 \stackrel{\mathrm{d}}{=} \alpha T_+(\infty) - I_+(\infty)$.

Proof. The second limit (28) follows from the first by (23) and the limit of the final term in (26) follows directly from (22), exploiting the continuous mapping theorem, Theorem 2.7 of Billingsley [1999], so it suffices to consider the first two terms in (26). Exploiting the assumption that the cycle length $T(1)$ has a nonlattice distribution, we can apply the *key renewal theorem*, p. 155 of Asmussen [2003], to deduce the *joint convergence* $(T_+(t), I_+(t)) \Rightarrow (T_+(\infty), I_+(\infty))$ as $t \to \infty$, where $(T_+(\infty), I_+(\infty))$ is a proper random vector, with $T_+(\infty)$ having the familiar equilibrium residual lifetime (or stationary-excess) distribution, which has mean $E[T_+(\infty)] = E[\tau_1^2]/2E[\tau_1]$ and second moment $E[T_+(\infty)^2] = E[\tau_1^3]/3E[\tau_1]$. We first apply the Portmanteau theorem, p. 15 of Billingsley [1999], to show that the desired limit is equivalent to the associated limit

$$E[g(T_+(t), I_+(t))] \to E[g(T_+(\infty), I_+(\infty))] \quad \text{as} \quad t \to \infty \tag{29}$$

for all continuous and bounded real-valued functions g. We then observe that, for any such function g, the expected value on the left side of (29) satisfies the renewal equation. Hence, it suffices to apply standard arguments in Asmussen [2003], exploiting direct Riemann integrability, just as in the proof of the closely-related Proposition 9 in Glynn and Whitt [1993]. □

We now show that the variance of the residual-cycle estimator decays as $1/t^2$ as $t \to \infty$ for any n, which is *faster than for the standard estimator by a factor of t.* For that, we need to impose some additional moment conditions.

THEOREM 3.3. (*variance asymptotics*) *If, in addition to the regenerative assumptions in force, $E[\tau_1^3] < \infty$ and $E[\tau_1 \tilde{U}_1^2] < \infty$, then*

$$t^2 \mathsf{Var}(\hat{\alpha}_n^r(t)) \to \frac{\mathsf{Var}(L_1)}{n} < \infty \quad as \quad t \to \infty \ , \tag{30}$$

for L_1 in Theorem 3.2.

PROOF. It suffices to separately treat the two components of $R(t)$ in (22): $T_+(t)$ and $I_+(t)$. (Again apply Lemma A.1.3 on p. 166 of Gut [1988].) In particular, we need to show that both $T_+(t)^2$ and $I_+(t)^2$ are UI, without further normalization, where we have already established the convergence in distribution in (26). Convergence in distribution extends immediately to the squares by the continuous mapping theorem, Theorem 2.7 of Billingsley [1999]. First, Gut [1988] has established convergence in distribution of $I_+(t)$ in his Theorem II.6.2, which implies convergence in distribution of $I_+(t)^2$ by the continuous mapping theorem. Gut [1988] has also established convergence of the moments $E[T_+(t)^2]$ in his Theorem II.6.3, under the assumed condition that $E[\tau_1^3] < \infty$. By Theorem 3.6 of Billingsley [1999], that implies UI of $T_+(t)^2$.

Hence it remains to treat $I_+(t)^2$, exploiting the condition $E[\tau_1 \tilde{Y}_1^2] < \infty$. Since $I_+(t)^2$ is nonnegative, it suffices to establish convergence of the moments. For that we can apply the key renewal theorem again. Let $A(t) \equiv E[I_+(t)^2]$ for $t \ge 0$. Then $A(t)$ satisfies the renewal equation

$$A(t) = a(t) + \int_0^t A(t-u)\, dF_{\tau_1}(u) \ , \tag{31}$$

where

$$a(t) = \int_{u=t}^{\infty} E\left[\left(\int_{s=t}^{u} f(X(s))\, ds \right)^2 \middle| \tau_1 = u \right] dF_{\tau_1}(u) \ . \tag{32}$$

In order to apply the key renewal theorem in Section V.4 of Asmussen [2003], we need $a(t)$ in (32) to be directly Riemann integrable (d.R.i.), for which conditions are given in Proposition 4.1 on p. 154 of Asmussen [2003]. For that purpose, we use

$$a_b(t) = \int_{u=t}^{\infty} E\left[\left(\int_{s=t}^{u} |f(X(s))|\, ds\right)^2 \middle| \tau_1 = u\right] dF_{\tau_1}(u) \; . \tag{33}$$

Note that $a_b(t)$ in (33) is nonincreasing in t. It is also Lebesgue integrable, because

$$\begin{aligned} \int_{t=0}^{\infty} a_b(t)\, dt &\le \int_{t=0}^{\infty} \left(\int_{u=t}^{\infty} E\left[\left(\int_{s=0}^{u} |f(X(s))|\, ds\right)^2 \middle| \tau_1 = u\right] dF_{\tau_1}(u)\right) dt \\ &= \int_{u=0}^{\infty} \left(\int_{t=0}^{u} E\left[\left(\int_{s=0}^{u} |f(X(s))|\, ds\right)^2 \middle| \tau_1 = u\right] dt\right) dF_{\tau_1}(u) \qquad (34) \\ &= \int_{u=0}^{\infty} uE\left[\left(\int_{s=0}^{u} |f(X(s))|\, ds\right)^2 \middle| \tau_1 = u\right] dF_{\tau_1}(u) = E[\tau_1 \tilde{Y}_1^2] < \infty \; . \end{aligned}$$

Thus $a_b(t)$ is d.R.i. by Proposition 4.1 (v) in Asmussen [2003]. But then $a(t)$ itself is d.R.i. by Proposition 4.1 (iv), because $0 \le a(t) \le a_b(t)$.

The key renewal theorem thus implies that $E[I_+(t)^2] \equiv A(t) \to E[I_+(\infty)^2]$ as $t \to \infty$, where $I_+(\infty)$ is the limit (convergence in distribution) in (26). Because of the nonnegativity, Theorem 3.6 of Billingsley [1999] implies that $I_+(t)^2$ is UI if $E[\tau_1 \tilde{U}_1^2] < \infty$. □

It would be nice to know the variance constants $\mathsf{Var}(L_1)$ in (30) and σ^2 in (8), (11) and (15), but there are no convenient expressions. It is natural to estimate these directly while performing the simulation.

4. A CONTROL-VARIABLE ESTIMATOR

Another way to arrive at the residual-cycle estimator of $\alpha(t)$ is to think of the steady-state limit α as a *control variable*; see Section 2.3 of Bratley et al. [1987]. With that in mind, let

$$\hat{Z}_n(t) = \frac{1}{n}\sum_{i=1}^{n} Z_i(t) \; , \tag{35}$$

where $Z_i(t)$ are IID with

$$Z_i(t) \stackrel{\mathrm{d}}{=} Z(t) \equiv t^{-1}\left[\sum_{i=1}^{N(t)+1} U_i - \alpha \sum_{i=1}^{N(t)+1} \tau_i\right] \; . \tag{36}$$

Given the relation (25), we can then think of $\hat{Z}_n(t)$ as a *control variate* for estimating $\alpha(t)$ by the standard estimator $\hat{\alpha}_n(t)$.

The associated *control statistic* would then be $\hat{\alpha}_{c,n}(t) \equiv \hat{\alpha}_n(t) - c_t\hat{Z}_n(t)$, where $\hat{Z}_n(t)$ is given in (35) and c_t is an an appropriate positive constant, called the *control weight*. The *optimal control estimator*, denoted by $\hat{\alpha}_{c^*,n}(t)$, is obtained using the *optimal control weight*

$$c_t^* \equiv \frac{\mathsf{Cov}(Y(t), Z(t))}{\mathsf{Var}(Z(t))} \ ; \tag{37}$$

see Section 2.3 of Bratley et al. [1987]. However, observe that

$$\hat{\alpha}_{c,n}(t) = \hat{\alpha}_n^r(t) \quad \text{when} \quad c_t = 1 \ , \tag{38}$$

where $\hat{\alpha}_n^r(t)$ is the residual-cycle estimator. Thus, the residual-cycle estimator $\hat{\alpha}_n^r(t)$ can be regarded as the control estimator with control weight $c_t = 1$, which is not the optimal control weight.

We now show that the residual-cycle estimator, when viewed as a control-variable estimator, is asymptotically equivalent to the optimal control-variable estimator as $t \to \infty$; we also determine a bound on the rate of convergence.

Theorem 4.1. (*control-weight bound*) *For all* $t > 0$,

$$t^{1/2}|c_t^* - 1| \le \sqrt{\frac{t\mathit{Var}(R(t))}{\mathit{Var}(Z(t))}} \to \frac{\mathit{Var}(L_1)}{\sigma^2} \quad \mathit{as} \quad t \to \infty \tag{39}$$

for each n, *where* L_1 *is in Theorem* 3.2.

Proof. First note that $Y(t) = Z(t) + R(t)$ and

$$\sqrt{t}Z(t) = \frac{\sum_{i=1}^{N(t)+1} (U_i - \alpha\tau_i)}{\sqrt{t}} \Rightarrow \sigma N(0,1)$$

for σ in (8) and (11); e.g., see Theorem 14.4 of Billingsley [1999] or Glynn and Whitt [2002]. By uniform integrability, it follows that $t\mathsf{Var}(Z(t)) \to \sigma^2$ as $t \to \infty$. Hence we can write

$$\begin{aligned} |c_t^* - 1| &= \left|\frac{\mathsf{Cov}(Y(t), Z(t))}{\mathsf{Var}(Z(t))} - 1\right| = \left|\frac{\mathsf{Cov}(Z(t) + R(t), Z(t))}{\mathsf{Var}(Z(t))} - 1\right| = \left|\frac{\mathsf{Cov}(R(t), Z(t))}{\mathsf{Var}(Z(t))}\right| \\ &\le \frac{\sqrt{\mathsf{Var}(R(t))\mathsf{Var}(Z(t))}}{\mathsf{Var}(Z(t))} = \sqrt{\frac{\mathsf{Var}(R(t))}{\mathsf{Var}(Z(t))}} \sim \sqrt{\frac{K}{t}} \ , \end{aligned}$$

for some constant K. In particular, from (15) and (30), $K = \sqrt{\mathsf{Var}(L_1)/\sigma^2}$. □

5. THE ASYMPTOTIC BIAS

In this section we justify the asymptotic bias limit in (3). The asymptotic form gives important insight about the relation between the unknown $\alpha(t)$ and the known α.

Theorem 5.1. (*asymptotic bias*) *If, in addition to the regenerative assumptions in force,* $E[\tau_1 \tilde{U}_1] < \infty$, *then*

$$t(\alpha(t) - \alpha) \to \beta \equiv \frac{\int_0^\infty a(t)\, dt}{E[\tau_1]} \quad \mathit{as} \quad t \to \infty \ , \tag{40}$$

where the limit in (40) *is finite and*

$$a(t) = -\int_{u=t}^\infty E\left[\int_{s=t}^u \{f(X(s)) - \alpha\} ds \,\middle|\, \tau_1 = u\right] dF_{\tau_1}(u) \ . \tag{41}$$

Proof. To justify (40), we modify (24) by subtracting αt:

$$\begin{aligned} C(t) - \alpha t &= \int_0^t \{f(X(u)) - \alpha\} du \\ &= \int_0^{T(N(t)+1)} \{f(X(u)) - \alpha\} du - \int_t^{T(N(t)+1)} \{f(X(u)) - \alpha\} du \\ &= \sum_{i=1}^{N(t)+1} Z_i - \int_t^{T(N(t)+1)} \{f(X(u)) - \alpha\} du \ , \end{aligned} \tag{42}$$

where $Z_i \equiv U_i - \alpha\tau_i$. Since $E[Z_i] = 0$, we can apply Wald's equation, p. 105 of Ross [1996], to obtain

$$E\left[\sum_{i=1}^{N(t)+1} Z_i\right] = E[N(t)+1]E[Z_i] = 0 \ . \tag{43}$$

Hence, taking expectations on both sides of equation (42), we see that

$$E[C(t) - \alpha t] = -E\left[\int_t^{T(N(t)+1)} \{f(X(u)) - \alpha\} du\right] \ . \tag{44}$$

Now we can apply the key renewal theorem to deduce that

$$E[C(t) - \alpha t] \to \beta \quad \text{as} \quad t \to \infty \ , \tag{45}$$

which is equivalent to (40). To carry out the analysis, we let

$$A(t) \equiv E[C(t) - \alpha t] = -E\left[\int_t^{T(N(t)+1)} \{f(X(u)) - \alpha\} du\right] \ . \tag{46}$$

Then $A(t)$ satisfies the renewal equation

$$A(t) = a(t) + \int_0^t A(t-u)\, dF_{\tau_1}(u) \ , \tag{47}$$

where $a(t)$ is given by (41).

Paralleling (33), let

$$a_b(t) = \int_{u=t}^{\infty} E\left[\int_{s=t}^{u} |f(X(s)) - \alpha|\, ds \middle| \tau_1 = u\right] dF_{\tau_1}(u) \ . \tag{48}$$

Note that $a_b(t)$ is nonincreasing and

$$\begin{aligned} \int_{t=0}^{\infty} a_b(t)\, dt &\le \int_{t=0}^{\infty} \left(\int_{u=t}^{\infty} E\left[\int_{s=0}^{u} |f(X(s)) - \alpha|\, ds \middle| \tau_1 = u\right] dF_{\tau_1}(u)\right) dt \\ &= \int_{u=0}^{\infty} \left(\int_{t=0}^{u} E\left[\int_{s=0}^{u} |f(X(s)) - \alpha|\, ds \middle| \tau_1 = u\right] dt\right) dF_{\tau_1}(u) \\ &= \int_{u=0}^{\infty} uE\left[\int_{s=0}^{u} |f(X(s)) - \alpha|\, ds \middle| \tau_1 = u\right] dF_{\tau_1}(u) \\ &\le \int_{u=0}^{\infty} uE\left[\int_{s=0}^{u} \{|f(X(s))| + \alpha\} ds \middle| \tau_1 = u\right] dF_{\tau_1}(u) \end{aligned}$$

$$= E[\tau_1 \tilde{U}_1] + \alpha E[\tau_1^2] < \infty \ , \tag{49}$$

so that $a_b(t)$ is Lebesgue integrable and thus d.R.i. by Proposition 4.1 (v) of Asmussen [2003].

We can then deduce that $a(t)$ itself is d.R.i. by breaking up $f(s) - \alpha$ into its positive and negative parts. Let $(x)^+ \equiv \max\{x, 0\}$ and $(x)^- \equiv -\min\{x, 0\}$. Then $a(t) = -a_+(t) + a_-(t)$, where

$$a_+(t) = \int_{u=t}^{\infty} E\left[\int_{s=t}^{u} (f(X(s)) - \alpha)^+ \, ds \middle| \tau_1 = u\right] dF_{\tau_1}(u) \tag{50}$$

and

$$a_-(t) = \int_{u=t}^{\infty} E\left[\int_{s=t}^{u} (f(X(s)) - \alpha)^- \, ds \middle| \tau_1 = u\right] dF_{\tau_1}(u) \ . \tag{51}$$

We see that $a_+(t)$ and $a_-(t)$ are both d.R.i. because $0 \le a_+(t) \le a_b(t)$ and $0 \le a_-(t) \le a_b(t)$. We can thus apply the key renewal theorem to $a_+(t)$ and $a_-(t)$ and put the results together to get the limit for $a(t)$. □

Explicit expressions for the asymptotic bias can be determined for special cases. e.g., for Markov chains and diffusion processes, see Whitt [1992] and references therein. For example, if the underlying stochastic process X is an irreducible finite-state CTMC with stationary probability vector π and fundamental matrix Z, then the asymptotic bias can be expressed directly as a matrix product by

$$\beta = \pi Z f^t \ , \tag{52}$$

where f^t is a representation of the function f in (1) as a column vector; see (33) of Whitt [1992]; see (49) of Whitt [1992] for the corresponding formula for a diffusion process. Equivalently, β can be represented as a solution of *Poisson's equation*; i.e., $\beta = x f^t$, where x is the unique solution to $xQ = -f^t + \bar{f} e^t$, where $\bar{f} \equiv \pi f^t$ and e^t is a column vector of $1's$; see Corollary 4 to Proposition 10 of Whitt [1992]. For birth-and-death processes and other skip-free Markov chains, β can be calculated recursively; see Remark 1 in Whitt [1992]. Thus, for these Markov processes, if t is large, then we may exploit this asymptotic relation as an approximation, and directly compute both α and β in order to generate a good approximation for $\alpha(t)$.

Moreover, the next term often is exponentially small, as in (74) in Section 8, so the asymptotic relation tends to yield a good approximation. For time-reversible irreducible finite-state CTMC's, which includes all birth-and-death processes, (74) is a refinement of the bias formula (40); see Section 3.2 of Keilson [1979]. More generally, the property is related to the notion of geometric ergodicity in Markov chains; see Chapters 15 and 16 of Meyn and Tweedie [1993].

6. OTHER EFFICIENCY-IMPROVEMENT TECHNIQUES

In this section we consider two other candidate efficiency-improvement techniques: overlapping cycles and splitting. We think of them being used together with the residual-cycle estimator. The overlapping-cycle technique can be considered with the standard estimator, but we found it to be ineffective in that way.

6.1 Overlapping-Cycle Estimators

This is another way to exploit the regenerative structure. We can extend each of the estimators above by developing new statistics, starting over at each regenerative time within each replication. Suppose that we do this for m cycles within each replication. The new *overlapping-cycle standard estimator* is

$$\hat{\alpha}_{o,m}(t) \equiv \hat{\alpha}_{n,o,m}(t) \equiv (mn)^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} Y_j(T_j(i-1), t) \ , \tag{53}$$

where n is the number of replications, m is the number of cycles per replication, $T_j(i-1)$ is the $(i-1)^{st}$ regeneration time in the j^{th} replication and $Y_j(T_j(i-1), t)$ is the statistic $Y(t)$ in (1) based on data after the $(i-1)^{st}$ regeneration time within the j^{th} replication, i.e.,

$$Y_j(T_j(i-1), t) \equiv t^{-1} \int_0^t f(X_j(T_j(i-1)+u))\, du, \quad t \geq 0 \ . \tag{54}$$

Because of the involved notation, henceforth we frequently omit the subscript n from our estimators. It is always understood that the estimators are based on n replications.

As with the residual-cycle estimator, the run length for each replication with the overlapping-cycle standard estimator is now random. The expected run length for each replication is $mE[T(1)] + t$, while the variance is $mVar(T(1))$. The overlapping-cycle standard estimator seems more promising than the ordinary standard estimator for situations in which t should be greater than the mean cycle time $E[T(1)]$. (Of course, we typically will not know $E[T(1)]$ in advance.) Under that condition, this alternative procedure produces many more samples per total run length, but improved efficiency is not automatic, because the m samples within each replication are dependent. On the other hand, if t is much less than $E[T(1)]$, than this procedure will tend to be less efficient than just using mn independent replications, because large portions of cycles are likely to be wasted.

We also consider an *overlapping residual-cycle estimator* of $\alpha(t)$ based on n replications and m cycles per replication, namely,

$$\hat{\alpha}^r_{o,m}(t) \equiv \hat{\alpha}^r_{n,o,m}(t) \equiv (mn)^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} R^o_{i,j}(t) \ , \tag{55}$$

where $R^o_{i,j}(t)$ is the overlapping-residual-cycle statistic starting from regeneration time $T_j(i-1)$ in the j^{th} replication; i.e., $R^o_{i,j}(t)$ is just the residual cycle statistic $R(t)$ based on the underlying stochastic process $\{X_j(T_j(i-1)+u) : u \geq 0\}$ which is regenerative with respect to the regeneration times $T_j(k) - T_j(i-1)$, $k \geq i-1$, where the subscript j indicates the j^{th} replication.

6.2 Splitting

As discussed in Hammersley and Handscomb [1964] and Bratley et al. [1987], simulation efficiency may often be enhanced by *splitting* a simulation run into independent sub-runs at some time, or after some state has been reached. If we split at time t, then p independent sub-runs are produced after time t, given the common history of the process realized up to time t. Clearly the portion up to time t is totally dependent in these p runs, but we only generate that initial portion once. We thus obtain more samples after time t, without having to re-simulate the initial portion.

In this paper, we consider splitting at time t to enhance the two residual-cycle estimators, $\hat{\alpha}^r(t)$ and $\hat{\alpha}^r_{o,m}(t)$, thereby obtaining estimators $\hat{\alpha}^r_{s,p}(t)$ and $\hat{\alpha}^r_{o,m,s,p}(t)$; e.g.,

$$\hat{\alpha}^r_{o,m,s,p}(t) \equiv \hat{\alpha}^r_{n,o,m,s,p}(t) \equiv (mnp)^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{k=1}^{p} R^{o,s}_{i,j,k}(t) \ , \tag{56}$$

where $R^{o,s}_{i,j,k}(t)$ is the k^{th} independent splitting starting at time t of the i^{th} overlapping-residual-cycle statistic starting from regeneration time $T_j(i-1)$ in the j^{th} replication; i.e., $R^{o,s}_{i,j,k}(t)$ is the k^{th} splitting of the residual cycle statistic $R(t)$ based on the underlying stochastic process $\{X_j(T_j(i-1)+t) : t \geq 0\}$ which is regenerative with respect to the regeneration times $T_j(l) - T_j(i-1)$, $l \geq i-1$, where the subscript j indicates the j^{th} replication.

We now investigate the effectiveness of splitting the residual-cycle estimator. The following analysis is standard, but important. To proceed, we now consider a general stochastic process $\{X(u) : u \geq 0\}$ and a filtration of sigma-fields (histories) generated by X, $\mathcal{F}_t \equiv \sigma(\{X(u) : 0 \leq u \leq t\})$. As before, define a residual run to be the process $\{X(u) : t \leq u \leq T(N(t)+1)\}$ and let the residual-cycle estimator $R(t)$ be as in (22). We generate p independent *residual runs* (splittings) contingent on the observation up to time t and obtain $R^s_i(t)$ for $1 \leq i \leq p$. Even though $R^s_i(t)$ and $R^s_j(t)$ are not independent for $i \neq j$, because they share a common portion up to time t, they are conditionally independent given $\mathcal{F}_t$.

The splitting residual estimator, based on a single replication of X on $[0,t]$, is defined by

$$\bar{R}^s(t) \equiv \frac{1}{p} \sum_{i=1}^{p} R^s_i(t) \ .$$

The conditional independence implies that

$$\mathsf{Var}\left(\frac{1}{p} \sum_{i=1}^{p} R^s_i(t) \ \middle| \ \mathcal{F}_t \right) = \frac{1}{p} \mathsf{Var}\left(R(t) \mid \mathcal{F}_t \right) \ , \tag{57}$$

so that

$$\begin{aligned} \mathsf{Var}\left(\bar{R}^s(t)\right) &= \mathsf{Var}\left(E\left[\bar{R}^s(t) \mid \mathcal{F}_t\right]\right) + E\left[\mathsf{Var}\left(\bar{R}^s(t) \mid \mathcal{F}_t\right)\right] \\ &= \mathsf{Var}\left(E\left[\frac{1}{p}\sum_{i=1}^{p} R^s_i(t) \ \middle| \ \mathcal{F}_t\right]\right) + E\left[\mathsf{Var}\left(\frac{1}{p}\sum_{i=1}^{p} R^s_i(t) \ \middle| \ \mathcal{F}_t\right)\right] \\ &= \mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + \frac{1}{p} E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right] \ . \end{aligned} \tag{58}$$

Since $\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right)$ and $E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]$ depend just on the stochastic process and are independent of p, we can clearly see the influence of increasing p. When we increase p, only the second term in (58) will decrease. The variance reduction achieved by splitting is

$$\frac{\mathsf{Var}\left(\bar{R}(t)\right)}{\mathsf{Var}\left(\bar{R}^s(t)\right)} = \frac{\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]}{\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + \frac{1}{p}E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]}\ , \tag{59}$$

so that the potential variance reduction is large, approaching a factor p, when the ratio

$$\frac{E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]}{\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right)} \tag{60}$$

is large, but not otherwise.

To proceed further, assume that the computational cost is proportional to the length of simulation horizon. Then the cost to get one split residual-cycle estimator with p splittings is proportional to $t + pE[T(N(t)+1) - t]$. To achieve a half-width HW, the total computational effort is proportional to

$$\frac{(t + pE[T(N(t)+1) - t])\mathsf{Var}\left(\bar{R}^s(t)\right)}{HW^2}\ .$$

To minimize this effort, it is sufficient to minimize

$$\begin{aligned} & (t + pE[T(N(t)+1) - t])\mathsf{Var}\left(\bar{R}^s(t)\right) \\ = \ & (t + pE[T(N(t)+1) - t])\left(\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + \frac{1}{p}E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]\right) \\ = \ & t\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + E[T(N(t)+1) - t]E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right] \\ & + pE[T(N(t)+1) - t]\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + \frac{1}{p}tE\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]\ . \end{aligned}$$

For any given t, we see that this is a *convex function* of p whose *minimum* is achieved at

$$p^*(t) = \sqrt{\frac{t}{E[T(N(t)+1) - t]} \times \frac{E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]}{\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right)}}\ . \tag{61}$$

Of course, $p^*(t)$ must be an integer, so we replace $p^*(t)$ in (61) by $\lfloor p^*(t) \rfloor$ or $\lfloor p^*(t) + 1 \rfloor$.

In applications we anticipate that the three expectation-and-variance terms in (61) will not change so rapidly with t, so that we might approximate the optimum number of splittings by $p^*(t) \approx \sqrt{Kt}$ for some constant K. We thus anticipate that $p^*(t)$ will grow with t proportional to $\sqrt{t}$. In practice, the constant K might be estimated by doing pilot simulation runs.

When t is large, the additional computational effort to perform splitting will be negligible, because $E[T(N(t)+1) - t]$ will be approximately constant and $t + pE[T(N(t)+1) - t]$ will be approximately t. Thus the critical factor for efficiency improvement from splitting is the ratio in (60), which of course depends on the application.

7. RELATIVE EFFICIENCY

In preparation for concrete examples, we now discuss ways to compare two estimators in the simulation experiments. Following the discussion so far, we look at the *ratio of the variances*, namely,

$$\mathcal{V}^r \equiv \mathcal{V}^r(t) \equiv \frac{\mathsf{Var}(\hat{\alpha}(t))}{\mathsf{Var}(\hat{\alpha}^r(t))} \tag{62}$$

and the associated estimated value

$$\widehat{\mathcal{V}}^r \equiv \widehat{\mathcal{V}}^r(t) \equiv \frac{\widehat{\mathsf{Var}}(\hat{\alpha}(t))}{\widehat{\mathsf{Var}}(\hat{\alpha}^r(t))} \ , \tag{63}$$

where a hat designates an estimator, the superscript r designates the residual-cycle estimator, and no superscript designates the standard estimator. (The subscript n is omitted throughout.) The limits in (15) and (30) imply that $\mathcal{V}^r(t)/t$ converges to a finite positive limit as $t \to \infty$ (for any n), showing the superiority of the residual-cycle estimator for large t.

However, the variance of the estimators is not the only issue, because the run length for the residual-cycle estimator is longer than the run length for the standard estimator, being $T(N(t)+1)$ instead of t. Moreover, the run length of the residual-cycle estimator is random. Thus, we really want to look at the *relative efficiency* of the two estimators, where the *efficiency* is taken as inversely proportional to the product of the sampling variance and the amount of labor expended in obtaining that estimate; see pp. 22 and 51 of Hammersley and Handscomb [1964] and Glynn and Whitt [1992].

Thus, we focus on the following *measure of relative efficiency*:

$$\mathcal{E}^r \equiv \mathcal{E}^r(t) \equiv \frac{\mathsf{Var}(\hat{\alpha}(t)) \times E[CT(t)]}{\mathsf{Var}(\hat{\alpha}^r(t)) \times E[CT^r(t)]} \tag{64}$$

and the associated estimated value

$$\widehat{\mathcal{E}}^r \equiv \widehat{\mathcal{E}}^r(t) \equiv \frac{\widehat{\mathsf{Var}}(\hat{\alpha}(t)) \times \widehat{CT}(t)}{\widehat{\mathsf{Var}}(\hat{\alpha}^r(t)) \times \widehat{CT}^r(t)} \ , \tag{65}$$

where $CT(t)$ and $CT^r(t)$ are the random amounts of CPU time used to obtain the standard and residual-cycle estimators, respectively, while $\widehat{CT}(t)$ and $\widehat{CT}^r(t)$ are the associated estimates, based on sample averages.

For simplicity, it is often convenient to use the expected total simulation run length as a surrogate for the amount of labor expended in obtaining the estimate. We thus also look at the following alternative measure of efficiency:

$$\mathcal{F}^r \equiv \mathcal{F}^r(t) \equiv \frac{\mathsf{Var}(\hat{\alpha}(t)) \times E[RL(t)]}{\mathsf{Var}(\hat{\alpha}^r(t)) \times E[RL^r(t)]} \ , \tag{66}$$

and associated estimated value

$$\widehat{\mathcal{F}}^r \equiv \widehat{\mathcal{F}}^r(t) \equiv \frac{\widehat{\mathsf{Var}}(\hat{\alpha}(t)) \times \widehat{RL}(t)}{\widehat{\mathsf{Var}}(\hat{\alpha}^r(t)) \times \widehat{RL}^r(t)} \ , \tag{67}$$

where $RL(t)$ and $RL^r(t)$ are the random simulation run lengths used to obtain the standard and residual-cycle estimators, respectively, while $\widehat{RL}(t)$ and $\widehat{RL}^r(t)$ are the associated estimates based on sample averages.

From the perspective of efficiency, the residual-cycle estimator is also more promising than the standard estimator as t gets larger, because then the random run length

$T(N(t)+1)$ of each replication will not differ greatly from t. Indeed, by Wald's equation and the elementary renewal theorem, we have

$$\frac{E[T(N(t)+1)]}{t} = \frac{E[N(t)+1]E[\tau_1]}{t} \to 1 \quad \text{as} \quad t \to \infty \ .$$

Indeed, the asymptotic relative efficiency favoring the residual-cycle estimator, as measured by $\mathcal{F}^r(t)$, is the same as the relative variance, being of order t as $t \to \infty$.

Hereafter, paralleling (23) and (62)–(67), we use the notation

$$\hat{\alpha}^r_{o,m,s,p}, \quad \mathcal{V}^r_{o,m,s,p}, \quad \mathcal{E}^r_{o,m,s,p}, \quad \mathcal{F}^r_{o,m,s,p}$$

to refer to the estimator, the variance ratio and the cpu-time (CT) and run-length (RL) efficiencies, respectively, for the combined estimator, using all three techniques. We suppress the subscripts o, m when overlapping is omitted; we suppress the subscripts s, p when splitting is omitted; and we suppress the subscripts s, p and the superscript r when considering the overlapping standard estimator.

8. A TEST MODEL: THE $M/G/1/0$ QUEUE

In order to see how these various simulation estimators perform, we conducted extensive simulation experiments with a relatively simple queueing model: the $M/G/1/0$ queue, which has a Poisson arrival process with rate λ, IID service times with a general cdf G having mean $1/\mu$, a single server and no extra waiting space. Because there is only one server and no extra waiting space, there can be either one customer in the system or none. We looked at the stochastic process $\{X(t) : t \geq 0\}$, where $X(t) = 1$ if a customer is present, and $X(t) = 0$ otherwise. We considered the alternative estimators for the expected value of the time average $Y(t) \equiv t^{-1} \int_0^t X(s)\, ds$, starting out empty at time 0, a regenerative state. That is a special case of (1) above with f being the identity function, i.e., $f(x) = x$ for all x. The random variable $Y(t)$ is the proportion of time that the server is busy during the interval $[0, t]$.

For the $M/G/1/0$ queue, the successive busy and idle times form an *alternating renewal process.* Thus the limit α is well known, namely,

$$\alpha = \frac{\mu^{-1}}{\lambda^{-1} + \mu^{-1}} = \frac{\lambda}{\lambda + \mu} \ .$$

For the general $M/G/1/0$ model, we have a two-state semi-Markov process, so that the various time-dependent performance measures can be computed using numerical transform inversion; e.g., see Duffield and Whitt [2000]. We consider this model, not because we need simulation efficiency for it, but because its tractability helps us understand what is happening.

8.1 Theory for the M/M/1/0 Case

Before actually considering the numerical experiments, we discuss closed-form analytical results that can be obtained in one case. Perhaps the most elementary nontrivial example is the queue-length process in the $M/M/1/0$ queueing model, because it is a two-state continuous-time Markov chain (CTMC). In addition to being an alternating renewal process, it is a birth-and-death process. As above, let the queue-length stochastic process be denoted by $X \equiv \{X(t) : t \geq 0\}$. Let the states be labelled 0 and 1. Let the birth (arrival) rate in state 0 be λ and let the death (service) rate in state 1 be μ. Let $P_{i,j}(t) \equiv P(X(t) = j|X(0) = i)$ denote the transition probability, i.e., the conditional probability of being in state j at time t given that the process started in state i at time 0.

For this example, the transition probabilities are easy to calculate by solving a system of two ordinary differential equations; see Example 6.11 on p. 364 of Ross [2003]. Here are the formulas:

$$P_{0,0}(t) \; = \; \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)t} \quad \text{and} \quad P_{1,0}(t) = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)t}, \quad t \geq 0 \; ,$$

with $P_{0,1}(t) = 1 - P_{0,0}(t)$ and $P_{1,1}(t) = 1 - P_{1,0}(t)$.

Recall that the initial state is 0 and the function f is the identity map. Then,

$$\alpha(t) = t^{-1} \int_0^t P_{0,1}(u)\, du = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{t(\lambda+\mu)^2}(1 - e^{-(\lambda+\mu)t}), \quad t \geq 0 \; . \tag{68}$$

We can directly calculate the variances $Var(Y(t))$ and $Var(R(t))$ for $Y(t)$ in (1) and $R(t)$ in (22). Thus we can calculate the variances of the standard and residual-cycle estimators. We say that $g(t)$ is $O(t)$ as $t \to \infty$ if $|g(t)/t| < M$ for some constant M for all t sufficiently large. In this example,

$$\begin{aligned} \mathsf{Var}(Y(t)) \; &= \; \frac{a_0}{t} + \frac{a_1}{t^2} + \frac{a_2 e^{-rt}}{t} + \frac{a_3 e^{-rt}}{t^2} + \frac{a_4 e^{-2rt}}{t^2} \; , \\ \mathsf{Var}(R(t)) \; &= \; \frac{b_0}{t^2} + \frac{b_1 e^{-rt}}{t^2} + \frac{b_2 e^{-2rt}}{t^2} \; , \\ \mathcal{V}^r(t) \equiv \frac{\mathsf{Var}(\hat{\alpha}_n(t))}{\mathsf{Var}(\hat{\alpha}_n^r(t))} = \frac{\mathsf{Var}(Y(t))}{\mathsf{Var}(R(t))} \; &= \; c_0 t + c_1 + c_2 t e^{-rt} + O(e^{-rt}) \; , \end{aligned} \tag{69}$$

as $t \to \infty$, where

$$\begin{aligned} a_0 \; &= \; \frac{2\lambda\mu}{(\lambda+\mu)^3}, \quad a_1 = \frac{\lambda(\lambda - 4\mu)}{\mu(\lambda+\mu)^4}, \quad a_2 = a_0(1 - \frac{\lambda}{\mu}), \\ a_3 \; &= \; \frac{4\lambda\mu}{(\lambda+\mu)^4}, \quad a_4 = \frac{\lambda^2}{(\lambda+\mu)^4}, \quad r = \lambda+\mu, \\ b_0 \; &= \; \frac{\lambda^2 + 4\lambda\mu + 2\mu^2}{(\lambda+\mu)^4}, \quad b_1 = \frac{2\lambda^2}{(\lambda+\mu)^4}, \quad b_2 = -b_1/2, \\ c_0 \; &= \; \frac{a_0}{b_0} = \frac{2\lambda\mu(\lambda+\mu)}{\lambda^2 + 4\lambda\mu + 2\mu^2}, \quad c_1 = \frac{a_1}{b_0} = \frac{\lambda^2 - 4\lambda\mu}{\lambda^2 + 4\lambda\mu + 2\mu^2}, \\ c_2 \; &= \; \frac{a_2 - a_0 b_1}{b_0}, \quad \frac{c_1}{c_0} = \frac{a_1}{a_0} = \frac{\lambda - 4\mu}{2\mu(\lambda+\mu)} \; . \end{aligned} \tag{70}$$

We can apply the variance formulas to calculate the relative efficiency of the two estimators, using the actual expected run length, instead of the estimated expected run length. We use the fact that

$$E[T(N(t)+1)] - t = \frac{1}{\mu} + \frac{P_{0,0}(t)}{\lambda} \to \frac{\lambda^2 + \lambda\mu + \mu^2}{\lambda\mu(\lambda+\mu)} \quad \text{as} \quad t \to \infty \; . \tag{71}$$

Then the run-length efficiency is

$$\mathcal{F}^r(t) = \frac{t\mathsf{Var}(Y(t))}{E[T(N(t)+1)]\mathsf{Var}(R(t))} = c_0 t + c_1 + O(te^{-rt}) \ , \tag{72}$$

where a_i, b_i and c_i are the parameters in (70).

This simple example shows us what we should look for more generally. From (68), we see that

$$\alpha(t) - \alpha = \frac{\beta}{t} - \frac{\beta}{t}e^{-\gamma t}, \quad t \geq 0 \ , \tag{73}$$

for $\beta = -\lambda/(\lambda+\mu)^2$ and $\gamma = \lambda + \mu$. Under suitable regularity conditions, we will have asymptotics like (73) for much more general models, but with equality replaced by asymptotic equivalence and for different parameters β and γ. In particular, we will have

$$\alpha(t) - \alpha \sim \frac{\beta}{t} - \frac{\delta}{t}e^{-\gamma t}, \quad t \geq 0 \ , \tag{74}$$

for appropriate parameters β, γ and δ, where $f(t) \sim g(t)$ means that $f(t)/g(t) \to 1$ as $t \to \infty$, and $f(t) \sim g(t) + h(t)$ means that $f(t) \sim g(t)$, $h(t)/g(t) \to 0$ and $f(t) - g(t) \sim h(t)$ as $t \to \infty$; i.e., we have an asymptotic expansion; see pp. 5-11 of Erdélyi [1956]. The parameter β in (73) and (74) is the asymptotic bias, as in (3) and (40).

8.2 Splitting in the $M/M/1/0$ Case

We now apply the splitting result in Section 6 to the $M/M/1/0$ model. For this model, the Markov and memoryless properties allow us to replace $\mathcal{F}_t$ by $X(t)$. Let U and V denote generic interarrival and service times, respectively, having means λ^{-1} and μ^{-1}, respectively. Then

$$\begin{aligned}
E[R(t)|X(t)=0] = \mu_0 &= E\left[\alpha + \frac{\alpha}{t}(U+V) - \frac{1}{t}Y\right] = \alpha + \frac{1}{t\mu} - \frac{1}{t\mu} = \alpha \ , \\
E[R(t)|X(t)=1] = \mu_1 &= E\left[\alpha + \frac{\alpha}{t}V - \frac{1}{t}V\right] = \alpha - \frac{1}{t(\lambda+\mu)} \ , \\
\mathsf{Var}[R(t)|X(t)=0] = \sigma_0^2 &= \mathsf{Var}\left(\alpha + \frac{\alpha}{t}(U+V) - \frac{1}{t}V\right) \\
&= \frac{1}{t^2}(\alpha^2\mathsf{Var}(U) + (\alpha-1)^2\mathsf{Var}(V)) = \frac{2}{t^2(\lambda+\mu)^2} \ , \\
\mathsf{Var}[R(t)|X(t)=1] = \sigma_1^2 &= \mathsf{Var}\left(\alpha + \frac{\alpha}{t}V - \frac{1}{t}V\right) = \frac{1}{t^2(\lambda+\mu)^2} \ .
\end{aligned}$$

Now letting $z(t) = P(X(t)=0)$ (given by $P_{0,0}(t)$ in (68)), we finally get

$$\begin{aligned}
\mathsf{Var}\left(\bar{R}^s\right) &= \mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right) + \frac{1}{p}E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right] \\
&= (\mu_0 - \mu_1)^2 z(t)(1-z(t)) + \frac{1}{p}(\sigma_0^2 z(t) + \sigma_1^2(1-z(t))) \\
&= \frac{z(t)(1-z(t))}{t^2(\lambda+\mu)^2} + \left(\frac{1}{p}\right)\left(\frac{1+z(t)}{t^2(\lambda+\mu)^2}\right) \ .
\end{aligned} \tag{75}$$

For this model, we also get the optimal splitting number $p^*(t)$ by

$$\begin{aligned} p^*(t) &= \sqrt{\frac{t}{E[T(N(t)+1)-t]} \times \frac{E\left[\mathsf{Var}\left(R(t) \mid \mathcal{F}_t\right)\right]}{\mathsf{Var}\left(E\left[R(t) \mid \mathcal{F}_t\right]\right)}} \\ &= \sqrt{\frac{t}{\frac{1}{\lambda}z(t)+\frac{1}{\mu}} \times \frac{1+z(t)}{z(t)(1-z(t))}}. \end{aligned} \tag{76}$$

Note that formula (76) simplifies greatly if we can regard $z(t)$ as approximately constant. Since $z(t)$ converges to a finite limit as $t \to \infty$, $z(t)$ should be approximately constant for large t. For example, for the special case in which $\lambda = \mu = 1$ and $z(t) \approx 1/2$, $p^*(t) = \sqrt{4t}$.

However, this example shows that there is the potential of significant variance reduction through splitting, because the ratio in (60) in this example is $(1 + z(t))/z(t)(1 - z(t))$, which becomes large as $z(t)$ approaches either 1 or 0.

9. $M/G/1/0$ NUMERICAL EXAMPLES

In this section we describe our experience with numerical examples for the $M/G/1/0$ model. We give a concise summary here; more details appear in Kang et al. [2005]. Without loss of generality, we fix the measuring units for time by letting the mean service time be 1. We considered three different service-time distributions: exponential (M), deterministic (D) and hyperexponential (H_2), which is a mixture of two exponential distributions, and thus more variable than a single exponential distribution. The specific H_2 density we considered was

$$g(x) = \frac{0.8}{17.0} \times 0.1e^{-0.1x} + \frac{16.2}{17.0} \times 1.8e^{-1.8x}, \quad x \geq 0 \ ,$$

which has SCV = 9.0 (i.e., squared coefficient of variation, defined as the variance divided by the square of the mean).

For the $M/M/1/0$ case, most of the results can be obtained directly from Section 8. In this case, the analytic results and simulation results are useful to provide a check on each other.

The analytic results are especially useful to show the form of the functions; e.g., as in equations (68), (69) and (72). On the other hand, the simulation results are the only available way to actually measure the computational effort through computer time expended, as characterized by the efficiency $\widehat{\mathcal{E}}^r_{o,s}$. We will see that run length is not always equivalent to cpu time.

9.1 The Experiment

We ran simulation experiments for three values of λ: 0.5, 1.0 and 1.5. Since we found no significant differences in the results for these three cases, we concentrated on the case $\lambda = 1$, for which $\alpha = 0.5$. For this case, the mean regenerative cycle length is $E[\tau_1] = 2.0$. We are primarily thinking of our procedures applying to time t greater than the mean cycle length, but we consider the full range of t in our experiments.

We performed simulation experiments for *six* different estimators: (1) standard, (2) standard plus overlapping cycles, (3) residual-cycle, (4) residual-cycle plus overlapping cycles, (5) residual-cycle plus splitting, and (6) residual-cycle plus both overlapping cycles and splitting. For each model and estimator, we considered 14 different time points t, changing in powers of 2. For $M/D/1/0$ and $M/M/1/0$, the times were 2^k for $-3 \leq k \leq 10$. Since $\alpha(t)$ approached α much more slowly with the highly variable H_2 service-time distribution, we used times 2^k for $0 \leq k \leq 13$ for the $M/H_2/1/0$ model. For both overlapping (m) and splitting (p), we considered 5 levels: 1 (not used), 10, 30, 50 and 100. When we combined all three techniques, we used the same number of splittings as overlapping cycles. The 6 different estimators and 5 levels of overlapping and/or splitting leads to 18 cases, not counting the time t. Thus, that produces $18 \times 14 = 252$ separate simulation experiments overall. In each separate simulation experiment we used $n = 1000$ independent replications. Thus we performed $252 \times 1000 = 252,000$ simulation runs.

9.2 The Results

We present some of those results in Tables I–IV below; the rest appear in Kang et al. [2005]. Since the $M/D/1/0$ results were similar to the $M/M/1/0$ results, we only display the $M/M/1/0$ results here. To save space, we reduced the number of time points from 14 to 10 in each case, putting five in each table below. We also reduced the number of cases here by not displaying the level $m = 50$ for the overlapping cycles and the level $p = 50$ for the splitting. That reduced the number from $14 \times 18 = 252$ to $10 \times 13 = 130$.

Tables I–IV clearly show that the overlapping cycles does not help the standard estimator. The efficiencies $\widehat{\mathcal{E}}_{o,m}$ and $\widehat{\mathcal{F}}_{o,m}$ are consistently less than or equal to 1 or only slightly above 1.

The situation looks much better when we consider the residual-cycle estimators, but even then the performance is not uniformly good. The time t evidently must exceed a *threshold* for there to be any efficiency benefit. For the $M/G/1/0$ queue with $\alpha = 0.5$, that threshold seems to occur when $0.40 \leq \alpha(t) \leq 0.45$, when $\alpha(t)$ is within between 20% and 10% of the limit $\alpha = 0.50$. When $\alpha(t) \geq 0.45$, we see consistent efficiency gains from the residual-cycle estimator and its overlapping-cycle and splitting refinements. When $\alpha(t) < 0.40$, the standard estimator is consistently more efficient. The observed threshold is only slightly greater than the mean cycle length $E[\tau_1] = 2.0$ for M service, but considerably greater for H_2 service, and somewhat less for D service. Our limited experimentation suggests that, as a general principle, the *threshold for simulation-efficiency gain* from the residual-cycle estimator might be approximately the limiting mean residual cycle, which is the mean of the equilibrium residual-lifetime of the cycle-time distribution, i.e., we might expect efficiency gain from using these techniques for times t with

$$t \geq E[T_+(\infty)] = E[\tau_1^2]/2E[\tau_1] = E[\tau_1](SCV(\tau_1) + 1)/2 \; . \tag{77}$$

This principle is borne out in this one $M/G/1/0$ example, where the threshold is 1 for D, 2 for M and 10 for H_2. More generally, this conjecture remains to be investigated.

We see dramatic improvement as t increases. From the perspective of the variance ratio, especially $\widehat{\mathcal{V}}^r_{o,s}$ for the combined estimator, the variance reduction looks very impressive for larger values of t, e.g., $\widehat{\mathcal{V}}^r_{o,100,s,100} = 326,000$ for the $M/M/1/0$ model with $t = 1024$. However, the story is less impressive when we look at the efficiencies,

Table I. The Efficiencies for $M/M/1/0$: $n = 1000$, $\lambda = 1.0$, $\alpha = 0.5$

t	0.1250	1.0	2.0	4.0	8.0
$\hat{\alpha}(t)$	0.0489	0.2997	0.3674	0.4294	0.4742
$HW_{n,1}$	0.0106	0.0197	0.0171	0.0138	0.0103
$\widehat{\mathcal{V}}_{o10}$	7.7	8.6	6.5	4.6	2.8
$\widehat{\mathcal{V}}_{o30}$	23.9	27.8	18.0	11.8	6.4
$\widehat{\mathcal{V}}_{o100}$	84.4	82.3	64.5	39.1	23.1
$\widehat{\mathcal{V}}^r$	9.8e-004	0.24	0.66	1.9	4.1
$\widehat{\mathcal{V}}^r_{o10}$	8.9e-003	1.5	3.9	9.6	22
$\widehat{\mathcal{V}}^r_{o30}$	0.03	4.6	11.2	25.2	63
$\widehat{\mathcal{V}}^r_{o100}$	0.10	14.4	37.6	80.3	193
$\widehat{\mathcal{V}}^r_{s10}$	6.5e-003	1.0	3.1	7.6	18
$\widehat{\mathcal{V}}^r_{s30}$	0.013	1.3	4.2	10.4	24
$\widehat{\mathcal{V}}^r_{s100}$	0.016	1.6	4.5	12.4	27
$\widehat{\mathcal{V}}^r_{o10,s10}$	0.063	8.3	24	65	140
$\widehat{\mathcal{V}}^r_{o30,s30}$	0.33	34	100	264	591
$\widehat{\mathcal{V}}^r_{o100,s100}$	1.5	134	380	929	2,302
$\widehat{\mathcal{E}}_{o10}$	0.9	1.1	1.0	1.0	0.8
$\widehat{\mathcal{E}}_{o30}$	1.0	1.2	1.0	1.0	0.7
$\widehat{\mathcal{E}}_{o100}$	1.0	1.1	1.1	1.0	0.9
$\widehat{\mathcal{E}}^r$	1.0e-003	0.2	0.7	1.9	4.1
$\widehat{\mathcal{E}}^r_{o10}$	1.0e-003	0.18	0.58	2.1	6.1
$\widehat{\mathcal{E}}^r_{o30}$	1.2e-003	0.20	0.60	2.0	7.0
$\widehat{\mathcal{E}}^r_{o100}$	1.1e-003	0.19	0.62	1.9	6.9
$\widehat{\mathcal{E}}^r_{s10}$	9.0e-004	0.18	0.73	2.3	8.2
$\widehat{\mathcal{E}}^r_{s30}$	6.2e-004	0.09	0.38	1.4	5.0
$\widehat{\mathcal{E}}^r_{s100}$	2.3e-004	0.03	0.13	0.5	1.9
$\widehat{\mathcal{E}}^r_{o10,s10}$	8.7e-004	0.16	0.63	2.6	8.8
$\widehat{\mathcal{E}}^r_{o30,s30}$	5.4e-004	0.079	0.32	1.3	4.8
$\widehat{\mathcal{E}}^r_{o100,s100}$	2.3e-004	0.029	0.11	0.44	1.8
$\widehat{\mathcal{F}}_{o10}$	0.05	0.45	0.66	0.83	0.87
$\widehat{\mathcal{F}}_{o30}$	0.05	0.47	0.60	0.77	0.77
$\widehat{\mathcal{F}}_{o100}$	0.05	0.41	0.65	0.77	0.90
$\widehat{\mathcal{F}}^r$	6.0e-005	0.096	0.37	1.4	3.4
$\widehat{\mathcal{F}}^r_{o10}$	5.5e-005	0.070	0.36	1.6	6.3
$\widehat{\mathcal{F}}^r_{o30}$	6.4e-005	0.076	0.36	1.6	7.5
$\widehat{\mathcal{F}}^r_{o100}$	6.0e-005	0.072	0.37	1.6	7.4
$\widehat{\mathcal{F}}^r_{s10}$	4.3e-005	0.060	0.37	1.6	6.4
$\widehat{\mathcal{F}}^r_{s30}$	2.8e-005	0.028	0.18	0.85	3.6
$\widehat{\mathcal{F}}^r_{s100}$	1.0e-005	0.010	0.059	0.32	1.3
$\widehat{\mathcal{F}}^r_{o10,s10}$	4.2e-005	0.050	0.31	1.6	6.9
$\widehat{\mathcal{F}}^r_{o30,s30}$	2.4e-005	0.024	0.14	0.77	3.5
$\widehat{\mathcal{F}}^r_{o100,s100}$	1.0e-005	0.009	0.05	0.25	1.2

because the variance reduction is obtained at the expense of some computational effort. Nevertheless, $\widehat{\mathcal{F}}^r_{o,30,s,30} = 35,000$ for the $M/M/1/0$ model with $t = 1024$. However, for the largest values of t, we might already be confident that $\alpha(t)$ will be close to the limit α. It may be best to judge the performance by looking at cases

Table II. The Efficiencies for $M/M/1/0$: $n = 1000$, $\lambda = 1.0$, $\alpha = 0.5$

t	16.0	64.0	128.0	256.0	1024.0
$\hat{\alpha}^r_{o100,s100}(t)$	0.48437	0.49608	0.49805	0.49903	0.49976
$HW_{n,1}$	7.50e-003	3.94e-003	2.72e-003	1.97e-003	9.80e-004
$\widehat{\mathcal{V}}_{o10}$	1.6	1.1	1.1	1.2	1.1
$\widehat{\mathcal{V}}_{o30}$	3.7	1.5	1.2	1.2	1.0
$\widehat{\mathcal{V}}_{o100}$	11.7	3.5	2.1	1.4	1.0
$\widehat{\mathcal{V}}^r$	9.3	40	73	130	567
$\widehat{\mathcal{V}}^r_{o10}$	42	188	329	668	3,086
$\widehat{\mathcal{V}}^r_{o30}$	125	521	998	2,167	8,683
$\widehat{\mathcal{V}}^r_{o100}$	349	1,814	3,248	6,583	29,095
$\widehat{\mathcal{V}}^r_{s10}$	38	174	317	666	2,587
$\widehat{\mathcal{V}}^r_{s30}$	51	222	411	871	3,505
$\widehat{\mathcal{V}}^r_{s100}$	57	253	476	1,042.1	3,935
$\widehat{\mathcal{V}}^r_{o10,s10}$	288	1,363	2,399	4,823	19,671
$\widehat{\mathcal{V}}^r_{o30,s30}$	1,174	5,269	9,910	19,407	83,507
$\widehat{\mathcal{V}}^r_{o100,s100}$	4,440.1	18,867	38,162	77,404	326,297
$\widehat{\mathcal{E}}_{o10}$	0.58	0.58	0.62	0.74	0.70
$\widehat{\mathcal{E}}_{o30}$	0.60	0.41	0.39	0.42	0.40
$\widehat{\mathcal{E}}_{o100}$	0.63	0.38	0.28	0.20	0.17
$\widehat{\mathcal{E}}^r$	9.3	41	74	131	574
$\widehat{\mathcal{E}}^r_{o10}$	14.8	100	191	406	1,935
$\widehat{\mathcal{E}}^r_{o30}$	19.5	146	321	778	3,363
$\widehat{\mathcal{E}}^r_{o100}$	17.9	191	411	953	4,629
$\widehat{\mathcal{E}}^r_{s10}$	18.2	145	280	637	2,464
$\widehat{\mathcal{E}}^r_{s30}$	15.3	138	279	753	3,301
$\widehat{\mathcal{E}}^r_{s100}$	6.6	83	226	676	3,484
$\widehat{\mathcal{E}}^r_{o10,s10}$	28.8	354	863	2,268	11,666
$\widehat{\mathcal{E}}^r_{o30,s30}$	16.0	240	779	2,595	22,412
$\widehat{\mathcal{E}}^r_{o100,s100}$	5.7	88	335	1,291	16,906
$\widehat{\mathcal{F}}_{o10}$	0.76	0.85	0.96	1.10	1.10
$\widehat{\mathcal{F}}_{o30}$	0.81	0.80	0.81	0.95	0.97
$\widehat{\mathcal{F}}_{o100}$	0.87	0.86	0.84	0.77	0.88
$\widehat{\mathcal{F}}^r$	8.5	39	72	129	567
$\widehat{\mathcal{F}}^r_{o10}$	19	145	285	620	3,028
$\widehat{\mathcal{F}}^r_{o30}$	26	269	682	1,759	8,205
$\widehat{\mathcal{F}}^r_{o100}$	26	441	1,267	3,702	24,345
$\widehat{\mathcal{F}}^r_{s10}$	19	141	283	630	2,550
$\widehat{\mathcal{F}}^r_{s30}$	13	130	304	740	3,355
$\widehat{\mathcal{F}}^r_{s100}$	5.4	75	220	658	3,428
$\widehat{\mathcal{F}}^r_{o10,s10}$	27	399	1,087	3,006	17,086
$\widehat{\mathcal{F}}^r_{o30,s30}$	14	236	849	3,057	35,831
$\widehat{\mathcal{F}}^r_{o100,s100}$	4.7	80	322	1,294.2	20,803

Table III. The Efficiencies for $M/H_2/1/0$: $n = 1000$, $\lambda = 1.0$, $\alpha = 0.5$

t	1.0	4.0	8.0	16.0	64.0
$\hat{\alpha}^r_{o100,s100}(t)$	0.083	0.186	0.263	0.361	0.4626
$HW_{n,1}$	0.0110	0.0162	0.0174	0.0174	0.0112
$\widehat{\mathcal{V}}_{o10}$	7.6	3.5	2.1	1.6	1.1
$\widehat{\mathcal{V}}_{o30}$	21.3	9.0	5.2	3.2	1.5
$\widehat{\mathcal{V}}_{o100}$	68.2	29.6	16.2	9.9	3.1
$\widehat{\mathcal{V}}^r$	3.3e-003	0.089	0.30	1.1	6.5
$\widehat{\mathcal{V}}^r_{o10}$	0.018	0.19	0.64	1.3	10.4
$\widehat{\mathcal{V}}^r_{o30}$	0.061	0.48	1.3	2.8	18.9
$\widehat{\mathcal{V}}^r_{o100}$	0.18	1.6	3.9	8.7	42.4
$\widehat{\mathcal{V}}^r_{s10}$	0.013	0.19	0.75	2.6	17.9
$\widehat{\mathcal{V}}^r_{s30}$	0.018	0.22	0.81	3.2	20.1
$\widehat{\mathcal{V}}^r_{s100}$	0.019	0.24	0.85	3.3	21.8
$\widehat{\mathcal{V}}^r_{o10,s10}$	0.092	0.65	1.8	5.5	34.6
$\widehat{\mathcal{V}}^r_{o30,s30}$	0.33	1.9	5.2	14.7	88.9
$\widehat{\mathcal{V}}^r_{o100,s100}$	1.2	6.7	14.8	46	282
$\widehat{\mathcal{E}}_{o10}$	1.1	0.88	0.75	0.67	0.62
$\widehat{\mathcal{E}}_{o30}$	1.2	1.00	0.76	0.60	0.47
$\widehat{\mathcal{E}}_{o100}$	1.2	0.99	0.79	0.67	0.39
$\widehat{\mathcal{E}}^r$	3.0e-003	0.089	0.52	1.0	6.6
$\widehat{\mathcal{E}}^r_{o10}$	2.7e-003	0.050	0.22	0.52	6.0
$\widehat{\mathcal{E}}^r_{o30}$	3.4e-003	0.053	0.21	0.49	5.9
$\widehat{\mathcal{E}}^r_{o100}$	2.8e-003	0.051	0.20	0.56	5.1
$\widehat{\mathcal{E}}^r_{s10}$	1.5e-003	0.032	0.19	0.71	9.7
$\widehat{\mathcal{E}}^r_{s30}$	7.5e-004	0.0127	0.062	0.34	5.6
$\widehat{\mathcal{E}}^r_{s100}$	2.3e-004	4.4e-003	0.021	0.11	2.2
$\widehat{\mathcal{E}}^r_{o10,s10}$	1.1e-003	0.012	0.049	0.19	3.5
$\widehat{\mathcal{E}}^r_{o30,s30}$	4.5e-004	4.1e-003	0.016	0.25	1.1
$\widehat{\mathcal{E}}^r_{o100,s100}$	1.5e-004	1.3e-003	3.9e-003	0.016	0.31
$\widehat{\mathcal{F}}_{o10}$	0.40	0.62	0.67	0.77	0.83
$\widehat{\mathcal{F}}_{o30}$	0.36	0.58	0.64	0.67	0.80
$\widehat{\mathcal{F}}_{o100}$	0.34	0.58	0.63	0.73	0.76
$\widehat{\mathcal{F}}^r$	9.1e-004	0.046	0.19	0.79	6.0
$\widehat{\mathcal{F}}^r_{o10}$	8.4e-004	0.03	0.17	0.51	7.6
$\widehat{\mathcal{F}}^r_{o30}$	1.0e-003	0.029	0.15	0.55	9.4
$\widehat{\mathcal{F}}^r_{o100}$	8.7e-004	0.030	0.147	0.64	10.1
$\widehat{\mathcal{F}}^r_{s10}$	4.8e-004	0.018	0.11	0.57	9.5
$\widehat{\mathcal{F}}^r_{s30}$	2.4e-004	6.9e-003	0.041	0.29	5.6
$\widehat{\mathcal{F}}^r_{s100}$	7.4e-005	2.4e-003	0.013	0.094	2.3
$\widehat{\mathcal{F}}^r_{o10,s10}$	3.6e-004	6.8e-003	0.031	0.16	3.6
$\widehat{\mathcal{F}}^r_{o30,s30}$	1.4e-004	2.2e-003	9.4e-003	0.048	1.1
$\widehat{\mathcal{F}}^r_{o100,s100}$	4.8e-005	6.8e-004	2.4e-003	0.0133	0.31

Table IV. The Efficiencies for $M/H_2/1/0$: $n = 1000$, $\lambda = 1.0$, $\alpha = 0.5$

t	128.0	256.0	512.0	1024.0	8192.0
$\hat{\alpha}^r_{o100,s100}(t)$	0.4816	0.4907	0.4953	0.49767	0.49971
$HW_{n,1}$	8.30e-003	5.93e-003	4.09e-003	3.04e-003	1.06e-003
$\widehat{\mathcal{V}}_{o10}$	1.0	1.1	1.0	1.1	1.0
$\widehat{\mathcal{V}}_{o30}$	1.2	1.1	1.0	1.0	0.9
$\widehat{\mathcal{V}}_{o100}$	2.0	1.4	1.0	1.1	1.0
$\widehat{\mathcal{V}}^r$	14	26	53	132	1064
$\widehat{\mathcal{V}}^r_{o10}$	27	47	80	191	1,383
$\widehat{\mathcal{V}}^r_{o30}$	44	89	149	275	2,424
$\widehat{\mathcal{V}}^r_{o100}$	96	223	377	1,030	7,584
$\widehat{\mathcal{V}}^r_{s10}$	39	82	147	325	2,588
$\widehat{\mathcal{V}}^r_{s30}$	45	92	175	394	3,013
$\widehat{\mathcal{V}}^r_{s100}$	48	94	186	401	3,109
$\widehat{\mathcal{V}}^r_{o10,s10}$	81	169	330	720	5,355
$\widehat{\mathcal{V}}^r_{o30,s30}$	196	395	765	1,712	12,958
$\widehat{\mathcal{V}}^r_{o100,s100}$	620	1,222	2,361	5,172	39,969
$\widehat{\mathcal{E}}_{o10}$	0.65	0.70	0.67	0.74	0.72
$\widehat{\mathcal{E}}_{o30}$	0.43	0.43	0.43	0.44	0.42
$\widehat{\mathcal{E}}_{o100}$	0.29	0.23	0.19	0.20	0.19
$\widehat{\mathcal{E}}^r$	15	25	52	131	1,093
$\widehat{\mathcal{E}}^r_{o10}$	17	29	53	130	966
$\widehat{\mathcal{E}}^r_{o30}$	16	35	61	117	1,087
$\widehat{\mathcal{E}}^r_{o100}$	14	37	66	190	1,493
$\widehat{\mathcal{E}}^r_{s10}$	27	66	121	305	2,624
$\widehat{\mathcal{E}}^r_{s30}$	18	54	124	334	3,011
$\widehat{\mathcal{E}}^r_{s100}$	8.3	27	83	246	2,948
$\widehat{\mathcal{E}}^r_{o10,s10}$	14	45	125	352	3,606
$\widehat{\mathcal{E}}^r_{o30,s30}$	4.3	16	57	217	4,499
$\widehat{\mathcal{E}}^r_{o100,s100}$	1.3	4.8	18	76	3,158
$\widehat{\mathcal{F}}_{o10}$	0.89	1.00	0.96	1.11	1.00
$\widehat{\mathcal{F}}_{o30}$	0.83	0.88	0.93	0.98	0.93
$\widehat{\mathcal{F}}_{o100}$	0.77	0.77	0.75	0.91	0.99
$\widehat{\mathcal{F}}^r$	14	25	52	132	1,064
$\widehat{\mathcal{F}}^r_{o10}$	23	43	77	187	1,379
$\widehat{\mathcal{F}}^r_{o30}$	29	71	133	259	2,406
$\widehat{\mathcal{F}}^r_{o100}$	37	124	269	861	7,399
$\widehat{\mathcal{F}}^r_{s10}$	27	68	132	307	2,570
$\widehat{\mathcal{F}}^r_{s30}$	19	55	132	338	2,952
$\widehat{\mathcal{F}}^r_{s100}$	9.0	29	89	257	2,904
$\widehat{\mathcal{F}}^r_{o10,s10}$	16	55	161	472	5,017
$\widehat{\mathcal{F}}^r_{o30,s30}$	4.9	19	70	289	7,997
$\widehat{\mathcal{F}}^r_{o100,s100}$	1.4	5.4	21	91	4,985

in which $\alpha(t) \approx 0.495$, which is within 1% of the limit. Then the efficiency gains can be considered from the cases $t = 64$ for $M/M/1/0$ and $t = 512$ for $M/H_2/1/0$; the maximum efficiency gains, considering only $\widehat{\mathcal{E}}$ is 354 for M service and 125 for H_2 service.

Consistent with our theoretical analysis of splitting in Sections 6 and 8, we see that the benefit of splitting is not monotone in the number p of splittings for each t, and the optimal level of splitting, $p^*(t)$ evidently is increasing in t; $p = 10$ seems best for small t, while $p = 30$ seems best for medium t, and $p = 100$ seems best for large t. (Analysis indicated $p^* = \sqrt{Kt}$.) In contrast, the residual-cycle plus overlapping-cycle efficiencies $\widehat{\mathcal{E}}^r_{o,m}$ and $\widehat{\mathcal{F}}^r_{o,m}$ seem to increase in m when t is not too small.

9.3 Uniformity of Effectiveness

The tables show that the efficiency gains for the H_2 service time occur for much larger values of t than for the other two service-time distributions. However, the time-dependent mean $\alpha(t)$ approaches the common limit $\alpha = 0.5$ much more slowly for the H_2 service-time distribution. We found a fascinating uniformity of effectiveness over time if instead of looking at the efficiency as a function of time t, we look at the efficiency as a function of $\alpha(t)$ or as a function of $\alpha(t)/\alpha$.

We also found it informative to plot the logarithm of the efficiency, with base 10. The logarithm with base 10 shows the order of magnitude effect; a value k corresponds to 10^k or k orders of magnitude improvement. To see the uniformity, we plotted the logarithm (base 10) of all the efficiency functions $\widehat{\mathcal{E}}(t)$ versus $\alpha(t)$ for the time points t considered.

By uniformity of effectiveness we mean that the three curves for the three service times tend to fall on top of each other. They all show steady improvement as t increases (with the exception of the overlapping-cycle modification of the standard estimator). We found that it is even more revealing to use a log scale on both axes; i.e., to plot the logarithm of the efficiency against $-\log_{10}(1 - (\hat{\alpha}(t)/\alpha))$ (the logarithm of the relative bias, regarding the estimator as an estimate of the limit α), with all logarithms to base 10. We then see a striking *linear relationship* when t is sufficiently large or, more precisely, when the estimator $\hat{\alpha}(t)$ is sufficiently close to $\alpha = 0.5$.

We illustrate in Figures 1 and 2. In Figure 1 we plot the logarithm of the efficiency $\widehat{\mathcal{E}}^r(t)$ for the pure residual-cycle estimator $\hat{\alpha}^r(t)$ versus $-\log(1 - (\hat{\alpha}^r(t)/\alpha))$. In Figure 2 we plot the logarithm of the efficiency $\widehat{\mathcal{E}}^r_{o,30,s,30}(t)$ for the combined estimator $\hat{\alpha}^r_{o,30,s,30}(t)$ versus $-\log(1 - (\hat{\alpha}^r_{o,30,s,30}(t)/\alpha))$. The plots for the other estimators are similar; see Kang et al. [2005].

The values 1 and 2 on the horizontal x axis of Figure 2 mean that $\alpha(t)$ differs from the limit α by 10% and 1%, respectively, while the values 1 and 2 on the vertical y axis mean that the estimated efficiency $\widehat{\mathcal{E}}^r_{o,30,s,30}(t)$ is $10^1 = 10$ and $10^2 = 100$, respectively. By definition, the threshold for efficiency gain occurs at 0 on the y axis, which evidently occurs when $\hat{\alpha}^r_{o,30,s,30}(t)$ is within about 10% of the limit α.

The linear relationship might be said to begin when $-\log_{10}(1 - (\hat{\alpha}^r_{o,30,s,30}(t)/\alpha)) = 0.5$, but certainly has begun when $-\log_{10}(1 - (\hat{\alpha}^r_{o,30,s,30}(t)/\alpha)) = 1.0$, when $\hat{\alpha}^r_{o,30,s,30}(t)$ is within 10% of α.

9.4 Implemented Efficiency Versus Run-Length Efficiency

The experiments nicely illustrate differences between different notions of "efficiency." First, we expect to see that the run-length and cpu-time efficiencies will be less than the variance ratios, because the variance reduction is usually obtained at some run-length and cpu-time cost, but sometimes that cost is small. To illustrate, first consider the largest time, $t = 1024$ for the $M/M/1/0$ model in Table II. Since a mean cycle time $E[\tau_1] = 2.0$ is negligible compared to $t = 1024$, it should not be surprising that $\widehat{\mathcal{F}}^r \approx \widehat{\mathcal{E}}^r \approx \widehat{\mathcal{V}}^r \approx 570$. In this case, the variance reduction is fully realized without run-length or cpu-time cost.

But in other cases, we expect to see a degradation of efficiency gain as we go from the variance ratio to the run-length and cpu-time efficiency measures. For example, for that same time $t = 1024$ with exponential (M) service in Table II, when we turn to the overlapping-cycle residual-cycle estimator $\hat{\alpha}_o^r$, we see that there is a bit more run-length cost, because now the m overlapping cycles make up a larger portion of the time $t = 1024$. Accordingly, we are not surprised to see that $\widehat{\mathcal{F}}^r_{o,100} \approx 24,300 < 29,100 \approx \widehat{\mathcal{V}}^r_{o,100}$.

The experiments also nicely illustrate differences between the "theoretical" run-length efficiency and the "implemented" cpu-time efficiency. Even though $\widehat{\mathcal{F}}^r_{o,100} \approx$

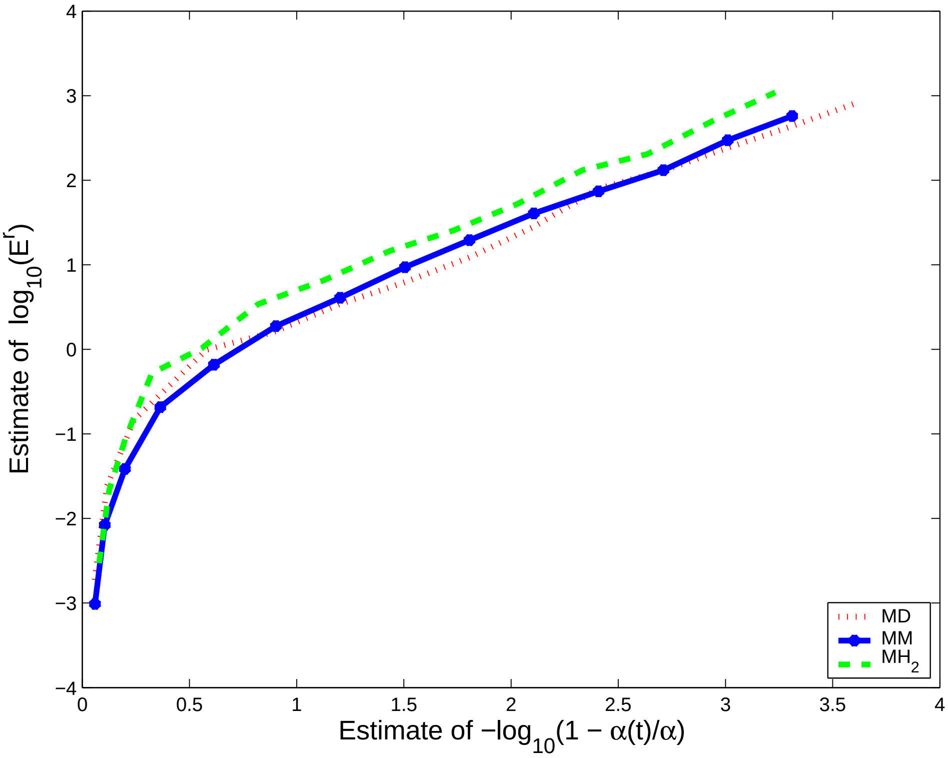

Fig. 1. The logarithm (base 10) of the cpu-time efficiency of the residual-cycle estimator, $\widehat{\mathcal{E}}^r(t)$, as a function of minus the logarithm (base 10) of the estimated $1 - (\hat{\alpha}^r(t)/\alpha)$.

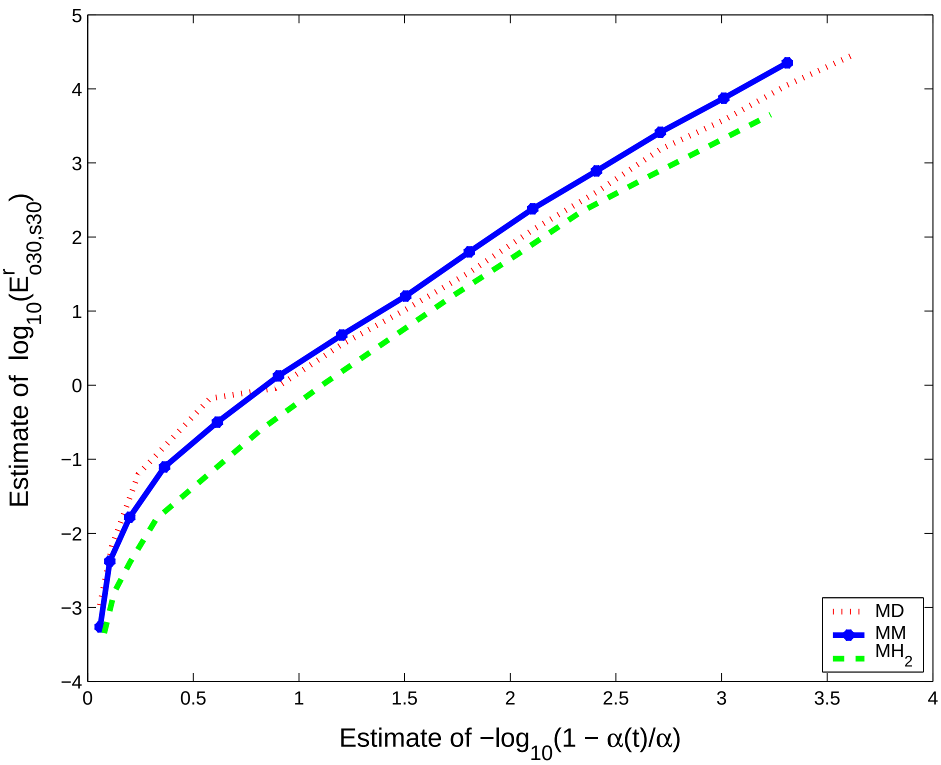

Fig. 2. The logarithm (base 10) of the cpu-time efficiency of the combined estimator, $\widehat{\mathcal{E}}^r_{o,30,s,30}(t)$, as a function of minus the logarithm (base 10) of the estimated $1 - (\hat{\alpha}^r_{o,30,s,30}(t)/\alpha)$.

$24,300$, the cpu-time efficiency is only $\widehat{\mathcal{E}}^r_{o,100} \approx 4,600$, a substantial difference. That occurs because it requires extra computation to implement the refined overlapping-cycle estimator. The much higher run-length efficiency indicates what might possibly be gained from a more efficient implementation, but the cpu-time efficiency is what was actually realized. The main point is to recognize that *it is important to substantiate efficiency gains through actual implementation.*

There are other interesting subtle differences as well. Consider the overlapping-cycle modification of the standard estimator, $\hat{\alpha}_{o,r}$ for M service at the smallest time point, $t = 0.125$. The variance ratios $\widehat{\mathcal{V}}_{o,m}$ show variance gains just slightly less than would occur with m multiple independent replications. However, the run-length efficiencies are very very low: $\widehat{\mathcal{F}}_{o,m} \approx 0.05$. That can be understood by recognizing that we waste a lot of time generating full cycles. The time t is only $1/16$ of the mean cycle time $E[\tau_1] = 2.0$, so we should expect the run-length efficiency loss we see.

What is startling to see, however, is that the cpu-time efficiencies show no such degradation: $\widehat{\mathcal{E}}_{o,m} \approx 1.0$. At first that suggests an error, but upon reflection it is not hard to understand. From the cpu-time perspective, there is relatively little waste during a cycle, because a full cycle is generated by generating only two random

Table V. Key Parameters in the $M/G/5/0$ Queue

λ	$E[\tau_1]$	α
1	2.716667	0.996933
2	3.633333	1.926606
3	6.133333	2.669837
4	10.716667	3.203733
5	18.283333	3.575661
6	29.966667	3.837597
7	47.109524	4.026964
8	71.258333	4.167934
9	104.161111	4.275828
10	147.766667	4.360478

variables: the initial interarrival time and then the following service time. Every cycle, no matter how short, will require generating at least one of these variables.

10. EXPERIMENTS FOR THE $M/G/5/0$ QUEUE

It is evident that the results for the $M/G/1/0$ model are roughly indicative of what will happen for other single-server $GI/G/1/r$ models, provided that either the traffic intensity $\rho = \lambda/\mu$ is not large (e.g., $\rho < 0.7$) or the traffic intensity is moderate (e.g., $\rho < 1.5$) and r is relatively small. Then the busy cycles will not be long. But the busy cycles will be much more variable, so we can expect $\alpha(t)$ to approach α more slowly.

As we indicated in Section 2, the situation can be much more complicated with multiserver $GI/G/s/r$ queues, when the regeneration epochs are the times arrivals come to an empty system, because the busy cycles can become very long. For large s, the empty state is visited so rarely that the residual-cycle estimator and its refinements are not promising, and we did not consider them.

To consider a second manageable case, we considered the $M/G/s/0$ loss model with $s = 5$ and $\mu = 1$. As before, we let $X(t)$ be the number of customers in the system at time t, and we focus on the time average, with the function f in (1) being the identity map. As before, arrivals to an empty system are the designated regeneration epochs.

The intervals τ_i between successive regeneration points are busy cycles. It is instructive to start to see how α and the mean busy cycle $E[\tau_1]$ depend on the arrival rate λ. Table V contains these results for λ varying from 1 to 10. The mean busy period $E[\tau_1]$ depends only on λ, μ and the steady-state (truncated Poisson) distribution π, which has the insensitivity property, so both $E[\tau_1]$ and α depend on the service-time distribution only through its mean. In particular, with I an idle period,

$$\pi_0 = \frac{E[I]}{E[\tau_1]} = \frac{\lambda^{-1}}{E[\tau_1]} \ . \tag{78}$$

10.1 The Experiment

We performed simulations for the three service-time distributions - D, M and H_2 - with $\mu = 1$ and for three values of λ - 1, 3 and 5. We choose the time horizons

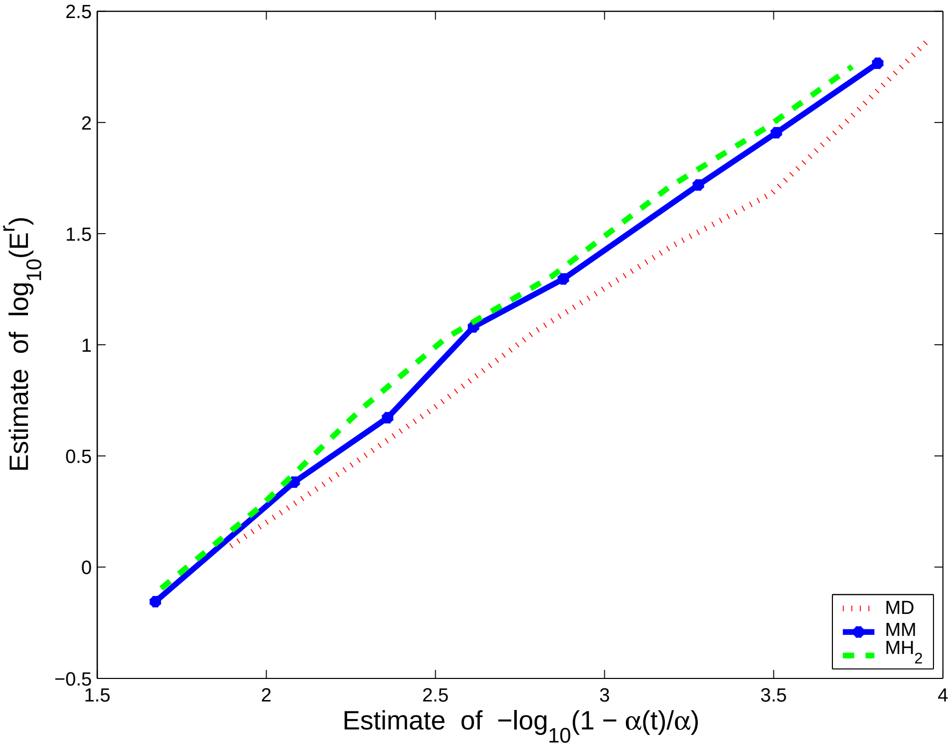

Fig. 3. The logarithm (base 10) of the cpu-time efficiency of the residual-cycle estimator, $\widehat{\mathcal{E}}^r(t)$, as a function of minus the logarithm (base 10) of the estimated $1 - (\hat{\alpha}^r(t)/\alpha)$ for the $M/G/5/0$ model with $\lambda = 5$, based on $n = 1000$ replications.

to roughly satisfy $1 \leq -\log(1 - \alpha(t)/\alpha) \leq 4$. For these experiments, we again considered $n = 1000$ replications.

10.2 The Results

Paralleling Figure 2, for each possible value of λ, we plotted the logarithm (base (10) of the observed cpu-time efficiency, $-\log_{10}\widehat{\mathcal{E}}^r$, versus $-\log_{10}(1 - (\hat{\alpha}^r(t)/\alpha))$ for the three service-time distributions. We present the plot for $\lambda = 5$ in Figure 3 below. Tables and plots for all other cases appear in Kang et al. (2005).

Once again we see, first, that there are substantial efficiency gains for large t, second, that the three curves for D, M and H_2 service fall on top of each other and, third, that the relationship between these two quantities is again approximately linear.

11. ESTIMATING THE ASYMPTOTIC BIAS

Given the asymptotic-bias relation (3) and (30), we can perform statistical analysis with our simulation results to see if the relation is approximately true for finite times of interest, and estimate the asymptotic-bias value β. We use our estimates of $\alpha(t)$ for larger times t.

Taking logarithms, we have

$$-\log_{10}(\alpha - \alpha(t)) \approx -\log_{10}(-\beta) + \log_{10} t \tag{79}$$

for $\beta > 0$. Thus we plot $(\log_{10}(t_i), -\log_{10}(\alpha - \alpha(t_i)))$ for all t_i, $i = 1, \ldots, n$ for the $M/G/1/0$ model in Figure 4, using 11 time points after discarding the initial 3 time points. (Recall that we used 14 time points of the form t^k; see Section 9.) The points evidently lie on a line of slope 1, as they should. The estimated slopes for the $M/G/1/0$ model, estimated β values and their R^2-values by regression analysis, are given in Table VI. Even though equation (79) holds only asymptotically, we see that the linearity is observed for the times we considered.

We obtain a better estimate of the asymptotic bias β by performing a linear regression, based on the relation

$$\frac{1}{\alpha(t) - \alpha} \sim \frac{t}{\beta} \; . \tag{80}$$

Those estimates are also given in Figure 4 and Table VI, under the linear columns. The two $M/M/1/0$ estimates of β in Table VI can be compared to the exact value $\beta = -\lambda/(\lambda + \mu)^2 = -0.2500$ from (73). The linear regression based on (80) is very accurate.

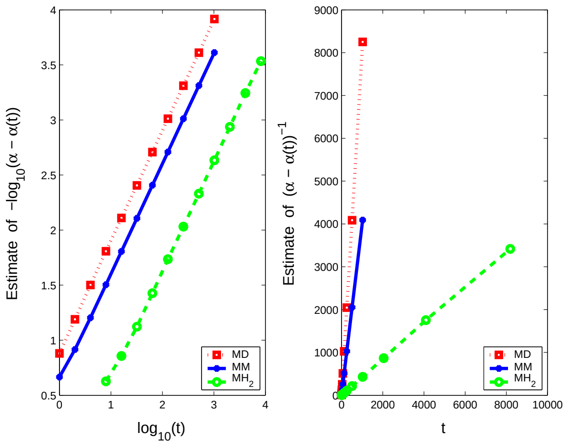

Fig. 4. Regression plots for M/G/1/0.

Table VI. Regression analysis for the $M/G/1/0$ model.

	log-log			linear	
	est. slope	$-\hat{\beta}$	R^2	$-\hat{\beta}$	R^2
MD	1.0058	0.1286	1.0000	0.1243	1.0000
MM	0.9892	0.2372	0.9997	0.2502	1.0000
MH_2	0.9827	2.1043	0.9993	2.3869	0.9999

The following is the same analysis for M/G/5/0. The fit is less precise for this

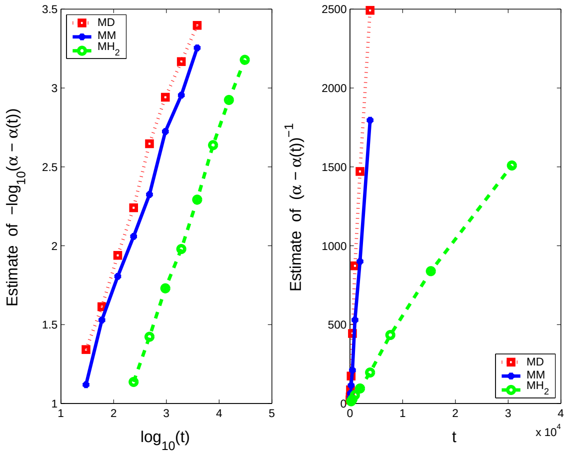

Fig. 5. Regression plots for M/G/5/0.

Table VII. Regression analysis for the $M/G/5/0$ model.

	log-log			linear	
	est. slope	$-\hat{\beta}$	R^2	$-\hat{\beta}$	R^2
MD	1.0111	1.4064	0.9947	1.5164	0.9856
MM	0.9924	1.95162	0.9966	2.1398	0.9978
MH_2	0.9823	16.2368	0.9990	20.0852	0.9965

more complicated model, but the high R^2 values confirm the effectiveness of the procedure.

12. CONCLUDING REMARKS

In this paper we considered how to efficiently apply computer simulation to estimate $\alpha(t) \equiv E[Y(t)]$ for the time average $Y(t)$ in (1), where the underlying stochastic process X is a regenerative process, when the limit α is known and time 0 is a regeneration time. We proposed the *residual-cycle estimator* $\alpha_n^r(t)$ in (23). We established limits for the variances of the standard estimator and the residual-cycle estimator in Theorems 2.1 and 3.3, implying that the residual-cycle estimator is *asymptotically more efficient* than the standard estimator as $t \to \infty$ by a factor of t. In Section 4 we showed that the residual-cycle estimator can be regarded as a *control-variable estimator* using the control variate in (35), and that it approaches the *asymptotically optimal* control-variable estimator as $t \to \infty$.

In Section 5 we established the asymptotic-bias limit in (3). In Section 11 we showed how our estimators of $\alpha(t)$, together with ordinary linear regression, can be used to efficiently estimate the bias parameter β. For our experiments, approximation (4) performed remarkably well. It would be interesting to see what happens for more complicated examples, e.g., with heavy-tailed distributions (where the moment conditions here are satisfied, without higher moments being finite). Then the two-term asymptotic expansion in (74) will not hold, even though (3) remains valid.

In Section 6 we proposed two other efficiency-improvement techniques to use with the residual-cycle estimator: *overlapping cycles* and *splitting*. Overlapping cycles can also be used with the standard estimator, but that combination did not provide any benefit in the simulation experiments. We analyzed the splitting of the residual-cycle estimator and showed, for any given model, that the efficiency is a *concave function* of the number p of splittings with a unique maximum, and derived an expression for the *optimal number of splittings* $p^*(t)$ in (61), which is a function of t, appearing to be roughly proportional to $\sqrt{t}$. In equation (76) we gave an explicit formula for $p^*(t)$ in the $M/M/1/0$ model. If, as an approximation, we assume that the time-dependent probability $z(t)$ there can be taken to be a constant for large t, then indeed we have $p^*(t) \approx \sqrt{Kt}$ for some constant K. The simulation experiments confirm that the efficiency gains by splitting the residual-cycle estimator, for fixed t first increase in p and then decrease, as predicted by the analysis. Moreover, the optimum number increases in t as well.

We conducted simulation experiments to study how the various simulation estimators perform. Extensive simulation experiments for the $M/G/1/0$ queue showed that these residual-cycle methods are indeed effective when t is sufficiently large. Using actual cpu time to represent the computational effort required, i.e., our efficiency measure $\mathcal{E}$, we deduce that the most efficient method for large t is the combination of all three efficiency-improvement techniques; see Tables II and IV. For more details, see Kang et al. [2005].

However, in that example, t had to be above a *threshold*, in particular, so that $\alpha(t)$ was within 10%-20% of the limit $\alpha = 0.5$, before any efficiency gain was realized. For very large t such as $t = 2^{10} = 1024$, the relative efficiency of the estimator using all three techniques was more than $20,000$. However, at that time, $|\alpha - \alpha(t)| \approx 1/4000$, so in practice we might already be confident that α is a good approximation for $\alpha(t)$. The relative efficiency was about 300 when $\alpha(t)$ was within 1% of the limit.

We also did limited experiments for the $M/G/5/0$ model, and found that the residual-cycle estimator again provides efficiency gains. Thus we conclude that these efficiency-improvement techniques are promising more generally, provided that t is suitably large. The methods here seem especially promising to estimate the rate of convergence of $\alpha(t)$ to α as $t \to \infty$.

It would be interesting to compare the techniques in this paper with previous efficiency-improvement techniques proposed by Glynn and Wong [1996] and Henderson and Glynn [2002].

12.1 Non-regenerative initial states.

We may want to consider the transient behavior of a system that does not start in a convenient regeneration state. We can still exploit the regenerative structure, but to do so we need to do it after the first regeneration time $T(1) > 0$. For that purpose, let the *residual-cycle statistic with non-regenerative initial state* $R^{(i)}(t)$ be just $Y(t)$ if $T(1) > t$. However, if $T(1) \leq t$, then we let

$$R^{(i)}(t) = t^{-1}[C(T(1)) + \alpha + \alpha T_+(t - T(1)) - I_+(t - T(1))] \; , \qquad (81)$$

where $C(t) \equiv tY(t)$ is the cumulative process in (24).

Just as with the residual-cycle estimator, the final residual-cycle estimator with non-regenerative initial state is the sample mean

$$\hat{\alpha}_n^{r,i}(t) \equiv n^{-1} \sum_{j=1}^{n} R_j^{(i)}(t) \ , \tag{82}$$

where $R_j^{(i)}(t)$ is the modified residual-cycle statistic from the j^{th} replication.

We expect this extension to perform well if t is large compared to $E[T(1)]$. Indeed, the theorems describing the asymptotic behavior as $t \to \infty$ can be extended to this setting. The unresolved issue is the actual performance for finite t. Unlike the residual-cycle estimator starting at a regeneration time, this estimator is biased for each t. As part of the asymptotic analysis, we can show that the bias is asymptotically negligible as $t \to \infty$, but the performance of this estimator for typical times of interest requires testing. In addition to the variance/efficiency issue, which we have focused on in this paper, we need to check the effect of the bias.

REFERENCES

ASMUSSEN, S. 2003. *Applied Probability and Queues*, second edition Wiley, New York.

BILLINGSLEY, P. 1999. *Convergence of Probability Measures*, second ed., Wiley, New York.

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

CALVIN, J. M. AND NAKAYAMA, M. K. 1998. Using permutations in regenerative simulations to reduce variance. *ACM Transactions on Modeling and Computer Simulation 8*, 153–193.

CALVIN, J. M. AND NAKAYAMA, M. K. 2000. Central limit theorems for permuted regenerative estimators. *Operations Research 48*, 776–787.

CHOW, Y. S., HSIUNG, C. A. AND LAI, T. L. 1979. Extended renewal theory and moment convergence in Anscombe's theorem. *Ann. Probab. 7*, 304–318.

CHUNG, K. L. 1974. *A Course in Probability Theory*, second edition, Academic Press.

CRANE, M. A. AND IGLEHART, D. L. 1975. Simulating stable stochastic systems III: regenerative processes and discrete event simulation. *Operations Research 23*, 33–45.

DUFFIELD, N. G. AND WHITT, W. 2000. Network design and control using on-off and multilevel source traffic models with heavy-tailed distributions. Chapter 17 in *Self-Similar Network Traffic and Performance Evaluation*, Kihong Park and Walter Willinger, eds., John Wiley and Sons, 421-445.

ERDÉLYI, A. 1956. *Asymptotic Expansions*, Dover.

GLYNN, P. W. AND IGLEHART, D. L. 1987. Consequences of uniform integrability for simulation. unpublished paper, Stanford University.

GLYNN, P. W. AND WHITT, W. 1992. The Asymptotic Efficiency of Simulation Estimators. *Operations Res.*, *40*, 505–520.

GLYNN, P. W. AND WHITT, W. 1993. Limit theorems for cumulative processes, *Stoch. Proc. Appl. 47*, 299–314.

GLYNN, P. W. AND WHITT, W. 2002. Necessary conditions in limit theorems for cumulative processes, *Stoch. Proc. Appl. 98*, 199–209.

GLYNN, P. W. AND WONG E. W. 1996. Efficient simulation via coupling. *Prob. Engr. Inf. Sci. 10*, 165–186.

GUT, A. 1988. *Stopped Random Walks*, Springer.

HAMMERSLEY, J. M. AND HANDSCOMB, D. C. 1964. *Monte Carlo Methods*, Methuen, London.

HENDERSON, S. G. AND GLYNN, P. W. 2002. Approximating martingales for variance reduction in Markov process simulation. *Math. Opns. Res. 27*, 253–271.

KANG, W., SHAHABUDDIN, P. AND WHITT, W. 2005. Exploiting regenerative structure to estimate finite time averages via simulation: supplementary material. Available at: http://www.columbia.edu/∼ww2040/

KEILSON, J. 1979. *Markov Chain Models - Rarity and Exponentiality*, Springer.

MEKETON, M. S. AND HEIDELBERGER, P. 1982. A renewal theoretic approach to bias reduction in regenerative simulations. *Management Science 28*, 173-181.

MEKETON, M. S. AND SCHMEISER, B. 1984. Overlapping batch means; something for nothing? *Proceedings of 1984 Winter Simulation Conference*, 227–230.

MEYN, S. P. AND TWEEDIE, R. L. 1993. *Markov Chains and Stochastic Stability*, Springer.

PAWLIKOWSKI, K. 1990. Steady state simulation of queueing processes; a survey of problems and solutions. *ACM Computing Surveys 22*, 123–170.

ROSS, S. M. 1996. *Stochastic Processes*, second edition, John Wiley and Sons.

ROSS, S. M. 2003. *Introduction to Probability Models*, eighth edition, Academic Press, New York.

SMITH, W. L. 1955. Regenerative stochastic processes. *Proceedings of Royal Society of London A232*, 6–31.

WHITT, W. 1972. Embedded renewal processes in the GI/G/s queue. *J. Appl. Prob. 9*, 650–658.

WHITT, W. 1992. Asymptotic formulas for Markov processes. *Operations Res. 40*, 279–291.

WONG E. W., GLYNN, P. W. AND IGLEHART, D. L. 1999. Transient simulation via empirically based coupling. *Prob. Engr. Inf. Sci. 13*, 147–167.

Received March 2005.