MULTILEVEL SPLITTING FOR ESTIMATING RARE EVENT PROBABILITIES

PAUL GLASSERMAN

403 Uris Hall, Columbia Business School, New York, New York 10027

PHILIP HEIDELBERGER

IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, New York 10598

PERWEZ SHAHABUDDIN

Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, perwez@ieor.columbia.edu

TIM ZAJIC

Lockheed Martin, P. O. Box 64525, MS U1P28, St. Paul, Minnesota 55164

(Received June 1996; revisions received April 1997, September 1997; accepted October 1997)

We analyze the performance of a *splitting* technique for the estimation of rare event probabilities by simulation. A straightforward estimator of the probability of an event evaluates the proportion of simulated paths on which the event occurs. If the event is rare, even a large number of paths may produce little information about its probability using this approach. The method we study reinforces promising paths at intermediate thresholds by splitting them into subpaths which then evolve independently. If implemented appropriately, this has the effect of dedicating a greater fraction of the computational effort to informative runs. We analyze the method for a class of models in which, roughly speaking, the number of states through which each threshold can be crossed is bounded. Under additional assumptions, we identify the optimal degree of splitting at each threshold as the rarity of the event increases: It should be set so that the expected number of subpaths reaching each threshold remains roughly constant. Thus implemented, the method is provably effective in a sense appropriate to rare event simulations. These results follow from a branching-process analysis of the method. We illustrate our theoretical results with some numerical examples for queueing models.

INTRODUCTION

The estimation of rare event probabilities poses some of the most difficult computational challenges for Monte Carlo simulation and, at the same time, some of the greatest opportunities for efficiency improvement through the use of variance reduction techniques. Current interest in rare events stems primarily from developments in computer and communications technology: Many industrial and scientific applications require highly reliable computer systems (with correspondingly small failure probabilities), and standards for emerging telecommunications systems call for extremely small buffer-overflow probabilities. The performance of these types of systems is frequently studied through simulation, but straightforward simulation can easily produce estimates that are off by orders of magnitude in estimating small probabilities. In these settings, variance reduction is essential.

Importance sampling, based on changing probability distributions to make rare events less rare, has been used to obtain dramatic improvements in efficiency in estimating small probabilities in queueing and reliability systems; see Asmussen and Rubinstein (1995), Heidelberger (1995), and Shahabuddin (1995) for surveys. But the effectiveness of importance sampling depends critically on the ability to find the right change of measure; indeed, used improperly, importance sampling is liable to produce worse results than straightforward simulation. Finding the right change of measure generally requires identifying at least the rough asymptotics of a rare event probability, often described by a large deviations result. This type of analysis can be formidable in complex models, so the domain of importance sampling, while substantial, does not include all problems of interest.

Villén-Altamirano and Villén-Altamirano (1991) describe an alternative method for rare event simulation that appears to require rather little analysis or model structure for its applicability. Their method, called RESTART, can be viewed as an application of a classical idea in variance reduction called *splitting* (see, e.g., Hammersley and Handscomb 1964, especially p. 131). It is closely related to methods described by Kahn and Harris (1951), Bayes (1970), and Hopmans and Kleijnen (1979). (Regrettably, Bayes called his version "importance sampling," in conflict with standard terminology.) Recent investigations are reported in Schreiber and Görg (1994), Villén-Altamirano et al. (1994), and Villén-Altamirano and Villén-Altamirano (1994); Shahabuddin (1995) gives a brief survey. The essence of the method is captured in the following description from Kahn

Subject classifications: Simulation, efficiency: simulation of rare events by splitting. Queues, simulation: Simulation of rare events in queues.
Area of review: SIMULATION

0030-364X/99/4704-0585 $05.00

and Harris (1951, p. 28), in the setting of particle transmission: "Whenever a particle passes from a less important to a more important region, it is split in two. Each of the resulting particles is given one-half the weight of the original particle and is treated independently from then on." The purpose of this paper is to describe a class of models and implementation conditions under which this type of method is provably effective and even optimal (in an asymptotic sense) for rare event simulation.

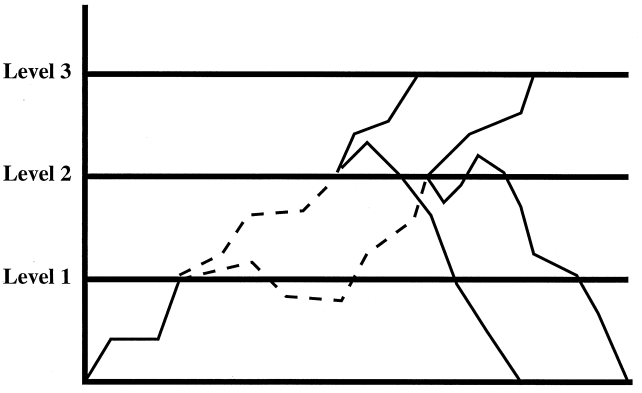
More recent analysis of particular splitting procedures used in nuclear physics appears in Dubi (1985), Dubi et al. (1986), and Burn (1990). The general idea is to derive an expression for the variance, which is quite complicated but which may be simplified by making certain approximations. Terms in the variance are then estimated in pilot studies. The splitting factors to minimize the estimated variance are then computed via numerical optimization.

Melas (1993, 1994) and Ermakov and Melas (1995) also consider particular forms of splitting for Markov chain simulations, similar to the setting considered here although their focus is not on rare event simulations. They derive "quasi-optimal" splitting factors, which are given by solutions to linear equations related to the distribution of the Markov chain over a regenerative cycle. Rare event asymptotics are not examined except in particular cases: the tail of the waiting time distribution in the $GI/G/1$ queue (Melas 1993) and the probability that the $M/M/1$ queue exceeds a fixed queue length as the traffic intensity approaches zero (Melas 1994).

The splitting method we consider is best described through a simple example. Consider the simulation of a nonnegative process that returns to the origin infinitely often—think of the queue-length process in a stable queue. Consider the probability that, starting from the origin, the process reaches some level b before returning to the origin. (As discussed in, e.g., Heidelberger 1995, efficient estimation of this type of probability is central to efficient estimation of the steady-state probability that a queue length exceeds b.) If b is large, this may well be a rare event; starting even a large number of sample paths at the origin may result in very few that reach b before returning, and thus little information about the probability of this event is obtained. To get around this problem we may partition the state space using intermediate thresholds as illustrated in Figure 1, where b corresponds to Level 3. Then, each time a sample path reaches a threshold higher than any it has reached before, we split it into a number of subpaths, which subsequently evolve independently of each other. A path is terminated when it reaches level b or returns to the origin.

Reaching an intermediate level is more likely than reaching b itself, and by splitting at each threshold we reinforce successful outcomes and end up allocating more effort to simulating more promising paths. Dividing the total number of paths that reach b before 0 by the total number potentially started at any level yields an unbiased estimate of the desired probability. (Villén-Altamirano and Villén-Altamirano 1991 and 1994 describe a slightly different implementation in which a path splits every time it crosses a threshold—even one it has reached before. Kahn and Harris 1951 mention both versions.)

Figure 1. Splitting with three levels and two split subpaths.



The central issues in implementing this method are choosing the thresholds and choosing the number of subpaths to generate when a path splits. In this paper, we address only the second issue. Some of our conditions may be interpreted as roughly requiring that the thresholds be eventually nearly evenly spaced. More precisely, we will require that the dynamics of the process between thresholds approach a limit at high thresholds. In a separate paper (Glasserman et al. 1998) we have examined necessary conditions on the choice of thresholds; that analysis involves rather different tools. Indeed, on more general state spaces than those we consider here the term "threshold" may be misleading; we require, in general, a nested sequence of subsets.

Our analysis in this paper is based on modeling the movement from one threshold to the next rather than explicitly modeling the underlying process. Thus, our results may be viewed as an exact analysis of processes for which these models apply literally and an approximate analysis for more general cases. Briefly, we consider three settings allowing for increasing levels of generality:

- On crossing a threshold, the underlying process has a fixed *success* probability p of achieving the next threshold before terminating, independent of its past. Hence, the process that records the highest threshold reached so far is Markov. The requirement that the success probability be independent of the past holds if the underlying process is itself a Markov chain and there is a single entry state for each threshold. If, in addition, the underlying process is spatially homogeneous and the thresholds are evenly spaced, then the success probability is indeed constant.
- The process that records the highest threshold reached so far becomes homogeneous Markov when augmented with a supplementary variable taking on finitely many values. If, for example, the underlying process is Markov and the number of entry states per

threshold is bounded, it suffices to record the highest threshold reached and the index of the state in which it was entered to get a Markov chain. In this setting, the movement from one threshold to the next is described by a matrix of transition probabilities.

- The movement from one threshold to the next is again described by transition probabilities, but we drop the requirement that a single transition matrix apply at all thresholds and replace it with the condition that the transition matrices converge to a limiting matrix.

The last setting is evidently the most general. For a specific example in which it applies, consider a queue in discrete time. Exactly one job is completed at each time increment so long as the system is not empty. Arrivals per time increment are i.i.d. and bounded. Take the underlying process to be the queue length and suppose the thresholds are at $\Delta, 2\Delta, 3\Delta, \ldots$ for some positive integer Δ larger than the greatest number of arrivals possible in a single time increment. Given that the queue length first achieved the threshold at $k\Delta$ by entering state $k\Delta + i$, for some $0 \leq i < \Delta$, the probability that it will achieve the next threshold (before returning to 0) by entering state $(k+1)\Delta + j$, for some $0 \leq j < \Delta$, is independent of the past. The movement from level k to $k+1$ can thus be described by a $\Delta \times \Delta$ transition matrix with entries $P_k(i,j)$, and it is easy to see that these matrices converge as $k \to \infty$.

For each of the settings above we show that appropriately choosing the degree of splitting at each threshold is critical to the effectiveness of the method. The choice must balance two competing concerns: excessive splitting creates an explosive computational burden, and insufficient splitting eliminates the advantage over straightforward simulation. But with just the right amount of splitting, the method becomes *asymptotically optimal* (in a sense reviewed in Section 1) and is thus in some respects as effective for rare event simulation as any method can be. Our main results identify the ideal level of splitting for the three settings above: in the first setting, each path should be split into approximately $1/p$ subpaths; in the second setting the splitting parameter should be the reciprocal of the spectral radius of the transition matrix; and in the third setting it should be the reciprocal of the spectral radius of the limiting transition matrix. Often, this entails randomizing the number of subpaths. We obtain these results by modeling the paths that reach each threshold as the population at subsequent generations of a branching process. They may be loosely interpreted as stating that when a path splits, the number of subpaths should be chosen so that on average one subpath makes it to the next threshold. This keeps the expected number of paths alive at each threshold roughly constant.

We analyze the three settings above in §§1–3. Section 4 reports numerical results supporting the theoretical analysis and exploring the robustness of the method. Section 5 contains some concluding remarks and cautionary observations. Indeed, whereas the results of this paper are essentially positive, it is important to emphasize that they are obtained under restrictions. Our purpose here is to show how well the method works under ideal conditions; in Glasserman et al. (1998) we address some of the limitations of the method, particularly in higher dimensional problems. For a nontechnical overview of the work in this paper and Glasserman et al. (1998), the reader is referred to Glasserman et al. (1998a).

1. THE SIMPLEST SETTING

In this section, we analyze the performance of multilevel splitting in a simplistic model. This setting provides insight into more general cases with minimal notation. Before proceeding with the analysis, we briefly review the general issue of rare event simulation. This discussion is relevant to later sections as well.

Consider a family of events $\{A_k, k = 1, 2, \ldots\}$ with $\gamma_k \triangleq P(A_k) \to 0$ as $k \to \infty$. Think of k as indexing rarity. The most obvious estimator of γ_k is the sample mean of independent copies of the indicator of A_k. By the central limit theorem, the width of an approximate confidence interval for γ_k based on m replications is proportional to the standard error $\sqrt{\gamma_k(1-\gamma_k)/m}$. For small γ_k, this is approximately $\sqrt{\gamma_k/m}$. It follows that the number of replications required to achieve a fixed *relative* error (i.e., to make the confidence interval width a fixed fraction of γ_k) is roughly proportional to $1/\gamma_k$, and thus increases without bound as the rarity parameter k increases.

Consider an alternative family of estimators $\{\hat{\gamma}_k, k = 1, 2, \ldots\}$ that is given by the sample mean of independent copies of some random variable $Y^{(k)}$. The estimator $\hat{\gamma}_k$ is said to have *bounded relative error* if

$$\limsup_{k\to\infty} \frac{\mathrm{Var}(Y^{(k)})}{\gamma_k^2} < \infty, \tag{1}$$

where $\mathrm{Var}(\cdot)$ denotes the variance of the random variable inside the parentheses. When this holds, the number of replications required to achieve a fixed relative error remains bounded as k increases. A weaker requirement is

$$\lim_{k\to\infty} \frac{\log E((Y^{(k)})^2)}{\log \gamma_k} = 2, \tag{2}$$

termed *asymptotic efficiency* or *asymptotic optimality*. This condition may be interpreted as stating that the exponential rate of decrease of the second moment of the estimator is twice that of the first moment. Nonnegativity of variance implies that this is the best possible rate—i.e., the limsup of the ratio on the left side of (2) can never exceed 2. See, e.g., Heidelberger (1995) or Shahabuddin (1995) for background.

The conditions in (1) and (2) reflect the impact of variance but not of computational effort. This is appropriate in comparing estimators with similar computational requirements; but in our setting the effort required can be vastly different depending on the amount of splitting, so it is important to reflect effort directly. A standard measure,

dating at least to Hammersley and Handscomb (1964) and formalized in a general framework by Glynn and Whitt (1992), compares estimators based on the product of variance and expected effort per run. This is the *work-normalized* variance. Comparing work-normalized variances is equivalent to comparing the variance resulting from a fixed computational budget. In light of these considerations, the following work-normalized notion of asymptotic efficiency seems most appropriate for our setting:

$$\lim_{k\to\infty} \frac{\log(\mathrm{Var}(Y^{(k)})w(k))}{\log \gamma_k} = 2, \tag{3}$$

with $w(k)$ denoting the expected computational effort to generate a sample of $Y^{(k)}$. This condition states that the exponential rate of decrease of the work-normalized variance is twice that of the probability γ_k itself. Straightforward simulation (generating replications of the indicator of A_k) has a variance per replication of $\gamma_k - \gamma_k^2$, which is approximately γ_k for large k. If the expected work per replication of an indicator is bounded in k (as would often be the case) then (3) compares the exponential rate of decrease in work-normalized variance for $Y^{(k)}$ and straightforward simulation. If the expected work to generate an indicator of A_k actually increases with k, then the work-normalized variance for straightforward simulation might decrease at a slower rate than that reflected in the denominator of (3). In this respect, (3) is a conservative measure of the performance of $Y^{(k)}$ compared with straightforward simulation.

With these definitions in mind, we now proceed with our first analysis of the effectiveness of multilevel splitting. Suppose we want to estimate $\gamma \triangleq P(A)$ for some event A. Consider intermediate events $A_1 \supset A_2 \supset \cdots \supset A_k = A$ and let $P(A_{i+1}|A_i) = p_i$, $i = 1, \ldots, k-1$, and $P(A_1) = p_1$, so that

$$\begin{aligned}\gamma = \gamma_k \triangleq P(A_k) &= P(A_1)P(A_2|A_1)\cdots P(A_k|A_{k-1}) \\ &= p_i \cdots p_k. \end{aligned} \tag{4}$$

We suppose we have a mechanism, perhaps only implicit, for generating independent Bernoulli trials with success parameter p_i, for each $i = 1, \ldots, k$. We estimate γ_k by first generating R_1 independent Bernoullis with parameter p_1. For each positive outcome of this first stage we generate R_2 independent Bernoullis, with parameter p_2. We continue in this fashion for k iterations and then form the sample

$$Y^{(k)} = \frac{1}{R_1 \cdots R_k} \sum_{i_1=1}^{R_1} \cdots \sum_{i_k=1}^{R_k} 1_{i_1} 1_{i_1 i_2} \cdots 1_{i_1 \cdots i_k}, \tag{5}$$

where $1_{i_1} = 1$ if the i_1th Bernoulli trial at stage 1 is successful, and 0 otherwise; $1_{i_1 i_2} = 1$ if both the i_1th trial at stage 1 and its i_2th subpath are successful, and 0 otherwise, and so on. It is readily verified that this provides an unbiased estimator of $P(A)$. Note that if $R_i = 1$ for $i = 1, \ldots, k$, then the procedure reduces to a single replication of a standard simulation of the rare event.

This is an exact description of the multilevel splitting estimator in the following setting. A real-valued Markov chain starts at the origin and takes only nonnegative values. The event A corresponds to the chain attaining a given value, b say, before returning to the origin. The events A_i correspond to the chain attaining a value b_i before returning to the origin, where $0 < b_1 < \cdots < b_k = b$. Suppose the chain must take the value b_i before taking any value greater than b_i, for $i = 1, \ldots, k$. (For example, the b_i could be integers and the chain integer-valued and *skip-free to the right*.) Then $P(A_i|A_{i-1})$ is the probability of achieving the value b_i, before returning to the origin, starting from b_{i-1}. Thus, we may generate Bernoulli trials with parameter p_{i-1} (without knowing p_{i-1}) by simulating the underlying Markov chain from state b_{i-1}, recording a success if the chain reaches b_i before 0 and a failure otherwise.

The estimator in (5) has a natural description in the language of branching processes. Think of an initial population of size one. The offspring distribution of this individual is Binomial(R_1, p_1). Each member of the first generation in turn has offspring distribution Binomial(R_2, p_2), and so on. The estimator in (5) can be interpreted as the number of individuals in the kth generation divided by the maximum possible number of individuals in the kth generation. (Kahn and Harris 1951 also mention a connection between splitting and branching processes but do not pursue it.)

To further simplify the setting, we assume that the p_i and R_i are constant, denoted by p and R respectively, so that $P(A) = p^k$. With $\mu = Rp$, classical results from branching process theory (see, for example, p. 6 of Harris 1963) yield

$$\mathrm{Var}(Y^{(k)}) = \begin{cases} \dfrac{1}{R^{2k}} \dfrac{\mu^k(\mu^k - 1)}{\mu^2 - \mu}(\mu - \mu p), & \mu \neq 1, \\ k\,\dfrac{\mu - \mu p}{R^{2k}}, & \mu = 1. \end{cases}$$

We will use the following terminology in this and later sections. A nonnegative function $f(k)$ is $O(g(k))$ (resp. $\underline{O}(g(k))$) if for all k large enough $f(k) \leqslant cg(k)$ (resp. $f(k) \geqslant cg(k)$), some constant c. A function $f(k)$ is $\Theta(g(k))$ if it is both $O(g(k))$ and $\underline{O}(g(k))$. Given this, we may write

$$\mathrm{Var}(Y^{(k)}) = \begin{cases} O(p^{2k}), & \mu > 1, \\ O(kp^{2k}), & \mu = 1, \\ O\left(\left(\dfrac{1}{\mu}\right)^k p^{2k}\right), & \mu < 1. \end{cases}$$

Thus, $Y^{(k)}$ is asymptotically optimal, in the sense of (2), in case $\mu \geq 1$.

If $\mu > 1$, the reduction in variance is achieved at the expense of a geometric growth in the number of paths to be simulated, so some accounting for computational effort is essential to a meaningful comparison. We consider two cost models. The simpler of the two assigns constant cost, taken to be unity, to each sample at each level. The second model takes the effort required per sample at level i to

increase linearly with i. (This case is motivated by the Markov example given earlier in this section, where most of the effort at level i might be devoted to simulating failing paths back to 0.) The total expected number of samples is given by

$$R\sum_{i=0}^{k-1}\mu^i=\begin{cases}R\dfrac{(\mu^k-1)}{\mu-1}, & \mu\neq 1,\\ Rk, & \mu=1,\end{cases}$$
$$=\begin{cases}O(\mu^k), & \mu>1,\\ O(k), & \mu=1\\ O(1), & \mu<1.\end{cases}$$

This is the cost under the constant cost model. The cost in the second model is bounded by a proportionality constant times k times the cost in the first model. In both cases, we see that $\hat{\gamma}_k$ performs best, as $k\to\infty$, when $\mu=1$. Indeed, with $\mu=1$ the work-normalized variance satisfies

$$\frac{\log\{\text{work}\times\text{variance}\}}{\log\gamma_k}=\frac{\log\{O(k^\alpha)O(kp^{2k})\}}{\log p^k}\to 2, \qquad (6)$$

with $\alpha=1$ or 2, depending on the cost model. With $\mu\neq 1$ the limit is less than 2 either because the variance is too large ($\mu<1$) or the effort is too great ($\mu>1$). Note that for standard simulation ($R=1$), $\mu=p$ and the work-normalized variance is of order p^k. Thus using (6), we see that the asymptotically optimal splitting estimator is of order $1/(p^k k^{\alpha+1})$ times more efficient than standard simulation. The result that the optimal R satisfies $R=1/p$ is consistent with a recommendation of Villén-Altamirano and Villén-Altamirano (1994) that p and R should be approximately e^{-2} and e^2, respectively, and also with the formulation of Kahn and Harris (1951), in which $p\approx 1/2$ and $R=2$.

Although we would therefore like to choose R so that $Rp=1$, we are constrained to choose R to be a positive integer. To circumvent this constraint, we randomize. Let $R_1,\ldots,R_k$ be random variables, independent of everything else, taking only positive integer values, and having means $m_1,\ldots,m_k$. Then the sample

$$Y^{(k)}=\frac{1}{m_1\cdots m_k}\sum_{i_1=1}^{R_1}\cdots\sum_{i_k=1}^{R_k}1_{i_1}1_{i_1 i_2}\cdots 1_{i_1\cdots i_k}$$

has an expected value of γ_k, as can be seen by a conditioning argument using induction and Wald's equation. Hence the estimator remains unbiased. If all R_i have the same distribution and $m_i=1/p$, we find that the variance becomes

$$\text{Var}(Y^{(k)})=kp^{2k}\times[(1-p)+p\ \text{Var}(R_i)],$$

and in fact (6) continues to hold. To minimize $\text{Var}(R_i)$, we should randomize between the two integers closest to $1/p$, when $1/p$ itself is not an integer, choosing $\lfloor 1/p\rfloor+1$ with probability $(1/p)$ mod 1 and $\lfloor 1/p\rfloor$ with the complementary probability. In a different but related setting, Fox (1997) proposes an alternative to randomization.

2. SPLITTING AS A MULTITYPE BRANCHING PROCESS

We continue to consider the estimation of a probability $\gamma_k=P(A_k)$ decomposed as in (4), but broaden the scope of splitting estimators. In the previous section, we assumed a mechanism for generating Bernoulli trials with each of the probabilities in the decomposition in (4) as parameter. This is motivated by the case in which each A_n represents the entry of a Markov chain into a set with the restriction that there be just one state through which the set can be entered, for then the Bernoulli trials can be generated implicitly by simulating the Markov chain.

We generalize this setting by supposing that each A_n can be expressed as the union of disjoint sets $A_n^{(1)},\ldots,A_n^{(r)}$ in such a way that we have a mechanism for generating independent trials taking the values $0,1,\ldots,r$ with probabilities

$$1-\sum_{j=1}^{r}P(A_{n+1}^{(j)}|A_n^{(i)}),\quad P(A_{n+1}^{(1)}|A_n^{(i)}),$$
$$P(A_{n+1}^{(2)}|A_n^{(i)}),\ \ldots,\ P(A_{n+1}^{(r)}|A_n^{(i)}), \qquad (7)$$

for each $i=1,\ldots,r$, for each level n. The main object of study in this setting becomes the matrix $\mathbf{P}_n$ with entries

$$P_n(i,j)=P(A_{n+1}^{(j)}|A_n^{(i)}).$$

In this section, we assume $\mathbf{P}_n$ does not change with n, and in the next section we assume merely that it converges as $n\to\infty$.

To make this setting more concrete, again think of A_n as the event that a Markov chain enters some set before returning to its initial state, designated 0. The framework above allows the set to be entered through r different states, and $A_n^{(i)}$ is just the event that entry at the nth level first occurs through the ith possible state, before a return to 0. We can generate a trial with outcomes $0,1,\ldots,r$ having the probabilities in (7) by simulating a path of the Markov chain out of the ith entry state at level n until it returns to 0 or reaches the $(n+1)$th level through one of its r possible entry states. Because this Markov chain example is the most vivid one, we will use its associated terminology of levels and states in the analysis that follows. Except for the requirement that r be finite, this setting seems general enough to encompass a substantial portion of the rare event probability problems one encounters in queueing and reliability; see, e.g., Heidelberger (1995).

We proceed, then, with the case in which the probability of going from level n to level $n+1$, before hitting level 0, depends on the state when the process hits level n, but not on n itself. At each level there are r possible entry states and for simplicity we assume that the initial state at level 0 is 1 (in the notation of (7), this corresponds to relabeling the sample space as $A_0^{(1)}$, though we could also allow an initial distribution over $\{1,\ldots,r\}$. In the matrix $\mathbf{P}=(P(i,j))$, interpret $P(i,j)$ as the probability of reaching level $n+1$ through its jth entry state, before hitting level 0, given that level n was reached through its ith entry state. Note

that $\mathbf{P}$ is a substochastic matrix, as starting from level k there may be a positive probability of not hitting level $n+1$ before hitting level 0.

Before proceeding further we introduce some vector and matrix notation. For any matrix $\mathbf{A}$ let $\mathbf{A}'$ denote its transpose and let $\mathbf{A}_{abs}$ be the matrix whose entries are the absolute values of those of $\mathbf{A}$. Let $\mathbf{1}$ denote a column vector of 1's and $\mathbf{e}_i$ a column vector with ith entry 1 and all other entries 0. In this notation, γ_k, the probability of hitting level k before hitting zero, is given by $\gamma_k = \mathbf{e}_1'\mathbf{P}^k\mathbf{1}$.

In the present setting, we model the splitting method as a *multitype* branching process. A standard multitype branching process starts with one individual of type 1 in generation 0. In generation 1 this individual produces a random number of offspring of r types and then dies. Each of these offspring lives for one generation, produces its own offspring, and then dies. The progeny distribution of each individual depends only on its type and not on the generation in which it lives.

In our application, types correspond to entry states. When we split a path into R independent subpaths, the number of subpaths that reach the next level through the ith entry state can be modeled as the number of progeny of type i. Let $(X_1^{(i)}, \ldots X_r^{(i)})$ be the progeny random vector, i.e., $X_j^{(i)}$ is the number of type-j offspring produced by a type-i individual. To specify the progeny distribution in the branching process succinctly, let $X_0^{(i)} = R - \sum_{j=1}^r X_j^{(i)}$ (in the simulation this is the number of subpaths that do not reach the next level before 0). Then the distribution of $(X_0^{(i)}, X_1^{(i)}, \ldots, X_r^{(i)})$ is Multinomial$(R, (1 - \sum_{l=1}^r P(i, l)), P(i, 1), \ldots, P(i, r))$; these are the same cell probabilities as in (7). Let $\mathbf{Z}^{(k)} = (Z_1^{(k)}, Z_2^{(k)}, \ldots, Z_r^{(k)})$ where $Z_i^{(k)}$ is the number of individuals of type i in generation k. In our case $\mathbf{Z}^{(0)} = \mathbf{e}_1$. Then the sample generated using splitting, $Y^{(k)}$, can be expressed as $\sum_{i=1}^r Z_i^{(k)}/R^k$: the total number of subpaths that reach the kth level, divided by the total potential number of paths started at any level.

We use the analogy with multitype branching processes to prove results about γ_k and the splitting estimator. Let $\|\mathbf{A}\|$ denote the sum of the elements of $\mathbf{A}_{abs}$, i.e., $\|\mathbf{A}\| = \mathbf{1}'\mathbf{A}_{abs}\mathbf{1}$. It is easy to show that $\|\cdot\|$ is a matrix norm (see, e.g., p. 291 of Horn and Johnson 1985). In particular, for any two matrices $\mathbf{A}$ and $\mathbf{B}$, $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$. All equalities and inequalities on vectors and matrices will be elementwise.

A fundamental result that we use frequently is a version of the Perron-Frobenius theorem from Harris (1963, p. 37). For convenience we restate the result here.

THEOREM 1. *Let* $\mathbf{A}$ *be an* $r \times r$ *nonnegative matrix such that* $\mathbf{A}^k > 0$ *for some integer* $k > 0$. *Then* $\mathbf{A}$ *has an eigenvalue* λ *that is positive and real and is bigger in absolute value than any other eigenvalue*; λ *corresponds to positive right and left eigenvectors* $\mu = (\mu_i)$ *and* $\nu = (\nu_j)$ *which are the only nonnegative eigenvectors. Moreover, we have* $\mathbf{A}^k = \lambda^k\tilde{\mathbf{A}} + \hat{\mathbf{A}}^k$ *where* $\tilde{\mathbf{A}} = \mu'\nu$ *with the normalization* $\sum_{i=1}^r \mu_i\nu_i = 1$. *Hence* $\tilde{\mathbf{A}}\tilde{\mathbf{A}} = \tilde{\mathbf{A}}$. *Furthermore* $\tilde{\mathbf{A}}\hat{\mathbf{A}} = 0$ *and* $\|\hat{\mathbf{A}}^k\|$ *is* $O(\lambda'^k)$ *where* $0 < \lambda' < \lambda$.

To use this theorem in our setting, we assume that

$$\text{there exists an integer } N > 0 \text{ such that } \mathbf{P}^N > 0. \tag{8}$$

This seems quite harmless in practice, given our interpretation of types. Let ρ be the spectral radius of $\mathbf{P}$—the maximal eigenvalue provided by Theorem 1. We use Theorem 1 to characterize the rate of decrease of γ_k, and for later comparison with the variance and cost of the splitting estimator.

LEMMA 1.

$$\lim_{k\to\infty} \frac{\gamma_k}{\rho^k} \to \mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1} > 0.$$

PROOF. The proof follows easily from Theorem 1. Note that

$$\frac{\gamma_k}{\rho^k} = \mathbf{e}_1'\left(\tilde{\mathbf{P}} + \frac{1}{\rho^k}\hat{\mathbf{P}}^k\right)\mathbf{1} = \mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1} + \frac{1}{\rho^k}\mathbf{e}_1'\hat{\mathbf{P}}^k\mathbf{1}.$$

Now $|\mathbf{e}_1'\hat{\mathbf{P}}^k\mathbf{1}| \leq \mathbf{e}_1'(\hat{\mathbf{P}}^k)_{abs}\mathbf{1} \leq \mathbf{1}'(\hat{\mathbf{P}}^k)_{abs}\mathbf{1} = \|\hat{\mathbf{P}}^k\| = O(\rho'^k)$ where $\rho' < \rho$. Also, since by Theorem 1, $\tilde{\mathbf{P}} = \mu'\nu > 0$, we get the result of the lemma. □

Next we examine the computational effort required by the splitting estimator. We assume that each split path takes unit effort until it hits either a higher level or level 0, though as in Section 1 this is by no means essential, and we could easily study other cost models. For any multitype branching process, we let $\mathbf{M} = (M(i, l))$ denote the expected number of type-l progeny produced by a type-i individual. In our context, it is easy to see that $\mathbf{M} = R\mathbf{P}$. Let $w(k)$ be the expected effort required to simulate for k levels. Then

$$\begin{aligned} w(k) &= R\mathbf{e}_1'\left(\sum_{j=0}^{k-1}\mathbf{M}^j\right)\mathbf{1} \\ &= R\left(\sum_{j=0}^{k-1} R^j\mathbf{e}_1'\mathbf{P}^j\mathbf{1}\right) \\ &= R + R\left(\sum_{j=1}^{k-1} R^j\mathbf{e}_1'(\rho^j\tilde{\mathbf{P}} + \hat{\mathbf{P}}^j)\mathbf{1}\right) \\ &= R + R\left(\sum_{j=1}^{k-1} R^j\rho^j\left(\mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1} + \frac{1}{\rho^j}O(\rho'^j)\right)\right). \end{aligned} \tag{9}$$

We use this to prove the following theorem.

THEOREM 2. (i) *For* $R < 1/\rho$, $w(k) = \Theta(1)$.
(ii) For $R = 1/\rho$, $w(k) = \Theta(k)$.
(iii) For $R > 1/\rho$, $w(k) = \underline{O}((R\rho)^k)$.

Hence, we see that in the first case the effort remains bounded as k increases, in the second case the effort grows linearly in k, and in the third case it grows exponentially.

PROOF. The bound in (ii) is obvious from (9) and the fact that $(\mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1} + \rho^{-j}O(\rho'^j))$ approaches a constant. The upper

bound in (i) follows from the additional fact that the sum of the geometric series $(R\rho)^i$, $i \geq 1$, is bounded for $R\rho < 1$. For the lower bound in (i) it suffices to consider the first term in the summation for $w(k)$, which yields $w(k) \geq R$. For the lower bound in (iii), it suffices to consider the last term in the summation for $w(k)$, i.e.,

$$w(k) \geq R(R\rho)^{k-1}\left(\mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1} + \frac{1}{\rho^k}O((\rho')^k)\right)$$

$$\geq R(R\rho)^{k-1}\frac{\mathbf{e}_1'\tilde{\mathbf{P}}\mathbf{1}}{2},$$

for sufficiently large k. □

To analyze the variance of $Y^{(k)}$, we use an expression for the variance of $\sum_{i=0}^r Z_i^{(k)}$ in a multitype branching processes (from p. 37 of Harris 1963) to get

$$\mathrm{Var}(Y^{(k)}) = \frac{1}{R^{2k}}\sum_{j=1}^{k}\mathbf{1}'(\mathbf{M}')^{k-j}\left(\sum_{i=1}^{r}\mathbf{V}_i(\mathbf{e}_1'\mathbf{M}^{j-1}\mathbf{e}_i)\right)(\mathbf{M})^{k-j}\mathbf{1}. \tag{10}$$

Here, $\mathbf{V}_i = (V_i(l, m)) = (\mathrm{Cov}(Z_l^{(1)}, Z_m^{(1)}|\mathbf{Z}^{(0)} = \mathbf{e}_i))$. When the progeny distribution is multinomial, $V_i(l, l) = RP(i, l)(1 - P(i, l))$ and $V_i(l, m) = -RP(i, l)P(i, m)$ for $l \neq m$. Hence, the $V_i(l, m)$ are functions only of R and $P(i, l)$, for $l = 1, \ldots, r$. The assumption in (8) implies that for each i there exists an l such that $P(i, l) > 0$. Hence, for each i there exists an l such that $\mathrm{Var}(Z_l^{(1)}|\mathbf{Z}^{(0)} = \mathbf{e}_i) = \mathrm{Cov}(Z_l^{(1)}, Z_l^{(1)}|\mathbf{Z}^{(0)} = \mathbf{e}_i) > 0$ so for all i, $\mathbf{V}_i$ is a positive-definite matrix for at least one i, i.e., for all vectors $\mathbf{x} \neq 0$, $\mathbf{x}'\mathbf{V}_i\mathbf{x} > 0$. The following result is proved in the appendix.

THEOREM 3. (*i*) *For* $R < 1/\rho$, $\mathbf{Var}(Y^{(k)})$ *is* $\underline{O}(\rho^{2k}(1/R\rho)^k))$.
(*ii*) For $R = 1/\rho$, $\mathbf{Var}(Y^{(k)})$ *is* $O(k\rho^{2k})$.
(*iii*) For $R > 1/\rho$, $\mathbf{Var}(Y^{(k)})$ *is* $\Theta(\rho^{2k})$.

Combining the asymptotic results for the variance $\mathrm{Var}(Y^{(k)})$ and the effort $w(k)$, we see that we have asymptotic efficiency for the case of $R = 1/\rho$. For $R \neq 1/\rho$ we do *not* have asymptotic efficiency: the simulation effort required to achieve a given relative error grows exponentially with k. Noting that standard simulation ($R = 1$) has a work-normalized variance of order ρ^k, these results imply that asymptotically optimal splitting is of order $1/(\rho^k k^2)$ times more efficient than standard simulation.

3. THE NONHOMOGENEOUS CASE

In the previous section we assumed that the transition probabilities $P(i, l)$ (and thus the progeny distribution in the corresponding branching process) did not depend on level k. As the queueing example in the introduction suggests, a more realistic model allows the transition matrices to depend on the level k but requires that they converge as $k \to \infty$. We now analyze this case.

3.1. Asymptotics for the Probability

As before, $\mathbf{P}_k$ denotes the transition matrix at the kth level. We assume that $\mathbf{P}_k \to \mathbf{P}$ elementwise as $k \to \infty$. We will also need to assume that:

ASSUMPTION 1. *There exists an integer* $N > 0$, *such that* $\mathbf{P}^N > 0$.

ASSUMPTION 2. *For* $j \geq 1$, *for each* i *there exists an* l *such that* $P_j(i, l) > 0$.

As before, we let ρ be the spectral radius of $\mathbf{P}$. The probability $\gamma_k = P(A_k)$ (in the notation of (4)) is now given by

$$\gamma_k = \mathbf{e}_1'\mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k\mathbf{1}.$$

As before, we interpret this as the probability that a Markov chain hits level k before level 0.

Assumption 1 ensures that for each k, the vector $\mathbf{e}_1'\mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k$ has at least one positive component so that $\gamma_k > 0$. In this more general setting, we have the following weaker counterpart of Lemma 1, proved in the appendix.

LEMMA 2.

$$\frac{1}{k}\log(\gamma_k) \to \log(\rho),$$

as $k \to \infty$.

3.2. Asymptotics for the Effort

Let R_j be the number of splits at the jth level and let us assume that $R_j > 0$ for all j and that $R_j \to R$ as $j \to \infty$. In practice, one would probably set $R_j = R$ for all j; however, it is worth considering the more general case, for which we obtain the following theorem.

THEOREM 4. (*i*) $\liminf_{k\to\infty} \log(w(k))/k > 0$ *if* $R > 1/\rho$.
(*ii*) $\lim_{k\to\infty}\log(w(k))/k = 0$ *if* $R = 1/\rho$.
(*iii*) $w(k)$ *is* $\Theta(1)$ *for* $R < 1/\rho$.

Hence in the first case we have an exponential growth in effort, in the second case we have at most a subexponential growth in effort, and in the third case the effort remains bounded. The proof of this theorem is given in the appendix.

3.3. Asymptotics for the Variance

We can modify the expression for the variance for the multitype branching process case to include the nonhomogeneous case as well. Proceeding along the same lines as the derivation in the homogeneous case (Harris 1963, p. 37), we get that

$$\mathrm{Var}(Y^{(k)}) = \frac{1}{(R_1R_2 \cdots R_k)^2}\sum_{j=1}^{k}\mathbf{1}'\mathbf{M}_k' \cdots \mathbf{M}_{j+1}'$$

$$\cdot\left(\sum_{i=1}^{r}\mathbf{V}_i^{(j)}(\mathbf{e}_1'\mathbf{M}_1 \cdots \mathbf{M}_{j-1}\mathbf{e}_i)\right)\mathbf{M}_{j+1} \cdots \mathbf{M}_k\mathbf{1}. \tag{11}$$

Here, $\mathbf{V}_i^{(j)}(l, m) = \mathrm{Cov}(Z_l^{(j)}, Z_m^{(j)}|\mathbf{Z}^{(j-1)} = \mathbf{e}_i)$. Similar to the homogeneous case, the elements of $\mathbf{V}_i^{(j)}$ are polynomial

functions of R_j and $P_j(i, l)$'s. Due to Assumption 2, $\mathbf{V}_i^{(j)}$ is positive definite for all i and j. Since $R_j \to R$ and $\mathbf{P}_j \to \mathbf{P}$ as $j \to \infty$, $\lim_{j\to\infty} \mathbf{V}_i^{(j)}$ exists and it is the same function of R and $P(i, l)$. Let us call the limit matrix $\mathbf{V}_i$. Due to Assumption 1, $\mathbf{V}_i$ is positive definite for all i. An alternative representation of the variance will also be useful:

$$\begin{aligned}\mathrm{Var}(Y^{(k)}) &= \frac{1}{(R_1 R_2 \cdots R_k)^2} \sum_{j=1}^{k} (R_1 \cdots R_{j-1})(R_{j+1} \cdots R_k)^2 \\ &\quad \cdot \mathbf{1}'\mathbf{P}_k' \cdots \mathbf{P}_{j+1}' \left(\sum_{i=1}^{r} \mathbf{V}_i^{(j)} (\mathbf{e}_1' \mathbf{P}_1 \cdots \mathbf{P}_{j-1} \mathbf{e}_i) \right) \\ &\quad \cdot \mathbf{P}_{j+1} \cdots \mathbf{P}_k \mathbf{1}. \end{aligned} \tag{12}$$

We can now summarize the asymptotic behavior of the variance, leaving the proof for the appendix.

THEOREM 5. (*i*) $\liminf_{k\to\infty} \log(\mathrm{Var}(Y^{(k)}))/k > 2\log(\rho)$ *if* $R < 1/\rho$.

(*ii*) $\limsup_{k\to\infty} \log(\mathrm{Var}(Y^{(k)}))/k \leq 2\log(\rho)$ *if* $R = 1/\rho$.

(*iii*) $\lim_{k\to\infty} \log(\mathrm{Var}(Y^{(k)}))/k = 2\log(\rho)$ *if* $R > 1/\rho$.

Hence in the first case we have an exponential growth in the $\mathrm{Var}(Y^{(k)})/\gamma_k^2$, and in the second and third cases we have at most subexponential growth in the $\mathrm{Var}(Y^{(k)})/\gamma_k^2$.

3.4. Asymptotic Efficiency

Using Lemma 2, Theorem 4, and Theorem 5, we see that for $R = 1/\rho$

$$\lim_{k\to\infty} \frac{\log(\mathrm{Var}(Y^{(k)})w(k))}{\log(\gamma_k)} = 2,$$

and for $R \neq 1/\rho$

$$\lim_{k\to\infty} \frac{\log(\mathrm{Var}(Y^{(k)})w(k))}{\log(\gamma_k)} < 2.$$

Hence we have asymptotic efficiency for $R = 1/\rho$ and we do not have asymptotic efficiency for $R \neq 1/\rho$.

Note that this generalizes and complements a result for the $M/M/1$ queue in Melas (1994), which considers estimating the probability that the $M/M/1$ queue exceeds a fixed level k during a cycle. As the traffic intensity $\rho \to 0$, one generates an average of $1/\rho$ splits upon each arrival transition.

3.5. Randomized Splitting

As discussed in Section 1, though ideally we want $R_i = R = 1/\rho$, in practice we are constrained to make the number of subpaths an integer when we split. We circumvent this problem by randomizing, using a level-dependent distribution for the number of subpaths. Let R_j denote a generic random variable having the distribution of the number of subpaths generated from each path that hits level j. The actual number generated from each path is sampled independent of everything else. We use

$$Y^{(k)} = \frac{\text{No. of paths that hit level } k}{E[R_1] \cdots E[R_k]} \tag{13}$$

as our sample.

We now show that $Y^{(k)}$ is unbiased and asymptotically efficient. We can again model its evolution using a branching process but with new $\mathbf{M}_j$ and $\mathbf{V}_i^{(j)}$ matrices corresponding to a new progeny distribution. In particular, $\mathbf{M}_j = E[R_j]\mathbf{P}_j$. Using the same notation as before we can express the $Y^{(k)}$ in (13) as

$$Y^{(k)} = \frac{\sum_{i=1}^{r} Z_i^{(k)}}{E[R_1] \cdots E[R_k]}.$$

Then

$$\begin{aligned} E[Y^{(k)}] &= \frac{1}{E[R_1] \cdots E[R_k]} E\left[\sum_{i=1}^{r} Z_i^{(k)} \right] \\ &= \frac{1}{E[R_1] \cdots E[R_k]} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_k \mathbf{1}) \\ &= \mathbf{e}_1' \mathbf{P}_1 \cdots \mathbf{P}_k \mathbf{1} \\ &= \gamma_k. \end{aligned}$$

Now assume that the R_i converge in distribution to a random variable R as $i \to \infty$. In practice, it is not a restriction to suppose that the support of all the R_i is contained in a finite set (typically, we would randomize between just two points) so that $E[R_i] \to E[R]$. Then it is easy to see that Theorems 4 and 5 hold with R_i replaced by $E[R_i]$ and R replaced by $E[R]$. For the effort, the only change in the basic Equation (17) is that the R_is will now be replaced by $E[R_i]$s. For the variance, the only change in the basic Equation (12) is that the R_i's are replaced by $E[R_i]$s and the $\mathbf{V}_i^{(j)}$ matrices are different from the deterministic number of splits case. In particular, $V_i^{(j)}(l, l) = E[R_j]P_j(i, l)(1 - P_j(i, l)) + \mathrm{Var}[R_j]P_j(i, l)$ and similarly for $V_i(l, l)$. Hence under Assumptions 1 and 2, the new $\mathbf{V}_i^{(j)}$s and $\mathbf{V}_i$s are still positive definite, so this change does not affect the proof of Theorem 5.

4. EXPERIMENTAL RESULTS

In this section, we empirically examine the sensitivity of the splitting estimator's performance with respect to the rarity of the event, number of levels, and number of subpaths per level for several different queueing network models. For the models studied, γ can be determined numerically by solving an appropriate system of linear equations (Chung 1967, §1.9), thereby permitting comparison of simulation with exact results. In this section, γ_b is the probability that a queue length reaches b before returning to the empty state. The thresholds are even spaced at multiples of some Δ, so the event of interest occurs when the kth threshold is reached, with $k = \lceil b/\Delta \rceil$. In the notation of previous sections, we should write γ_k for this probability, but in the context of specific examples it is more convenient to write γ_b with b indicating the queue length

that must be reached, rather than the threshold. Throughout this section, γ_b was numerically computed by solving linear equations and for the purpose of these studies is considered known.

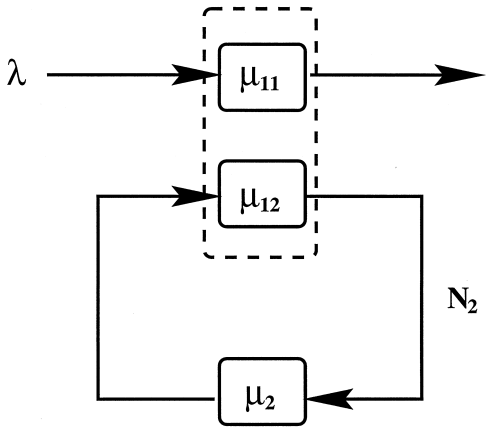
To perform an experiment, i.i.d. replicates $\{Y_j^{(k)}, j \geq 1\}$ of $Y^{(k)}$ may be generated. The point estimate from m such replications is the usual sample average: $\hat{\gamma}_k = \sum_{j=1}^{m} Y_j^{(k)}/m$. We can then build confidence intervals by estimating the variance of $Y^{(k)}$, by using its replicates. We adapt a slightly different procedure that allows us to make more efficient use of the finite time simulation runs that we conduct. For example, in the simplest setting of Section 1, we can interpret $Y^{(k)}$ in (5) as $\sum_{j=1}^{R_1} Y_j^{(k,1)}/R_1$ where $Y_j^{(k,1)}$'s are i.i.d. replicates of $Y^{(k,1)} \equiv (\sum_{i_2=1}^{R_2} \cdots \sum_{i_k=1}^{R_k} 1_1 1_{1i_2} \cdots 1_{1i_2 i_k})/(R_2 \cdots R_k)$; the $Y_j^{(k,1)}$'s track all the split paths originating from a single path starting at level 0. Hence $\hat{\gamma}_k$ may be interpreted as the sample mean of mR_1 i.i.d. replicates of the random variable $Y^{(k,1)}$. We can then build confidence intervals by estimating the variance of $Y^{(k,1)}$ by using its replicates. Henceforth we will call each copy of $Y^{(k,1)}$ as a "replication" (instead of each copy of $Y^{(k)}$). Similar procedures were applied for the more complex settings.

The embedded discrete time Markov chains of the models were simulated and the CSIM (Schwetman 1986) package was used to coordinate the execution of the different trials. CSIM is a C language function library that enables users to develop "process-oriented" models. Specifically, we had a different CSIM process for each threshold and used CSIM messages and mailboxes to exchange information between levels. We used a "depth first" scheduling algorithm in which no pending trial at level k is executed until all the trials at level $k + 1$ are complete. This guarantees that only a linear amount (in the number of levels) of storage will be used. If a "breadth first" algorithm is used, i.e., do not simulate a level $k + 1$ trial until all the level k trials are completed, then the storage requirements will grow exponentially in the supercritical case.

4.1. Mixed Open and Closed Queueing Network

We first consider a mixed open and closed queueing network as shown in Figure 2. The open jobs arrive to service center 1 with rate λ. There are N_2 closed jobs that circulate between service centers 1 and 2. Closed jobs have pre-emptive priority over open jobs at service center 1. All service and interarrival times are exponentially distributed. The service rates at center 1 are μ_{11} and μ_{12} for open and closed jobs, respectively. The service rate at center 2 is μ_2. States are denoted by (i, j) where i is the number of open jobs at center 1 and j is the number of closed jobs at center 1. As j is bounded by N_2, this model fits into the class of nonhomogeneous Markovian models analyzed earlier. Notice that if $N_2 = 1$, this model is equivalent to the $M/M/1$ queue with server breakdowns or vacations. In this model, γ_b is the probability that the number of open jobs reaches b before returning to the state $(0, 0)$, given that the system starts in $(0, 0)$. The thresholds are sets of the form $\{(k\Delta, j), j = 0, 1, \ldots, N_2\}$; thus, there are $N_2 + 1$ entry states for each threshold, making $r = N_2 + 1$ in the branching process formulation.

Figure 2. Mixed open and closed queueing network.



All experiments for this model were run for approximately 500 seconds on a dedicated RS/6000 workstation. More specifically, after a group of replications completed, the CPU time was checked. An experiment terminated when the total CPU time used first exceeds 500 seconds. The group size was chosen (depending on the splitting factor R) so that the CPU time was checked every few seconds, typically every 3 to 5 seconds. This balances the overhead of checking the CPU time against the desire to stop all experiments after exactly the same amount of CPU time has been expended (which would have been harder to implement). Note that at the time the experiment terminates, there are no replications or split paths in progress. The number of replications thus obtained depends on the splitting factor. For example, in the last part of Table 1, 490,000 replications from level 0 are completed when the splitting factor is appropriately chosen ($R = 5$) but only 1,150 replications from level 0 are completed when the splitting factor is too large ($R = 8$).

In Tables 1 and 2, for a given set of model parameters and R, the estimates $\hat{\gamma}_{20}$ and $\hat{\gamma}_{40}$ were obtained from the same (approximately) 500 second run. In Table 3, separate runs were done for each value of b.

As γ_b was obtained numerically, we can use its value to approximate ρ, the spectral radius of the limiting 1 level transition matrix $\mathbf{P}$. This was done as follows. In Table 1, with $b = 40$, we compute $\hat{\rho}$ so that $\gamma_b = \hat{\rho}^b$, i.e., $\log(\hat{\rho}) = \log(\gamma_b)/b$. We use $\hat{\rho}$ as an approximation to ρ (since $\hat{\rho} \to \rho$ as $b \to \infty$). While such numerically obtained estimates of ρ would not be available in most models, we use it here so as to enable sensitivity studies as R deviates from its asymptotically optimal value. The threshold spacing Δ was fixed at 2. Thus $\mathbf{Q}_k = \mathbf{P}_{2k-1}\mathbf{P}_{2k}$ denotes the transition matrix of the process embedded at threshold crossing times. The asymptotically optimal splitting factor is thus one over the spectral radius of the matrix $\mathbf{Q} = \lim_{k\to\infty}\mathbf{Q}_k = \mathbf{P}^2$. Since $\rho(\mathbf{P}^2) = \rho(\mathbf{P})^2 = \rho^2$, the optimal splitting factor is $1/\rho^2$.

Table 1. Results for the mixed open and closed network.

μ_2	b	γ_b	$\hat{R}$	R	$\hat{\gamma}_b$	Relative Error
1.0	20	5.96×10^{-7}	4.1	2	6.35×10^{-7}	±11.5%
				3	5.98×10^{-7}	±3.5%
				4	5.83×10^{-7}	±3.5%
				5	5.76×10^{-7}	±9.7%
				6	5.21×10^{-7}	±31%
				7	1.58×10^{-6}	±130%
1.0	40	5.68×10^{-13}	4.1	2	0	—
				3	5.44×10^{-13}	±15.7%
				4	5.49×10^{-13}	±5.0%
				5	5.40×10^{-13}	±10%
				6	4.92×10^{-13}	±33%
				7	1.52×10^{-12}	±131%
0.5	20	3.91×10^{-8}	5.4	3	3.90×10^{-8}	±8.9%
				4	3.91×10^{-8}	±4.0%
				5	3.86×10^{-8}	±3.6%
				6	3.91×10^{-8}	±6.1%
				7	3.88×10^{-8}	±12%
				8	2.92×10^{-8}	±33%
				9	4.42×10^{-8}	±115%
0.5	40	2.02×10^{-15}	5.4	3	7.68×10^{-16}	±182%
				4	2.10×10^{-15}	±17%
				5	1.98×10^{-15}	±6.2%
				6	2.06×10^{-15}	±7.1%
				7	1.96×10^{-15}	±13%
				8	1.49×10^{-15}	±34%
				9	2.92×10^{-15}	±113%

$\Delta = 2$. Note: Network parameters are $\lambda = 1.0$, $\mu_{11} = 4.0$, $\mu_{12} = 2.0$, $N_2 = 1$.

Thus, as an approximation to the optimal splitting parameter, with $\Delta = 2$, we should start $\hat{R} = 1/\hat{\rho}^2$ trials per level. The table examines the performance of the method for different values of R. In agreement with theory, the best results are obtained when R is close to $\hat{R}$.

In Table 2, we vary the number of levels by varying Δ, the amount by which the queue length has to increase to reach the next level. We attempt to use a near optimal number of subpaths per level by setting $R = 1/\hat{\rho}^{\Delta}$, rounded to the nearest integer. Note that, with one exception, the results are uniformly good and relatively insensitive to Δ. The exception is for $\mu_2 = 0.5$ and $b = 40$ with $\Delta = 1$. The problem in this case is one of rounding. Here $\hat{R} = 1/\hat{\rho} = 2.33$. Rounding $\hat{R}$ to the nearest integer yields $R = 2$. However, with this value of R and so many levels, reaching level 40 is a relatively rare event. The estimated $E[Z_k] = 0.001$ for $b = 40$, suggesting that the number of subpaths surviving to successive levels is dropping to zero. The problem is corrected by using a random, state-independent, number of splits per level. For example, if we fix $E[R] = 2.33$ and select $R = 3$ with probability 0.33 and $R = 2$ with probability 0.67, then our estimate is $\hat{\gamma}_b = 2.06 \times 10^{-15} \pm 6.3\%$. With this random splitting parameter, $E[Z_k]$ increases to 0.44, suggesting that the number of subpaths surviving to successive levels remains roughly $O(1)$.

Table 2. Results with a near optimal splitting parameter for the mixed open and closed network.

μ_2	b	γ_b	Δ	$\hat{\gamma}_b$	Relative Error
1.0	20	5.96×10^{-7}	1	5.98×10^{-7}	±3.9%
			2	5.83×10^{-7}	±3.5%
			4	5.90×10^{-7}	±4.2%
1.0	40	5.68×10^{-13}	1	5.67×10^{-13}	±5.5%
			2	5.49×10^{-13}	±5.0%
			3	5.56×10^{-13}	±5.4%
			4	5.46×10^{-13}	±6.1%
0.5	20	3.91×10^{-8}	1	3.84×10^{-8}	±4.4%
			2	3.86×10^{-8}	±3.6%
			3	3.78×10^{-8}	±6.1%
0.5	40	2.02×10^{-15}	1	2.05×10^{-16}	±20.2%
			2	1.98×10^{-15}	±6.2%
			3	2.00×10^{-15}	±6.8%
			4	1.89×10^{-15}	±8.3%

Note: Network parameters are $\lambda = 1.0$, $\mu_{11} = 4.0$, $\mu_{12} = 2.0$, $N_2 = 1$.

In most problems ρ is not known and must be estimated. Furthermore, the problem of nonintegral ρ is generic so that an asymptotically optimal procedure must use a random number of splits. Table 3 indicates the effect that errors in estimating ρ would have within this context. In this table, $\Delta = 2$ and R has the two point distribution concentrating its mass on the two integers on either side of $E[R] = 1/[\delta\hat{\rho}]^2$ for differing values of δ. Here δ represents the error in an estimate of the near optimal value of $\hat{\rho}$ that might be obtained from a pilot study ($\delta = 1.0$ represents no error). Since $\hat{\rho}$ was obtained numerically, no pilot studies actually needed to be performed; however, the effect of

Table 3. Results for the mixed open and closed network $\Delta = 2$ and $E[R] = 1/[\delta\hat{\rho}]^2$.

δ	$b = 40$ $\gamma_b = 2.02 \times 10^{-15}$	$b = 60$ $\gamma_b = 1.04 \times 10^{-22}$	$b = 80$ $\gamma_b = 5.40 \times 10^{-30}$
0.85	$1.90 \times 10^{-15} \pm 25\%$	$2.49 \times 10^{-22} \pm 141\%$	$1.06 \times 10^{-29} \pm 258\%$
0.90	$1.98 \times 10^{-15} \pm 12\%$	$1.03 \times 10^{-22} \pm 39\%$	$5.72 \times 10^{-30} \pm 107\%$
0.95	$2.01 \times 10^{-15} \pm 7.3\%$	$1.02 \times 10^{-22} \pm 14\%$	$6.28 \times 10^{-30} \pm 25\%$
1.00	$1.97 \times 10^{-15} \pm 5.9\%$	$1.06 \times 10^{-22} \pm 9.2\%$	$5.30 \times 10^{-30} \pm 12\%$
1.05	$2.02 \times 10^{-15} \pm 6.6\%$	$1.04 \times 10^{-22} \pm 12\%$	$5.30 \times 10^{-30} \pm 19\%$
1.10	$2.01 \times 10^{-15} \pm 9.9\%$	$1.11 \times 10^{-22} \pm 24\%$	$7.31 \times 10^{-30} \pm 57\%$
1.15	$2.08 \times 10^{-15} \pm 16\%$	$1.53 \times 10^{-22} \pm 61\%$	$1.32 \times 10^{-30} \pm 57\%$

Note: Network parameters are $\lambda = 1.0$, $\mu_{11} = 4.0$, $\mu_{12} = 2.0$, $\mu_2 = 0.5$, $N_2 = 1$.

poor estimates of ρ can be studied by appropriately setting δ. In this table, a separate run is made for each value of b. For the near optimal runs ($\delta = 1.0$), notice the slow increase in the relative errors as b increases. In addition, the relative error is roughly comparable to the corresponding entry in Table 1 in which a near optimal constant splitting parameter is used ($b = 40$, $\mu_2 = 0.5$, $R = 5$). Also, notice that the performance of the method is more sensitive to δ as b increases. Thus, it becomes more important to obtain good estimates of ρ as the event of interest becomes rarer.

The results reported in Tables 1–3 all used a truncation procedure to discard unpromising trials. Specifically, consider a trial that starts with a queue length of i. If that trial ever reaches a queue length of $(i - d)$ for some specified value of d, then the trial is discarded and counted as a failure. (Our experiments used $d = 10$.) This introduces some error since there is always some possibility that the trial will rebound and reach the next level. This error, which was analyzed for special cases in Glasserman et al. (1996), is small since the confidence intervals all contain the true value of γ_b when the number of splits is appropriately chosen. If d is fixed, the expected amount of work per trial is constant. Without truncation, the expected amount of work per trial from level k is of order k since, with positive probability bounded away from zero, the queue will return to 0 before reaching the next level. In an asymptotically optimal splitting procedure, truncation then reduces the total expected work per replication from order b^2 to order b.

To see the numerical effect of truncation, we re-ran the near optimal ($\delta = 1.0$) simulations of Table 3 without truncation (i.e., $d = \infty$). Because we held the total CPU time fixed, truncation produced a larger number of replications. Compared to the $d = 10$ run, the $b = 40$ relative error increases from 5.9% to 8.0%, while the $b = 80$ relative error increases from 12% to 22%. For queueing models the benefit of truncation increases with the number of levels. For $b = 80$ the run with truncation executed 3.4 times as many replications (from $(0, 0)$), and reduced the average number of transitions per trial from 75 to 17. Interestingly, the $d = \infty$ run simulated a total of 87 million transitions compared to only 66 million transitions for the $d = 10$ run. With $d = 10$ a greater percentage of the time is spent in overheads such as process switching and state space copying. However, that is time well spent compared to executing an unpromising trial on its long way back to $(0, 0)$.

4.2. A Queue with On-Off Sources

The second model we consider is a queueing model with multiple Markov-modulated sources of the type arising in models of ATM (Asynchronous Transfer Mode) networks. There are N on-off sources that operate as follows. If a source is in the on state, the packet arrival rate is λ while if it is in the off state, the arrival rate is 0. The source remains in the off (on) state for an exponentially distributed amount of time with rate α_0 (α_1). The service rate is μ. The state of the system is denoted by (i, j) where i is the total queue length and j is the number of sources in the on state. The steady-state average number of sources in the on state is $\bar{m} = N\alpha_0/(\alpha_0 + \alpha_1)$ and the overall utilization is $u = \bar{m}\lambda/\mu$. We fix $\mu = 10$, $\alpha_0 = 1$, $N = 20$, and $u = 0.25$ and vary the "burstiness" of the sources by varying λ (this uniquely determines α_1). This is a model fitting our theoretical framework.

We let the initial state be $(0, \lfloor \bar{m} \rfloor)$ and consider estimating γ_b, the event that the queue length reaches b before returning to the initial state. Again, γ_b is computed numerically allowing us to approximate ρ by $\hat{\rho}$ where $\gamma_b = \hat{\rho}^b$. Table 4 examines the method's sensitivity to $E[R]$, again representative of the situation in which the ρ is estimated with some error from pilot studies. We fix $\Delta = 2$ and use a random number of splits with the two-point distribution on either side of $E[R] = 1/[\delta\hat{\rho}]^2$. All runs were for approximately 500 seconds on an RS/6000 workstation. As in the mixed open and closed network, the results are very satisfactory if there is no error in estimating $\hat{\rho}$ ($\delta = 1.0$) but degrade as the error increases.

4.3. Tandem Jackson Network

We next consider a queueing network model that has received considerable attention in the importance sampling literature (Anantharam et al. 1990, Frater et al. 1991, Glasserman and Kou 1995, Heidelberger 1995, Parekh and Walrand 1989). The model is an open tandem Jackson network with two queues. Let λ be the arrival rate and let ρ_i denote the utilization at queue i. The state space is denoted by (i, j) where i (j) is the number of jobs at queue 1 (2). Since i and j can both be simultaneously large, our earlier results do not strictly apply. In this example, $\gamma_b =$

Table 4. Results for the queue with on-off sources, $\Delta = 2$ and $E[R] = 1/[\delta\hat{\rho}]^2$.

λ	γ_b	δ	$\hat{\gamma}_b$	Relative Error
0.5	2.33×10^{-21}	0.8	1.96×10^{-21}	±155%
		0.9	3.03×10^{-21}	±39%
		1.0	2.29×10^{-21}	±16%
		1.1	2.47×10^{-21}	±19%
		1.2	1.94×10^{-21}	±51%
4.0	2.33×10^{-20}	0.8	1.29×10^{-20}	±129%
		0.9	2.18×10^{-20}	±34%
		1.0	2.33×10^{-20}	±13%
		1.1	2.14×10^{-20}	±17%
		1.2	2.93×10^{-20}	±42%

Note: Model parameters are b = 40, N = 20, μ= 10.0, and $\alpha_0 = 1.0$. $\alpha_1 = 3.0$ if $\lambda = 0.5$ and $\alpha_1 = 31.0$ if $\lambda = 4.0$.

$P(A_b)$ with A_b denoting the event that the number of jobs in the second queue reaches b before the system empties, given that the system is initially empty.

The runs reported in Table 5 all used $\Delta = 2$ and $R = 4$ since the numerically computed $\hat{\rho}$ is very close to $\rho_2 = 0.5$ for all cases considered. All runs were for approximately 1000 seconds on a dedicated RS/6000 workstation, with a different run for each b.

If the second queue is the bottleneck, i.e., $\rho_1 < \rho_2$, we do not expect the first queue to be large on A_b. Thus the behavior of the system is approximated by one in which i is bounded and we therefore anticipate that splitting will work well when queue 2 is the bottleneck.

However, if queue 1 is the bottleneck, results in Anantharam et al. (1990), Frater et al. (1991), and Heidelberger (1995) based on time-reversal strongly suggest that the way A_b happens is for queue 1 to first build up to a certain large level and then for queue 2 to build up as queue 1 drains down. If this is indeed the most likely overflow path, then the assumption that the state space is finite in one dimension breaks down, and we should have $E[\bar{Q}_1|A_b] \to \infty$ where $\bar{Q}_1$ denotes the maximum length of queue 1 during a cycle. Furthermore, selecting intermediate sets of the form $A_i = \{Q_2 \geq i\}$ (where Q_j denotes the queue length at node $j, j = 1, 2$) would be inconsistent with the presumed large deviations behavior. More specifically, the distribution of queue 1 upon entrance to $A_{[b/2]}$ would be concentrated near 0, however the distribution of queue 1 upon entrance to $A_{[b/2]}$ given that A_b is eventually hit would be concentrated about the point cb for some positive c. Thus, the splitting procedure would be starting most of its trials from level $[b/2]$ with Q_1 near 0 when in fact it should be starting them from near level cb. We therefore do not expect splitting to work well in this situation. We verify this empirically in this paper and study this issue theoretically in a different paper, Glasserman et al. (1998).

Indeed, if queue 2 is the bottleneck, the method works well. However, if queue 1 is the bottleneck, the relative errors increase dramatically with b. For $b = 100$ the point estimate is also too low by a factor of 40.

5. CONCLUDING REMARKS

We have analyzed the use of multilevel splitting in estimating rare event probabilities. Our results show that for problems with a certain structure, choosing the degree of splitting correctly produces asymptotically optimal estimates; whereas too much splitting results in explosive computational requirements, and too little splitting eliminates any reduction in variance. Numerical results support these conclusions, but also suggest that the method is reasonably robust to approximating the optimal splitting parameter. Robustness is important because in practice one would not know the spectral radius that determines the optimal level of splitting. This suggests that pilot runs could be used to get a rough estimate of γ_k from which ρ can be approximated. Some promising experimental results on the effectiveness of using pilot studies in the context of the RESTART procedure are reported in Kelling (1996), although this is an area requiring further study.

Of the restrictions we imposed to obtain our results, the most significant is the requirement that there be only finitely many ways of achieving each threshold. As the examples of Section 4 indicate, this generally restricts us to models whose state spaces are infinite in only one dimension. In these examples, the event of interest becomes rare along the one unbounded dimension and the thresholds are defined along this dimension as well. This structure is

Table 5. Results for the two queue tandem Jackson networks, $\Delta = 2$ and $R = 4$.

ρ_1	ρ_2	b	γ_b	$\hat{\gamma}_b$	Relative Error
0.25	0.50	20	1.27×10^{-6}	1.25×10^{-6}	±2.4%
0.25	0.50	60	1.16×10^{-18}	1.10×10^{-18}	±9%
0.25	0.50	100	1.05×10^{-30}	9.73×10^{-31}	±20%
0.75	0.50	20	3.82×10^{-6}	3.85×10^{-6}	±3.9%
0.75	0.50	60	3.47×10^{-18}	9.61×10^{-19}	±73%
0.75	0.50	100	3.16×10^{-30}	7.85×10^{-32}	±188%

central to our analysis and, perhaps, to the method itself. On more complex state spaces—even the two-queue example of Section 4.3—it is not always clear how intermediate thresholds should be defined to estimate a rare event probability. Results in Glasserman et al. (1998) suggest that the proper implementation of splitting in higher dimensions requires an understanding of the way rare events occur, i.e., the large deviations behavior, not unlike what is often needed to use importance sampling.

APPENDIX: PROOFS

PROOF OF THEOREM 3. To simplify notation, we define $\mathbf{M}_c \equiv \mathbf{P}/\rho$. The spectral radius of $\mathbf{M}_c$ is 1. From Theorem 1 we find that $\mathbf{e}_1'\mathbf{M}_c^j\mathbf{e}_i \leq \mathbf{1}'\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^j)$, with $0 < \lambda < 1$. Define $\mathbf{V}_{abs} = (V_{abs}(l, m))$ where $V_{abs}(l, m) = r \max_{1\leq i\leq r} V_{i,abs}(l, m)$.

First consider the case where $R = 1/\rho$, so that $\mathbf{M} = \mathbf{M}_c$. then from (10) we get

$$\begin{aligned}\mathrm{Var}(Y^{(k)}) &= \rho^{2k}\sum_{j=1}^{k}\mathbf{1}'(\mathbf{M}_c')^{k-j} \\ &\quad\cdot\left(\sum_{i=1}^{r}\mathbf{V}_i(\mathbf{e}_1'\mathbf{M}_c^{j-1}\mathbf{e}_i)\right)(\mathbf{M}_c)^{k-j}\mathbf{1} \qquad (14)\\ &\leq \rho^{2k}\sum_{j=1}^{k}\|\mathbf{M}_c'^{\,k-j}\mathbf{V}_{abs}\mathbf{M}_c^{k-j}\|(\mathbf{a}'\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^j)) \qquad (15)\\ &\leq \rho^{2k}\sum_{j=1}^{k}\|\mathbf{V}_{abs}\|\cdot\|\mathbf{M}_c^{k-j}\|^2(\mathbf{1}'\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^j)) \\ &\leq \rho^{2k}\sum_{j=1}^{k}\|\mathbf{V}_{abs}\|(\mathbf{1}'\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^{k-j})) \\ &\quad\cdot(\mathbf{1}'\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^j)).\end{aligned}$$

Note that the $O(\lambda^j)$ terms are bounded for all $j \geq 0$. Hence the $O((\lambda)^{k-j})$ terms are bounded for all $k \geq 1$ and $0 \leq j \leq k$. Hence, we have $\mathrm{Var}(Y^{(k)}) = O(\rho^{2k}k)$.

Now consider the case $R \neq 1/\rho$. In that case we can express $\mathrm{Var}(Y^{(k)})$ as

$$\begin{aligned}\mathrm{Var}(Y^{(k)}) &= \left(\frac{1}{R}\right)^{2k}\sum_{j=1}^{k}(R\rho)^{2k-j-1}\mathbf{1}'(\mathbf{M}_c')^{k-j} \\ &\quad\cdot\left(\sum_{i=1}^{r}\mathbf{V}_i(\mathbf{e}_1'\mathbf{M}_c^{j-1}\mathbf{e}_i)\right)(\mathbf{M}_c)^{k-j}\mathbf{1} \\ &= \rho^{2k}\frac{1}{R\rho}\sum_{j=1}^{k}\left(\frac{1}{R\rho}\right)^{j}\mathbf{1}'(\mathbf{M}_c')^{k-j} \\ &\quad\cdot\left(\sum_{i=1}^{r}\mathbf{V}_i(\mathbf{e}_1'\mathbf{M}_c^{j-1}\mathbf{e}_i)\right)(\mathbf{M}_c)^{k-j}\mathbf{1}. \qquad (16)\end{aligned}$$

Writing it in this form allows comparison with (15). From the steps used there, we find that for $R > 1/\rho$, we have $\mathrm{Var}(Y^{(k)}) = O(\rho^{2k})$.

Next we prove the lower bounds in (i) and (iii). Because the $\mathbf{V}_i$ matrices are positive definite and $\mathbf{M}_c \geq 0$, all the terms in the summation in (16) are nonnegative. Hence, for the lower bound in (iii) it suffices to consider the first term in the summation for $\mathrm{Var}(Y^{(k)})$; i.e.,

$$\mathrm{Var}(Y^{(k)}) \geq \rho^{2k}\left(\frac{1}{R\rho}\right)^2\mathbf{1}'(\mathbf{M}_c^{k-1})'\left(\sum_{i=1}^{r}\mathbf{V}_i\right)(\mathbf{M}_k^{k-1})\mathbf{1}.$$

Then, using Theorem 1 we have that

$$\mathrm{Var}(Y^{(k)}) \geq \rho^{2k}\left(\frac{1}{R\rho}\right)^2\left(\mathbf{1}'\tilde{\mathbf{M}}_c'\left(\sum_{i=1}^{r}\mathbf{V}_i\right)\tilde{\mathbf{M}}_c\mathbf{1} + O(\lambda^{k-1})\right),$$

where $\lambda < 1$. Now using the fact that the $\mathbf{V}_i$ is a positive definite matrix for all i and the fact that $\mathbf{M}_c > 0$ (recall that $\tilde{\mathbf{M}}_c = \mu'\nu$ where μ and ν are positive left and right eigenvectors of $\mathbf{M}_c$), we get the result we want. For (i) we can bound the $\mathrm{Var}(Y^{(k)})$ in (16) from below by the last term in the summation and use the fact that

$$\begin{aligned}\mathrm{Var}(Y^{(k)}) &\geq \rho^{2k}\left(\frac{1}{R\rho}\right)\left(\frac{1}{R\rho}\right)^k\sum_{i=1}^{r}\mathbf{1}'\mathbf{V}_i\mathbf{1}(\mathbf{1}'\mathbf{M}_c^{k-1}\mathbf{e}_i) \\ &= \rho^{2k}\left(\frac{1}{R\rho}\right)\left(\frac{1}{R\rho}\right)^k\sum_{i=1}^{r}\mathbf{1}'\mathbf{V}_i\mathbf{1}(\mathbf{1}'\tilde{\mathbf{M}}_c\mathbf{e}_i + O(\lambda^{k-1})).\end{aligned}$$

PROOF OF LEMMA 2. First consider the case when $N = 1$. For all $\delta > 0$ there exists $k_0 \equiv k_0(\delta)$ such that for all $k > k_0$, $(1 - \delta)\mathbf{P} \leq \mathbf{P}_k \leq (1 + \delta)\mathbf{P}$. Hence, for all $k \geq k_0$,

$$\begin{aligned}\gamma_k &\leq \mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\mathbf{P}^{k-k_0}\mathbf{1}(1+\delta)^{k-k_0} \\ &= \mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}(\rho^{k-k_0}\tilde{\mathbf{P}} + \hat{\mathbf{P}}^{k-k_0})\mathbf{1}(1+\delta)^{k-k_0} \\ &= (1+\delta)^{k-k_0}\rho^{k-k_0} \\ &\quad\cdot\left(\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\tilde{\mathbf{P}}\mathbf{1} + \frac{1}{\rho^{k-k_0}}\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\hat{\mathbf{P}}^{k-k_0}\mathbf{1}\right).\end{aligned}$$

The last equality follows from Theorem 1. Taking logarithms we have that

$$\begin{aligned}\frac{1}{k}\log(\gamma_k) &\leq \frac{(k-k_0)}{k}\log(1+\delta) + \frac{(k-k_0)}{k}\log(\rho) \\ &\quad + \frac{1}{k}\log(\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\tilde{\mathbf{P}}\mathbf{1} \\ &\quad + \frac{1}{\rho^{k-k_0}}\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\hat{\mathbf{P}}^{k-k_0}\mathbf{1}) \\ &\leq \log(1+\delta) + \log(\rho) - \frac{k_0}{k}\log(1+\delta) \\ &\quad - \frac{k_0}{k}\log(\rho) + \frac{1}{k}\log\Big(\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\tilde{\mathbf{P}}\mathbf{1} \\ &\quad + \frac{1}{\rho^{k-k_0}}\mathbf{e}_1'\mathbf{P}_1\cdots\mathbf{P}_{k_0}\hat{\mathbf{P}}^{k-k_0}\mathbf{1}\Big).\end{aligned}$$

Now for any $\epsilon > 0$, there exists an integer $k_1 > 0$ such that for $k \geq k_1$ the absolute value of each of the last three terms is less than $\epsilon/3$. Also, what we want is that $\log(1 + \delta) = \epsilon/3$ or $\delta = e^{\epsilon/3} - 1$. Then for all $k \geq k_2 \equiv \max(k_1, k_0)$ we will have that $\log(\gamma_k)/k \leq \log(\rho) + \epsilon$.

Similarly, we can show that for all $\epsilon > 0$, there will exist k_3 such that for all $k > k_3$, $\log(\gamma_k)/k \geq \log(\rho) - \epsilon$. Hence, we will have the result for $N = 1$.

Note that for the case of general N (recall that we assume $\mathbf{P}^N > 0$ for some $N > 0$), we will define $\mathbf{B}_k =$

$\mathbf{P}_{(k-1)N+1} \cdots \mathbf{P}_{kN}$. Then clearly $\mathbf{B}_k \to \mathbf{P}^N$ as $k \to \infty$. Hence, for every $\delta > 0$ there exists $k(\delta)$, such that for $k > k(\delta)$, $(1 - \delta)\mathbf{P}^N \leqslant \mathbf{B_k} \leqslant (1 + \delta)\mathbf{P}^N$. Then using the same method as above we can show that a subsequence of $\{\log(\gamma_k)/k; k \geqslant 1\}$, i.e., $\{\log(\gamma_{kN})/kN : k \geqslant 0\}$, converges to $\log(\rho)$. We can show the same thing for all the N nonoverlapping subsequences $\{\log(\gamma_{kN+j})/(kN + j) : k \geqslant 0\}$ for $j = 0, 1 \ldots N - 1$, which when put together form the complete sequence. Hence, the overall sequence converges to $\log(\rho)$. □

PROOF OF THEOREM 4. We will need the following lemma for proving the theorem.

LEMMA 3. *Consider a sequence* $\{a_i : i \geqslant 0\}$ *with* $a_i > 0$ *and define* $d_k = \sum_{i=0}^{k} a_i$. *If* $\log_{k\to\infty} \log(a_k)/k = 0$ *then* $\lim_{k\to\infty} \log(d_k)/k = 0$.

This simply states that if the terms of a summation have a subexponential growth rate then the summation has a subexponential growth rate in the number of terms.

PROOF OF LEMMA 3. For all $\delta > 0$, there exists $k_0 \equiv k_0(\delta)$ such that for $k > k_0$, $-\delta < \log(a_k)/k < \delta$ or $e^{-\delta k} \leqslant a_k \leqslant e^{\delta k}$. Now for all $k \geqslant k_0$,

$$d_k = \sum_{i=1}^{k_0} a_i + \sum_{i=k_0+1}^{k} a_i \leqslant \sum_{i=1}^{k_0} a_i + (k - k_0)e^{k\delta}.$$

Therefore,

$$\frac{1}{k}\log(d_k) \leqslant \frac{1}{k}\log\left(\frac{\sum_{i=1}^{k_0} a_i + (k - k_0)e^{k\delta}}{(k - k_0)e^{k\delta}}\right) + \frac{1}{k}\log(k - k_0) + \delta.$$

Note that the first two terms on the right side approach zero as $k \to \infty$. For any $\epsilon > 0$, let k_1 (k_2) be such that the absolute value of the first term (second term) is less than $\epsilon/3$. Also choose $\delta = \epsilon/3$. Then for $k \geqslant \max\{k_0, k_1, k_2\}$, $\log(d_k)/k \leqslant \epsilon$. We can similarly show that $\log(d_k)/k \geqslant -\epsilon$. □

PROOF OF THEOREM 4. The proof makes use of the asymptotic property of γ_k given in Lemma 2. Define $\gamma_0 = 1$. Using the fact that $\mathbf{M}_j = R_j\mathbf{P}_j$ we can express

$$w(k) = \sum_{i=0}^{k-1} \mathbf{e}_1'\left(\prod_{j=1}^{i} \mathbf{M}_j\right)\mathbf{1}R_{i+1} = \sum_{i=0}^{k-1} \mathbf{e}_1'\left(\prod_{j=1}^{i} \mathbf{P}_j\right)\mathbf{1}\prod_{j=1}^{i+1} R_j = \sum_{i=0}^{k-1} \gamma_i \prod_{j=1}^{i+1} R_j. \quad (17)$$

It is also easy to show that since $R_j \to R$,

$$\frac{1}{k}\log\left(\prod_{j=1}^{k} R_j\right) \to \log(R). \quad (18)$$

Hence, from Theorem 2 we get that

$$\lim_{k\to\infty} \frac{1}{k}\log\left(\gamma_k \prod_{j=1}^{k+1} R_j\right) = \log(\rho) + \log(R) = \log(R\rho). \quad (19)$$

Now for the case of $R > 1/\rho$, since all the terms in the summation for $w(k)$ are positive,

$$\liminf_{k\to\infty} \frac{1}{k}\log(w(k)) \geqslant \liminf_{k\to\infty} \frac{1}{k}\log\left(\gamma_{k-1} \prod_{j=1}^{k} R_j\right) = \log(R\rho) > 0.$$

The last equation follows from (19). For the case of $R = 1/\rho$, the proof follows straight from (19) and Lemma 3. For the case of $R < 1/\rho$, to show that $w(k) = O(1)$, we use the root test, i.e., we need to show that $\limsup_i(\gamma_i \prod_{j=1}^{i+1} R_j)^{1/i} < 1$ or $\limsup_i \log(\gamma_i \prod_{j=1}^{i+1} R_j)/i < 0$. This follows from (19). To show that $w(k) = \underline{O}(1)$ we can just use the fact that $w(k) \geqslant R_1$ from (17). □

PROOF OF THEOREM 5. As before, from the positive definiteness of the $\mathbf{V}_i^{(j)}$ matrices and the fact that $\mathbf{M} \geqslant 0$, we find that all the terms in the summation in (11) are nonnegative. To show (i) we will consider only the last term of (12), i.e.,

$$\mathrm{Var}(Y^{(k)}) \geqslant \frac{R_1 \cdots R_{k-1}}{(R_1 \cdots R_k)^2}\left(\sum_{i=1}^{r} \mathbf{1}'\mathbf{V}_i^{(k)}\mathbf{1}\right) \cdot (\mathbf{e}_1'\mathbf{P}_1 \cdots \mathbf{P}_{k-1}\mathbf{e}_i) \geqslant \rho^{2k}\frac{1}{\rho^{2k}(R_1 \cdots R_k)}\frac{1}{R_k}\left(\sum_{i=1}^{r} \mathbf{1}'\mathbf{V}_i^{(k)}\mathbf{1}\right) \cdot (\mathbf{e}_1'\mathbf{P}_1 \cdots \mathbf{P}_{k-1}\mathbf{e}_i) > 0. \quad (20)$$

Using exactly the same method as for Lemma 2, we can show that $\log(\mathbf{e}_1'\mathbf{P}_1 \cdots \mathbf{P}_{k-1}\mathbf{e}_i))/k \to \log(\rho)$ as $k \to \infty$. Taking logarithms of both sides of (20), dividing by k and using (18), we get that $\liminf_{k\to\infty}(\log(\mathrm{Var}(Y^{(k)}))/k) \geqslant 2\log(\rho) + \log(1/(R\rho)) > 2\log(\rho)$.

For the case where $R \geqslant 1/\rho$, we will only consider the case where $\mathbf{P} > 0$ and therefore $\mathbf{M} \equiv R\mathbf{P} > 0$. Then one can work along lines similar to Lemma 2 to extend it to the case where $\mathbf{P}^N > 0$ for some $N > 0$. Since $R_k \to R$ and $\mathbf{P}_k \to \mathbf{P}$ as $k \to \infty$, we have $\mathbf{M}_k \to \mathbf{M}$. Also, $\mathbf{V}_{abs}^{(k)} \to \mathbf{V}_{abs}$. Hence, for any $\delta > 0$, there exists j_0 such that for $j \geqslant j_0$, $(1 - \delta)\mathbf{M} \leqslant \mathbf{M}_j \leqslant (1 + \delta)\mathbf{M}$ and $\|\mathbf{V}_{abs}^{(j)}(l, m) - \mathbf{V}_{abs}(l, m)\| < \delta$ for all l and m.

First we will show that $\limsup_{k\to\infty} \log(\mathrm{Var}(Y^{(k)}))/k \leqslant 2\log(\rho)$ for $R \geqslant 1/\rho$. Note that for this case since $\mathbf{M}_c \equiv \mathbf{P}/\rho$, $\mathbf{M} \equiv (R\rho)\mathbf{M}_c$. For $k > j_0$, we will divide the summation in (11) into two parts; the first is summation from $j = 1$ to j_0 which we call Var1(k), the other from $j = j_0 + 1$ to k which we call Var2(k).

$$\begin{aligned}
\text{Var1}(k) &\equiv \sum_{j=1}^{j_0} \mathbf{1}'(\mathbf{M}_k' \cdots \mathbf{M}_{j+1}') \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_i^{(j)} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j-1} \mathbf{e}_i) \right) (\mathbf{M}_{j+1} \cdots \mathbf{M}_k) \mathbf{1} \\
&\leqslant \sum_{j=1}^{j_0} (1+\delta)^{2k-2j_0} \mathbf{1}'(\mathbf{M}')^{k-j_0} (\mathbf{M}_{j_0}' \cdots \mathbf{M}_{j+1}') \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_{abs}^{(j)} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j-1} \mathbf{e}_i) \right) (\mathbf{M}_{j+1} \cdots \mathbf{M}_{j_0}) (\mathbf{M})^{k-j_0} \mathbf{1}' \\
&\leqslant (1+\delta)^{2k-2j_0} \sum_{j=1}^{j_0} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j-1} \mathbf{1}) \\
&\quad \cdot \| (\mathbf{M}')^{k-j_0} (\mathbf{M}_{j_0}' \cdots \mathbf{M}_{j+1}' \mathbf{V}_{abs}^{(j)} \mathbf{M}_{j+1} \cdots \mathbf{M}_{j_0}) (\mathbf{M})^{k-j_0} \| \\
&\leqslant (1+\delta)^{2k-2j_0} \| (\mathbf{M}_c)^{k-j_0} \|^2 (R\rho)^{2k-2j_0} \\
&\quad \cdot \sum_{j=1}^{j_0} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j-1} \mathbf{1}) \| \mathbf{M}_{j_0}' \cdots \mathbf{M}_{j+1}' \\
&\quad \cdot \mathbf{V}_{abs}^{(j)} \mathbf{M}_{j+1} \cdots \mathbf{M}_{j_0} \| \\
&\leqslant (1+\delta)^{2k} c_1 (R\rho)^{2k},
\end{aligned}$$

where c_1 is a positive constant (independent of k, possibly depending on j_0). The last inequality follows because $\|\mathbf{M}_c^{k-j_0}\| = \mathbf{1}'(\tilde{M}_c + \hat{M}_c^{k-j_0})\mathbf{1} \leqslant \mathbf{1}'\tilde{M}_c\mathbf{1} + \|\hat{M}_c^{k-j_0}\| = \mathbf{1}'\tilde{M}_c\mathbf{1} + O(\lambda^{k-j_0})$ where $\lambda < 1$.

$$\begin{aligned}
\text{Var2}(k) &\equiv \sum_{j=j_0+1}^{k} \mathbf{1}'(\mathbf{M}_k' \cdots \mathbf{M}_{j+1}') \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_i^{(j)} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j_0} \cdots \mathbf{M}_{j-1} \mathbf{e}_i) \right) (\mathbf{M}_{j+1} \cdots \mathbf{M}_k) \mathbf{1} \\
&\leqslant \sum_{j=j_0+1}^{k} (1+\delta)^{2k-2j} \mathbf{1}'(\mathbf{M}')^{k-j} \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_{abs}^{(j)} (\mathbf{e}_1' \mathbf{M}_1 \cdots \mathbf{M}_{j-1} \mathbf{e}_i) \right) (\mathbf{M})^{k-j} \mathbf{1}' \\
&\leqslant \| \mathbf{M}_1 \cdots \mathbf{M}_{j_0} \| \sum_{j=j_0+1}^{k} (1+\delta)^{2k-2j} \\
&\quad \cdot \| \mathbf{M}_{j_0+1} \cdots \mathbf{M}_{j-1} \| \\
&\quad \cdot \| (\mathbf{M}')^{k-j} (\mathbf{V}_{abs} + \mathbf{1}\mathbf{1}'\delta) (\mathbf{M})^{k-j} \| \\
&\leqslant \| \mathbf{V}_{abs} + \mathbf{1}\mathbf{1}'\delta \| \cdot \| \mathbf{M}_1 \cdots \mathbf{M}_{j_0} \| \sum_{j=j_0+1}^{k} (1+\delta)^{2k-2j} \\
&\quad \cdot (1+\delta)^{j-1-j_0} \| (\mathbf{M})^{k-j} \|^2 \| \mathbf{M}^{j-1-j_0} \| \\
&\leqslant \| \mathbf{V}_{abs} + \mathbf{1}\mathbf{1}'\delta \| \cdot \| \mathbf{M}_1 \cdots \mathbf{M}_{j_0} \| \sum_{j=j_0+1}^{k} (1+\delta)^{2k-2j} \\
&\quad \cdot (1+\delta)^{j-1-j_0} \| (\mathbf{M}_c)^{k-j} \|^2 \| \mathbf{M}_c^{j-1-j_0} \| (R\rho)^{2k-j-1-j_0} \\
&\leqslant \sum_{j=j_0+1}^{k} (1+\delta)^{2k-2j} (1+\delta)^{j} c_2 (R\rho)^{2k-j-1-j_0} \\
&\leqslant \sum_{j=j_0+1}^{k} (1+\delta)^{2k} c_2 (R\rho)^{2k} \\
&\leqslant k(1+\delta)^{2k} c_2 (R\rho)^{2k},
\end{aligned}$$

where c_2 is a positive constant. Hence we have that for $k > j_0$,

$$\begin{aligned}
\frac{1}{k} \log(\text{Var}(Y^{(k)})) &\leqslant -2 \frac{1}{k} \log(R_1 \cdots R_k) \\
&\quad + \frac{1}{k} \log(c_1 + kc_2) + 2 \log(1+\delta) + 2 \log(R\rho). \qquad (21)
\end{aligned}$$

Note that as $k \to \infty$, the first term on the right converges to $-2\log(R)$ and the second term converges to 0. So for any $\epsilon > 0$, we can choose k_1 and k_2, such that the first term is less than $2 \log(\rho) + \epsilon/3$ for all $k \geqslant k_1$, and the second term is less than $\epsilon/3$ for all $k \geqslant k_2$. Choose a δ such that the third term is equal to $\epsilon/3$. Then for all $\epsilon > 0$, there exists $k_{\max} = \max(k_1, k_2, j_0)$ such that for all $k \geqslant k_{\max}$, $\log(\text{Var}(Y^{(k)}))/k \leqslant 2 \log(\rho) + \epsilon$. Thus the lim sup results of the theorem, for $R \geqslant 1/\rho$, follow.

Finally, we will show that $\liminf_{k\to\infty} \log(\text{Var}(Y^{(k)}))/k \geqslant 2 \log(\rho)$ for $R > 1/\rho$. We will just need to look at the first term in the summation of (11). Hence, for $k > j_0$,

$$\begin{aligned}
&\text{Var}(Y^{(k)}) \\
&\geqslant \frac{1}{(R_1 R_2 \cdots R_k)^2} \mathbf{1}' \mathbf{M}_k' \cdots \mathbf{M}_2' \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_k \mathbf{1} \\
&\geqslant \frac{(1-\delta)^{2(k-j_0)}}{(R_1 R_2 \cdots R_k)^2} \mathbf{1}' \mathbf{M}'^{k-j_0} \mathbf{M}_{j_0}' \cdots \mathbf{M}_2' \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_{j_0} \mathbf{M}^{k-j_0} \mathbf{1} \\
&\geqslant \frac{(1-\delta)^{2(k-j_0)}}{(R_1 R_2 \cdots R_k)^2} (R\rho)^{2k-2j_0} \mathbf{1}' \mathbf{M}_c'^{k-j_0} \mathbf{M}_{j_0}' \cdots \mathbf{M}_2' \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_{j_0} \mathbf{M}_c^{k-j_0} \mathbf{1} \\
&\geqslant \frac{(1-\delta)^{2(k-j_0)}}{(R_1 R_2 \cdots R_k)^2} (R\rho)^{2k-2j_0} \mathbf{1} (\tilde{\mathbf{M}}_c' + \hat{M}_c'^{k-j_0}) \mathbf{M}_{j_0}' \cdots \\
&\quad \cdot \mathbf{M}_2' \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_{j_0} (\tilde{\mathbf{M}}_c + \hat{\mathbf{M}}_c^{k-j_0}) \mathbf{1} \\
&\geqslant \frac{(1-\delta)^{2(k-j_0)}}{(R_1 R_2 \cdots R_k)^2} (R\rho)^{2k-2j_0} (\mathbf{1}' \tilde{\mathbf{M}}_c' \mathbf{M}_{j_0}' \cdots \mathbf{M}_2' \\
&\quad \cdot \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_{j_0} \tilde{\mathbf{M}}_c \mathbf{1} + O(\lambda^{k-j_0}),
\end{aligned}$$

where $\lambda < 1$. Note that due to the fact that $\tilde{\mathbf{M}}_c > 0$, $\mathbf{M}_i > 0$ for all i, and the positive definiteness of $\mathbf{V}_i^{(1)}$ for at least one i, we have that

$$1' \tilde{\mathbf{M}}_c' \mathbf{M}_{j_0}' \cdots \mathbf{M}_2' \left(\sum_{i=1}^{r} \mathbf{V}_i^{(1)} \right) \mathbf{M}_2 \cdots \mathbf{M}_{j_0} \tilde{\mathbf{M}}_c \mathbf{1} > 0.$$

Taking logarithms of both the sides and using similar techniques as in the case where $R = 1/\rho$, we get (iii). □

ACKNOWLEDGMENTS

This work is supported by NSF National Young Investigator Award grant DMI-94-57189, NSF University-Industry Cooperative Research Program grant DMS-95-08709 and NSF Career Award grant DMI-96-25297. The work of T.

Zajic was performed while he held a joint appointment at IBM Research and Columbia University. The authors are grateful to Herbert Rief for referring them to the work of Burn, Dubi, and Goldfeld, and to a referee for referring them to the work of Melas and Ermakov.

REFERENCES

Anantharam, V., P. Heidelberger, P. Tsoucas. 1990. Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. IBM Research Report RC 16280. Yorktown Heights, NY.

Asmussen, S., R. Rubinstein. 1995. Steady-state rare events simulation in queueing models and its complexity properties. J. Dhashalow, ed. *Advances in Queueing*. CRC Press, Boca Raton, FL.

Bayes, A. J. 1970. Statistical techniques for simulation models. *Australian Comput. J.* **2** 180–184.

Burn, K. 1990. Optimizing cell importances using an extension of the DSA—Theory, implementation, preliminary results. *Progress in Nuclear Energy* **24** 39–54.

Chung, K. L. 1967. *Markov Chains with Stationary Transition Probabilities*, Second Edition. Springer-Verlag, New York.

Dubi, A. 1985. General statistical model for geometrical splitting in Monte Carlo—I. *Transport Theory and Statist. Physics* **14** 167–193.

——, A. Goldfeld, K. Burn. 1986. Application of the direct statistical approach on a multisurface splitting problem in Monte Carlo calculations. *Nuclear Sci. and Engrg.* **93** 204–213.

Ermakov, S. M., V. B. Melas. 1995. *Design and Analysis of Simulation Experiments*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Fox, B. L. 1997. Derandomizing splitting and Russian roulette. Working Paper. Sim-Opt Consulting, Boulder, CO.

Frater, M. R., T. M. Lenon, B. D. O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Trans. Automatic Control* **36** 1395–1405.

Glasserman, P., P. Heidelberger, P. Shahabuddin, T. Zajic. 1996. Splitting for rare event simulation: analysis of simple cases. *Proc. 1996 Winter Simulation Conf*., IEEE, Piscataway, NJ.

——, ——, ——, ——. 1998. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans. Automat. Control,* **43** 1666–1679.

——, ——, ——, ——. 1998a. A look at multilevel splitting. H. Niederreiter, P. Hellekalek, G. Larcher, P. Zinterhof, eds. *Monte Carlo and Quasi Monte Carlo Methods 1996*. Springer-Verlag, New York. 98–108.

——, S. G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling and Computer Simulation* **5** 22–42.

Glynn, P. W., W. Whitt. 1992. The asymptotic efficiency of simulation estimators. *Oper. Res.* **40** 505–520.

Hammersley, J., D. Handscomb. 1964. *Monte Carlo Methods*. Methuen & Co., Ltd., London, UK.

Harris, T. 1963. *The Theory of Branching Processes*. Springer-Verlag, New York.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling and Computer Simulation* **5** 43–85.

Hopmans, A. C. M., J. P. C. Kleijnen. 1979. Importance sampling in systems simulation: a practical failure? *Math. and Computers in Simulation* **21** 209–220.

Horn, R. A., C. R. Johnson. 1985. *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Kahn, H., T. E. Harris. 1951. Estimation of particle transmission by random sampling. *National Bureau of Standards Appl. Math. Series* **12** 27–30.

Kelling, C. 1996. A framework for rare event simulation of stochastic Petri nets using "RESTART." *Proc. 1996 Winter Simulation Conf*., IEEE, Piscataway, NJ.

Melas, V. B. 1993. Optimal simulation design by branching technique. W. G. Muller, H. P. Wynn, A. A. Zhigljavsky, eds. *Model Oriented Data Analysis*. Physica-Verlag, Heidelberg, Germany, 113–128.

——. 1994. Branching technique for Markov chain simulation (finite state case). *Statistics* **25** 159–171.

Parekh, S., J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Control* **34** 54–56.

Schreiber, F., C. Görg. 1994. Rare event simulation: a modified RESTART-method using LRE-Algorithm. J. Labetoulle, J. W. Roberts, eds. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*. Elsevier Science Publishers, Amsterdam, 787–796.

Schwetman, H. 1986. CSIM: a C-based, process oriented simulation language. *Proc. 1986 Winter Simulation Conf*., IEEE, Piscataway, NJ.

Shahabuddin, P. 1995. Rare event simulation in stochastic models. *Proc. 1995 Winter Simulation Conf*., IEEE, Piscataway, NJ.

Villén-Altamirano, M., A. Martínez-Marrón, J. Gamo, F. Fernández-Cuesta. 1994. Enhancements of the accelerated simulation method RESTART by considering multiple thresholds. J. Labetoulle, J. W. Roberts, eds. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*. Elsevier Science Publishers, Amsterdam, 797–810.

——, J. Villén-Altamirano. 1991. RESTART: a method for accelerating rare event simulations. J. W. Cohen, C. D. Pack, eds. *Queueing, Performance and Control in ATM*. Elsevier Science Publishers, Amsterdam, 71–76.

——, ——. 1994. RESTART: a straight-forward method for fast simulation of rare events. *Proc. Winter Simulation Conf*., Society for Computer Simulation, San Diego, CA.