RC 20521 (08/07/96)
Computer Science

IBM Research Report

Splitting For Rare Event Simulation: Analysis of Simple Cases

Paul Glasserman

Columbia Business School
New York, New York 10027

Philip Heidelberger

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York 10598

Perwez Shahabuddin, Tim Zajic

IEOR Department
Columbia University
New York, New York 10027

Research Division
Almaden · T.J. Watson · Tokyo · Zurich

Splitting For Rare Event Simulation: Analysis of Simple Cases *

Paul Glasserman
Columbia University

Philip Heidelberger
IBM Research

Perwez Shahabuddin†
Columbia University

Tim Zajic
Columbia University and IBM Research‡

July 1996

ABSTRACT

An approach to rare event simulation uses the technique of splitting. The basic idea is to split sample paths of the stochastic process into multiple copies when they approach closer to the rare set; this increases the overall number of hits to the rare set for a given amount of simulation time. This paper analyzes the bias and efficiency of some simple cases of this method.

1 INTRODUCTION

Estimations of the small probabilities of rare events are required in the design and operation of many engineering systems. Consider the case of a telecommunications network. It is customary to model such systems as a network of queues, with each queue having a buffer of finite capacity. Information packets that arrive to a queue when its buffer is full are lost. The rare event of interest may be the event of a packet being lost. Current standards stipulate that the probability of packet loss should not exceed 10^{-9}. Or in a reliability model of a space craft computer, we may be interested in estimating the probability of the event that the system fails before the mission completion. Naturally, one would want this to be extremely low. The main problem with using standard simulation to estimate such small probabilities is that a large number of events have to be simulated in the model before any samples of the rare event may occur. Hence special simulation techniques are needed to make the events of interest occur more frequently.

Importance sampling is a technique that has been widely used for this purpose. The reader is referred to Heidelberger 1995 and Shahabuddin 1995 for some surveys. In importance sampling, the stochastic model is simulated with a new probability dynamics (called a change of measure), that makes the events of interest occur more frequently. The sample value is then adjusted to make the final estimate unbiased. However, choosing any change of measure that makes the event of interest occur frequently is not enough; *how* it is made to happen more frequently is also very important. For example, an arbitrary change of measure that makes the rare event happen more frequently may give an estimator with an infinite variance. Thus the main problem in importance sampling is

*This paper will appear in the *Proceedings of the 1996 Winter Simulation Conference.*

†Currently on a leave of absence from IBM Research

‡Author Addresses. Glasserman: 403 Uris Hall, Columbia Business School, New York, NY 10027. Heidelberger: IBM T.J. Watson Research Center, PO Box 704, Yorktown Heights, NY 10598. Shahabuddin and Zajic: IEOR Department, Columbia University, New York, NY 10027.

to come up with an appropriate change of measure for the rare event simulation problem in hand. This may be difficult or almost impossible for complicated models. Hence, even though importance sampling works very well for a large class of stochastic models, the scope of application of importance sampling is limited to systems with "nice" structure.

This paper deals with an alternate approach to rare event simulation that uses the simulation technique of splitting (see, e.g., Hammersley and Handscomb 1965). In standard simulation, the stochastic process being simulated, spends a lot of time in a region of the state space which is "far away" from the rare set of interest, i.e, from where the chance of it entering the rare set is extremely low. In splitting a region of the state space that is "closer" to the rare set is defined. Each time the process reaches this region, from the "far away" region, many identical copies of this process are generated. This way we get more instances of the stochastic process spending time in a region where the rare event is more likely to occur. The boundary between the far away region and the closer region is called a threshold. The above described a case with one-threshold; one can easily extend it to the case where we have multiple thresholds. This approach to rare event simulation was introduced in Kahn and Harris (1951) and used later in Bayes (1970) and Hopmans and Kleijnen(1979). Recently it was revisited in a significant way by Villén-Altamirano and Villén-Altamirano (1991), Villén-Altamirano et al. (1994) and Villén-Altamirano and Villén-Altamirano (1994), who used it for estimating the probability of rare events in computer and communication systems. They called their version of this approach RESTART. A software package called ASTRO (Villén-Altamirano and Villén-Altamirano 1994) was created that implements their method. They also did some approximate efficiency analysis that gave some insights into threshold selection and number of split paths generated at each threshold. But a formal and thorough analysis was lacking.

Glasserman, Heidelberger, Shahabuddin, Zajic (1996) (henceforth referred to as GHSZ) describe a unifying 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. The theory of branching processes (see, e.g., Harris 1989) was used to derive the unbiasedness and efficiency results. Experimental results supporting the theoretical analysis and exploring the robustness of the splitting method, are also reported in GHSZ. In this paper we introduce and derive some biasedness and efficiency results that supplement those in that paper. We begin with a simple setting, and give conditions under which the splitting method is optimal. We then give reasons why deviations from this simple setting result in difficulties. Some of these have been handled in GHSZ, whereas others are currently being investigated. Some analytical results on the optimal selection of thresholds are introduced next. Finally we give an analysis of the bias introduced in one implementation of this method that truncates sample paths to save simulation effort.

2 A SIMPLE SETTING

Consider the problem of estimating $\gamma = P(A)$, thinking of A as a rare event. Let $A = A_k \supset A_{k-1} \cdots \supset A_1$ be a nested sequence of events which we think of as intermediate thresholds. Let $p_1 = P(A_1)$ and $p_{i+1} = P(A_{i+1}|A_i)$, $i = 1, \ldots, k-1$; then

$$\gamma \equiv \gamma_k = p_1 p_2 \cdots p_k.$$

We think of k increasing to infinity and $\gamma \to 0$ (this would happen, for example, if $p_i = p$ for all i, where p is some fixed constant between 0 and 1).

To motivate the above setting, consider a single server queueing system with a finite buffer B. Define the state of the system to be the number of jobs in the queue. The problem may be to estimate the probability that starting from state 0, the system reaches state B before visiting 0. We can think of this event as the event A. Estimating probabilities of this type are crucial to the simulation based estimation of performance measures like the steady state probability of packet loss (see, e.g., Heidelberger 1995). Clearly, if the overall arrival rate is smaller than the overall service rate (which is a requirement for the stability of the queue), and B is large, then the event A is a rare event. Suppose now that we place $k-1$ intermediate thresholds between 0 and B (with B being the kth threshold). Let A_i be the event that starting from state 0, the number of jobs in the system reaches threshold i before reaching 0. Then clearly $A_{i+1} \subset A_i$ and we have an example of the setting mentioned in the previous paragraph.

Suppose that for each i we can generate n_i Bernoulli random variables with parameter p_i, all independent of each other. These are the building blocks of a splitting estimator in this simple setting. From each successful Bernoulli outcome at stage i, we generate n_{i+1} stage-$(i+1)$ Bernoullis. Thus, at the first stage we have Bernoullis

$$\mathbf{1}_1, \mathbf{1}_2, \ldots, \mathbf{1}_{n_1};$$

the jth of these, if successful, spawns

$$\mathbf{1}_{j1}, \mathbf{1}_{j2}, \ldots, \mathbf{1}_{jn_2},$$

and so on. The estimator is

$$I_k = \frac{1}{n_1 \cdots n_k} \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} \mathbf{1}_{i_1} \cdots \mathbf{1}_{i_1 \cdots i_k}.$$

It is easy to show that I_k is an unbiased estimator. By conditioning on $\mathcal{F}_k$ we mean conditioning on the outcomes of all Bernoullis up to stage k. Then

$$\begin{aligned}
E(I_k) &= E(E(I_k|\mathcal{F}_{k-1})) \\
&= E\left(\frac{1}{n_1 \cdots n_k} \sum_{i_1=1}^{n_1} \cdots \sum_{i_{k-1}=1}^{n_{k-1}} \mathbf{1}_{i_1} \cdots \mathbf{1}_{i_1 \ldots i_{k-1}} \right. \\
&\quad \left. \sum_{i_k=1}^{n_k} E(\mathbf{1}_{i_1 \cdots i_k}|\mathcal{F}_{k-1}).\right) \\
&= E(I_{k-1}\frac{1}{n_k} n_k p_k) = p_k E(I_{k-1}).
\end{aligned}$$

Doing this iteratively we get that $E(I_k) = p_1 \ldots p_k = \gamma_k$.

Returning to the simple queueing example introduced above, a Bernoulli random variable might be the indicator that a simulated process reaches the next threshold from the current one, without visiting state 0. In particular, the p_i should be considered unknown, so the n_{i+1} Bernoullis from each of the successful outcomes at stage i would typically be generated implicitly by simulating n_{i+1} samples of the underlying process starting from stage i, until it either hits threshold $i+1$, or it hits 0 (see Figure 1). If a sample path hits threshold $i+1$ before hitting 0, then the corresponding Bernoulli random variable is set to 1; else it is set to zero. Of course, since the Bernoullis are all independent, the queuing process must satisfy the assumption that the dynamics of the process after it hits the ith threshold, is independent of the past and depends only on i. A simple example where this is true is the M/M/1/N queue.

We now calculate the variance of this estimator:

$$\begin{aligned}
\sigma_{k+1}^2 &\equiv \mathrm{Var}[I_{k+1}] \\
&= \mathrm{Var}[E[I_{k+1}|\mathcal{F}_k]] + E[Var[I_{k+1}|\mathcal{F}_k]] \\
&= \frac{1}{n_{k+1}^2}\{\mathrm{Var}[I_k n_{k+1} p_{k+1}] \\
&\quad + E[I_k n_{k+1} p_{k+1}(1-p_{k+1})]\} \\
&= p_{k+1}^2 \sigma_k^2 + \frac{p_1 \cdots p_k p_{k+1}(1-p_{k+1})}{n_1 \cdots n_{k+1}}.
\end{aligned}$$

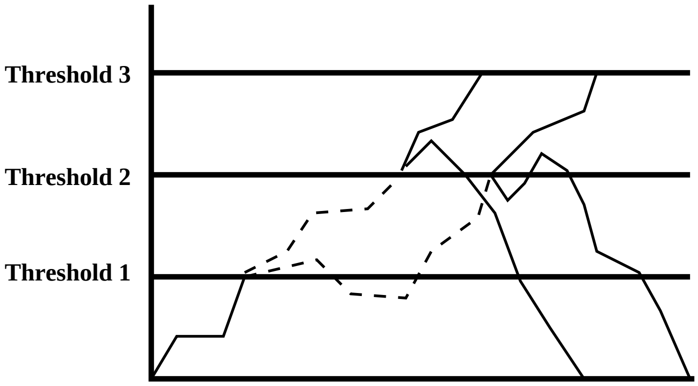

Figure 1: Splitting with three thresholds and two split subpaths at each intermediate threshold.

This recurrence relation can be solved to get

$$\sigma_k^2 = \sum_{j=1}^{k} \left(\prod_{i=j+1}^{k} p_i^2 \right) \frac{p_1 \cdots p_j (1-p_j)}{n_1 \cdots n_j},$$

which can also be written as

$$\sigma_k^2 = (p_1 \cdots p_k)^2 \sum_{j=1}^{k} \frac{1-p_j}{(p_1 n_1) \cdots (p_j n_j)}. \tag{1}$$

The next result examines the behavior of this variance as k increases.

Lemma 1 *(i) If* $\liminf_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \log(n_i p_i) > 0$,

$$\sigma_k^2 = O((p_1 \cdots p_k)^2); \tag{2}$$

(ii) if $-\infty < \liminf_{j\to\infty} \sum_{i=1}^{j} \log(n_i p_i)$,

$$\sigma_k^2 = O(k(p_1 \cdots p_k)^2);$$

(iii) if $\limsup_{k\to\infty} p_k < 1$ *and* $\liminf_{j\to\infty} \frac{1}{j} \sum_{i=1}^{j} \log(n_i p_i) < 0$,

$$(p_1 \cdots p_k)^2 = o(\sigma_k^2).$$

Proof. The three cases are related to the convergence of the sum on the right side of (1). By the root test, the series

$$\sum_{j=1}^{\infty} \frac{1-p_j}{(p_1 n_1) \cdots (p_j n_j)} \tag{3}$$

converges if

$$\limsup_{j\to\infty} \left[\frac{1-p_j}{(p_1 n_1) \cdots (p_j n_j)} \right]^{1/j} < 1; \tag{4}$$

i.e., if

$$\limsup_{j\to\infty} \frac{1}{j} \left[\log(1-p_j) + \sum_{i=1}^{j} \log\left(\frac{1}{n_i p_i}\right) \right] < 0.$$

The condition in (i) is equivalent to

$$\limsup_{j\to\infty} \frac{1}{j}\sum_{i=1}^{j} \log(\frac{1}{n_i p_i}) < 0 \tag{5}$$

which ensures convergence of (3) and proves (2). The reverse inequality in (iii) similarly ensures divergence of (3). For (ii), notice that

$$\sum_{j=1}^{k} \frac{1-p_j}{(p_1 n_1)\cdots(p_j n_j)} = O(k)$$

if

$$\limsup_{j\to\infty}(1-p_j)\prod_{i=1}^{j}\left(\frac{1}{p_i n_i}\right) < \infty,$$

which holds under the condition in (ii). □

The conditions in this lemma simplify in the important special case that the p_i approach some limit p and all n_i equal some n for sufficiently large i. In this setting, the three cases in the lemma can be replaced with $np > 1$, $np = 1$, and $np < 1$. The corresponding results for this special cases may be found in GHSZ.

In case (i) of the lemma, the second moment of the estimator is also $O((p_1\cdots p_k)^2)$, because the first moment is $p_1\cdots p_k$. Nonnegativity of variance makes this the best possible rate of decrease for the second moment. In contrast, straightforward simulation (corresponding to a single Bernoulli with parameter $p_1\cdots p_k$) has variance

$$(p_1\cdots p_k)[1-(p_1\cdots p_k)] = O(p_1\cdots p_k)$$

per replication and a second moment of the same order.

We now supplement results for the variance with an assessment of the computational effort. We assume, for simplicity, that the work per sample is constant across stages. (In many cases this may not be true. For example, in many queueing models the expected cost of simulating a trial from threshold i grows proportionally with i. This is because, with positive probability bounded away from zero, the system never reaches threshold $i+1$ and therefore many trials consist of simulating the queue until it empties again. However, these other cases can be handled similarly and lead to similar conclusions.) Then the expected work is proportional to the expected number of samples, which is

$$\begin{aligned} n_1 + (n_1 p_1) n_2 + \cdots + (n_1 p_1 \cdots n_{k-1} p_{k-1}) n_k \\ = n_1 \textstyle\sum_{j=0}^{k-1} \prod_{i=1}^{j} p_i n_{i+1}. \end{aligned} \tag{6}$$

For the expected work we have:

Lemma 2 *(i) If* $\limsup_{j\to\infty} \frac{1}{j}\sum_{i=1}^{j}\log(p_i n_{i+1}) > 0$, *the expected work per run grows exponentially in* k.

(ii) If $\sum_{i=1}^{\infty}\log(p_i n_{i+1}) < \infty$, *then the expected work per run is* $O(k)$.

(iii) If $\limsup_{j\to\infty} \frac{1}{j}\sum_{i=1}^{j}\log(p_i n_{i+1}) < 0$, *then the expected work per run is* $O(1)$.

Proof. For case (i), note that

$$\begin{aligned}\frac{1}{k}\log\left(\sum_{j=0}^{k-1}\prod_{i=1}^{j}p_in_{i+1}\right) &\geq \frac{1}{k}\log\left(\prod_{i=1}^{k-1}p_in_{i+1}\right)\\ &= \frac{1}{k}\sum_{i=1}^{k-1}\log(p_in_{i+1})\end{aligned}$$

so a positive limsup for this expression indicates exponential growth of expected work. The expected work is $O(k)$ if

$$\frac{1}{k}\sum_{j=0}^{k-1}\prod_{i=1}^{j}p_in_{i+1}$$

converges, and a sufficient condition for this is the condition in case (ii) above. The condition in case (iii) above ensures that the series

$$\sum_{j=0}^{\infty}\prod_{i=1}^{j}p_in_{i+1}$$

converges, by the root test. □

As in Lemma 1, the conditions here can be replaced with $np >, =,$ or < 1 in the case of $p_i \to p$ and fixed $n_i = n$. The corresponding results for these special cases may also be found in GHSZ.

The work-normalized variance, balancing computational effort and estimator variance, is the product of the variance and the expected work per run; see Glynn and Whitt (1989) for full justification of this criterion. Combining Lemmas 1 and 2 yields a condition for optimal splitting:

Theorem 1 *If*

$$-\infty < \liminf \sum_{i=1}^{j}\log(p_in_i)$$

and

$$\sum_{i=1}^{\infty}\log(p_in_{i+1}) < \infty,$$

then I_k is asymptotically efficient in the sense that

$$\lim_{k\to\infty}\frac{\log O(k^2(p_1\cdots p_k)^2)}{\log E[I_k]} = 2.$$

We interpret this result to mean that splitting is most effective when $n_i \approx 1/p_i$. GHSZ discuss the use of a random number of splits in order to get the expected number of subpaths equal to $1/p_i$ when p_i is not the reciprocal of an integer.

The analysis above is based on a very simple model of splitting in which the success probabilities p_i are constant at each threshold, regardless of what may have happened at previous thresholds. Consider estimating the probability that a Markov chain reaches some rare set before returning to its initial state. We label the initial state 0 and assume it is recurrent. Imagine introducing intermediate thresholds in the state space of the Markov chain and splitting each path that reaches a threshold into some number of subpaths. In general, the probability that the chain will reach the ith threshold before 0, given that it has reached the $(i-1)$th threshold before 0, will depend on the state of the chain when it reached the $(i-1)$th threshold. The assumption of constant p_i would hold if, say, there were just one state through which the $(i-1)$th threshold could be reached; but, more typically, p_i would vary depending on the entry state.

The case where we have a fixed and finite number of entry states into each threshold, and the probability dynamics of the process is homogeneous (in some limiting sense) with respect to the thresholds, is further analyzed in GHSZ. To get a sense of the possible impact of the variability of the p_i's in a more general setting (i.e., uncountably infinite number of entry states into thresholds), we consider a simple two-threshold problem. Our objective is to estimate $\gamma = p_1 p_2$ where, now, $p_2 = E[\tilde{p}_2]$, with $\tilde{p}_2$ stochastic. The mechanism we have in mind is this: each path has probability p_1 of reaching the first threshold; upon reaching that threshold, its offspring are randomly assigned a sample of $\tilde{p}_2$ as their common second-stage success probability. This accurately represents the Markovian setting described above. The estimator is

$$I_2 = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \mathbf{1}_{i_1} \mathbf{1}_{i_1 i_2},$$

with each $\mathbf{1}_{i_1} \sim \text{Bernoulli}(p_1)$ and each $\sum_{i_2=1}^{n_2} \mathbf{1}_{i_1 i_2}$ conditionally Binomial$(n_2, \tilde{p}_2)$, independent for different i_1.

Let $N_2 = \sum_{i_2=1}^{n_2} \mathbf{1}_{i_1 i_2}$ for some i_1. Then

$$\text{Var}[N_2] = n_2 p_2 (1 - p_2) + (n_2^2 - n_2)\text{Var}[\tilde{p}_2].$$

Proceeding as we did for (1), we get

$$\text{Var}[I_2] = \frac{1}{n_1^2 n_2^2} \left\{ n_2^2 p_2^2 n_1 p_1 (1 - p_1) + n_1 p_1 \text{Var}[N_2] \right\}.$$

This becomes

$$\frac{p_2^2 p_1 (1 - p_1)}{n_1} + \frac{p_1 p_2 (1 - p_2)}{n_1 n_2} + \frac{p_1 (n_2^2 - n_2)\text{Var}[\tilde{p}_2]}{n_1 n_2^2}. \tag{7}$$

The last term gives the effect of a random $\tilde{p}_2$ compared with a fixed p_2. To get a sense of its impact, divide through by $\gamma^2 = p_1^2 p_2^2$, and suppose that $n_i \approx 1/p_i$, $i = 1, 2$. The contribution of the first two terms is then $O(1)$ whereas the new variability term contributes $O(n_2^2 Var[\tilde{p}_2]) = O(Var[\tilde{p}_2]/p_2^2)$.

This simple observation has important implications for the effectiveness of multithreshold splitting procedures: splitting will be most effective if there is little variability in the success probability at each threshold. This further suggests (at least heuristically) that the thresholds should be chosen in a way that is consistent with the most likely path to a rare set. For then each subpath will draw a success probability close to that for the most likely path, resulting in little variation across subpaths. Understanding the large deviations behavior of a rare event may therefore be useful in designing a splitting procedure.

3 OPTIMAL PARTITIONS

We now return to the simple setting from the start of Section 2. In particular, the p_i are constant at each threshold and we want to estimate $\gamma \equiv \gamma_k = p_1 \cdots p_k$, with $p_1 = P(A_1)$ and $p_i = p(A_i | A_{i-1})$, continuing to think of $k \to \infty$ and $\gamma_k \to 0$. We consider the problem of choosing the intermediate events $A_0, A_1, \ldots, A_{k-1}$ and make two observations: choosing these events so that the p_i converge as $i \to \infty$ has an asymptotic optimality property, and there is a connection between being able to choose the thresholds so that the p_i converge and being able to analyze the large deviations behavior of a rare event.

We begin by examining the optimal choice of $p_1, \ldots, p_k$ for fixed γ. Based on the analysis in Section 2, we restrict attention to the case $n_i = 1/p_i$, ignoring the integrality constraint on the n_i. In this case, the variance becomes

$$\sigma_k^2 = \gamma^2 \sum_{j=1}^{k} (1 - p_j),$$

and the expected work per run becomes

$$\sum_{j=1}^{k} \frac{1}{p_j}.$$

Our objective is then to minimize

$$g(p_1, \ldots, p_k) = \sum_{i,j}^{k} \frac{1-p_i}{p_j}$$

subject to $p_1 \cdots p_k = \gamma$. Rewriting the constraint and appending it with a Lagrange multiplier yields

$$\sum_{i,j}^{k} \frac{1-p_i}{p_j} + \lambda \left(\sum_{j=1}^{k} \log p_j - \log \gamma \right).$$

The first-order conditions

$$-\frac{1}{p_i^2} \sum_{j=1}^{k} (1-p_j) - \sum_{j=1}^{k} \frac{1}{p_j} + \frac{\lambda}{p_i} = 0, \quad i = 1, \ldots, k,$$

and $\sum_{j=1}^{k} \log p_j - \log \gamma = 0$, are solved by taking $p_i = p \equiv \gamma^{1/k}$ and $\lambda = k/p$. Moreover, the objective g is convex because it is a sum of terms $(1-p_i)/p_j$, each of which is convex in (p_i, p_j) (or simply p_i in case $j = i$). Thus, it is optimal to make the p_i equal.

It is now a simple matter to conclude that partitioning so that the p_i converge is asymptotically optimal (at least among schemes with $n_i = 1/p_i$). For each k let $q_1^{(k)}, \ldots, q_k^{(k)}$ be any probabilities multiplying to γ_k. We claim that if $p_k \to p$ as $k \to \infty$, then

$$\limsup_{k \to \infty} \frac{g(p_1, \ldots, p_k)}{g(q_1^{(k)}, \ldots, q_k^{(k)})} \leq 1.$$

In light of the optimization carried out above,

$$\frac{g(\gamma_k^{1/k}, \ldots, \gamma_k^{1/k})}{g(q_1^{(k)}, \ldots, q_k^{(k)})} \leq 1$$

for all k. In addition, we now argue that

$$\frac{g(p_1, \ldots, p_k)}{g(\gamma_k^{1/k}, \ldots, \gamma_k^{1/k})} \to 1, \tag{8}$$

whenever the p_i converge. To see this, notice that

$$\frac{1}{k^2} g(p_1, \ldots, p_k) = \frac{1}{k} \sum_{j=1}^{k} (1-p_j) \frac{1}{k} \sum_{i=1}^{k} \frac{1}{p_i} \to \frac{1-p}{p}.$$

Also, since $\prod_{j=1}^{k} p_j = \gamma_k$, $\gamma_k^{1/k} \to p$, so

$$\frac{1}{k^2} g(\gamma_k^{1/k}, \ldots, \gamma_k^{1/k}) = (1 - \gamma_k^{1/k}) \frac{1}{\gamma_k^{1/k}} \to \frac{1-p}{p},$$

as $k \to \infty$, which verifies (8). We conclude that choosing the p_i so that they converge to a limit is asymptotically as effective (as $k \to \infty$) as using the optimal partition at each k.

What does choosing the p_k to converge entail for the sets A_k? We now point out that the availability of a convergent p_k sequence is related to the A_k satisfying a limit theorem of the large deviations type. More specifically, if the p_k converge then the $P(A_k)$ have a logarithmic limit; and if the $P(A_k)$ have an asymptotically exponential decay, then the p_k converge. For if $p_k \to p$ then

$$\lim_{k\to\infty} \frac{1}{k} \log P(A_k) = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} \log p_j = \log p.$$

And if

$$P(A_k) \sim Ce^{-\alpha k}, \quad \text{for some } \alpha > 0,$$

as $k \to \infty$ through integer values, then

$$p_k = \frac{P(A_k)}{P(A_{k-1})} \to e^{-\alpha}.$$

This gives another sense in which knowing something about the large deviations behavior of a rare event could be useful in designing a splitting procedure. Knowing the large deviations behavior should be useful in setting thresholds for which the resulting p_k converge.

4 TRUNCATION BIAS

As mentioned before, in many queueing models the expected cost of simulating a trial from threshold i grows proportionally with i. This is because, with positive probability bounded away from zero, the system never reaches threshold $i+1$ and therefore many trials consist of simulating the queue until it empties again. As such unsuccessful trials do not contribute positive weight to the estimation of γ_k, it seems wasteful to devote significant computing resources to them. Therefore, it is desirable to "throw away" trials that have dropped many thresholds from the starting threshold and thus are very unlikely to reach the next highest threshold. However, doing so introduces some bias in the estimator. In this section we analyze this "truncation" bias for a simple example, which should nevertheless yield insight into more complex situations.

We assume there is a truncation threshold d. If a trial started at threshold i where $i \leq d$, then we simulate the sample path the same way as in the case without truncation. If a trial started from threshold i, where $i > d$, ever drops to threshold $(i-d)$, that trial is counted as a failure and discarded. We analyze this bias for the simplest possible queueing system, the M/M/1 queue. We let λ and μ denote the arrival and service rates, respectively, and define $\rho = \lambda/\mu < 1$. We assume $\lambda + \mu = 1$. The embedded discrete time Markov chain is a random walk with increments that take on the value +1 with probability λ and -1 with probability μ. In this case we let the thresholds correspond to queue sizes of $1, 2, \ldots k$. To estimate the bias we first need to calculate p_i, which is the probability that the queue length ever reaches $(i+1)$ before emptying, given that the initial queue length is i. Such probabilities are known from analysis of the "gambler's ruin" problem; see pages 344-348 of Feller (1968). More generally, if $r_j = P\{\text{hit } 0 \text{ before } n |\text{ start at } j\}$, then

$$r_j = \frac{1-\rho^{n-j}}{1-\rho^n}. \tag{9}$$

Specializing (9) to $j = i$ and $n = (i+1)$, we have $p_i = \rho[1-\rho^i]/[1-\rho^{i+1}]$. Now let p_i' denote the probability of reaching threshold $i+1$ before threshold $i-d$, given an initial queue length of i. Splitting using truncation yields an unbiased estimate of $\gamma_k' = \prod_{i=1}^{k-1} p_i'$. Note that $p_i' = p_i$ for $0 \leq i \leq d$. The formula for p_i' for $i \geq d$ is determined from the right-hand-side of (9) with $j = d$ and $n = (j+1)$: $p_i' = \rho[1-\rho^d]/[1-\rho^{d+1}]$.

We wish to compare γ'_k to γ_k and in particular wish to know how $d = d_k$ should be chosen so that $\gamma'_k/\gamma_k \to 1$ as $k \to \infty$.

$$\frac{\gamma'_k}{\gamma_k} = \prod_{i=d+1}^{k-1} \frac{p'_i}{p_i} = \left[\frac{1-\rho^d}{1-\rho^{d+1}}\right]^{k-d-1} \prod_{i=d+1}^{k-1} \frac{1-\rho^{i+1}}{1-\rho^i}. \tag{10}$$

The product term on the right-hand-side of (10) telescopes to $[1-\rho^k]/[1-\rho^{d+1}]$ which $\to 1$ provided both k and $d \to \infty$. Thus we require $[(1-\rho^d)/(1-\rho^{d+1})]^{k-d} \to 1$. This will be true provided $(1-\rho^d)^{k-d} \to 1$ or, equivalently, $(k-d)\log(1-\rho^d) \to 0$. Using the Taylor series expansion $\log(1-\epsilon) \approx -\epsilon$ for small ϵ, we then require that $k\rho^d \to 0$ (since $d\rho^d \to 0$ as $d \to \infty$). That is, we require that $d \to \infty$ and that k not grow too quickly with respect to d, specifically:

$$k = o(\rho^{-d}). \tag{11}$$

In an asymptotically optimal splitting procedure the expected cost to simulate all of the offspring from a single trial from threshold 0 without truncation is of order $w = k^2$. With truncation, this is reduced to order $w' = d \times k$. Thus $w/w' = k/d$ can grow arbitrarily large and still satisfy (11), i.e., by appropriately choosing the truncation threshold we obtain significant computational savings without introducing significant bias. As a numerical example, when Equation 10 is computed with $\rho = 0.5$, $k = 20$ and $d = 5$, $\gamma'_k/\gamma_k = 0.81$ representing a truncation bias of about 20%, but when d is increased to 10, γ'_k/γ_k increases to 0.996. Even with $k = 50$ (and $d = 10$), $\gamma'_k/\gamma_k = 0.98$, representing only 2% bias.

ACKNOWLEDGEMENT

This work is supported by NSF grants DMI-94-57189 and DMS-9508709.

REFERENCES

Bayes, A.J. 1970. Statistical techniques for simulation models. *The Australian Computer Journal* 2: 180-184.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications, Volume I*, Third Edition. New York: John Wiley & Sons, Inc.

Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic. 1996. Multilevel Splitting for Estimating Rare Event Probabilities. IBM Research Report, Yorktown Heights, New York.

Glynn, P.W., and W. Whitt. 1992. The Asymptotic Efficiency of Simulation Estimators. *Operations Research* 40: 505–520.

Hammersley, J., and D. Handscomb. 1965. *Monte Carlo Methods.* Methuen & Co. Ltd., London.

Harris, T. 1989. *The Theory of Branching Processes.* Dover, New York.

Heidelberger, P. 1993. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5, 43-85.

Hopmans, A.C.M., and J.P.C. Kleijnen. 1979. Importance sampling in system simulation: A practical failure? *Mathematics and Computing in Simulation XXI*: 209-220.

Kahn, H., and T.E. Harris. 1951. Estimation of Particle Transmission by Random Sampling. *National Bureau of Standards Applied Mathematics Series* 12, 27-30.

Shahabuddin, P. 1995. Rare Event Simulation in Stochastic Models. In *Proceedings of the 1995 Winter Simulation Conference*, 178-185, IEEE Press.

Villén-Altamirano, M., and J. Villén-Altamirano. 1991. RESTART: A method for accelerating rare events simulation. In *Proceedings of the 13th International Teletraffic Congress, Queuing performance and control in ATM*, 71-76, North Holland Publishing Company.

Villén-Altamarino, M., A. Martinez Marron, J. Gamo and F. Fernandez-Cuesta. 1994. Enhancement of acclerated simulation method RESTART by considering multiple threshholds. In *Proceedings of the 14th International Teletraffic Congress, The fundamental role of teletraffic in the evolution of telecommunication networks*, 787-810, Elsevier.

Villén-Altamarino, M., and J. Villén-Altamirano. 1994. RESTART: A straightforward method for fast simulation of rare events. In *Proceedings of the 1994 Winter Simulation Conference*, 282-289, IEEE Press.