Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

OPTIMAL IMPORTANCE SAMPLING FOR QUICK SIMULATION OF HIGHLY RELIABLE MARKOVIAN SYSTEMS

Stephen G. Strickland

Dept. of Systems Engineering
University of Virginia
Charlottesville, VA 22903, U.S.A.

ABSTRACT

We develop necessary and sufficient conditions for importance sampling measures to yield estimates with bounded relative error. We use these conditions to examine the properties of existing methods for estimating failure probabilities in highly reliable systems. We then propose a new approach which we show has bounded relative error and is asymptotically optimal.

1 Introduction

We consider estimating the mean-time-to-failure (MTTF) of highly reliable systems using simulation. Analytical difficulties often require the use of simulation to estimate MTTF, even in cases where component failure times are exponentially distributed. Unfortunately, for highly reliable systems, the events of interest (system failures) occur very infrequently, so long simulations are required to estimate the MTTF with reasonable confidence. Equivalently, simulations of a given "length" typically have high variance. Importance sampling is a variance reduction technique which provides potentially dramatic (orders of magnitude) reductions in estimate variance, or equivalent reductions in the computation required to obtain estimates of a specified confidence.

Unfortunately, the optimal application of importance sampling requires in some sense "knowing the answer." Moreover, misapplication can *increase* the variance, rather than reducing it. Existing applications in the reliability area have been based substantially on heuristics.

In this paper, we develop new expressions for the relative error of estimates based on importance sampling and use them to characterize two existing heuristics. Then we propose a new approach which we show yields estimates with bounded relative error, and which is asymptotically optimal (relative error → 0).

2 Background on Importance Sampling

The basic idea of importance sampling is to modify the system so that the events of interest are more frequent. The observations made on this modified system are then transformed to obtain an estimate for the original system. Basic references are Hammersley (1964) and Glynn & Iglehart (1989). See also Bratley, et. al. (1987) for a succinct presentation of the key ideas. To see how importance sampling works, consider the following simple example.

Let X be a random variable with

$$X = \begin{cases} 1 & \text{with probability } .99 \\ 10000 & \text{with probability } .01 \end{cases}$$

and assume that we want to estimate its (presumably unknown) mean $\mu_X := \mathsf{E}[X]$ via simulation. To do this, we generate N independent samples of X, denote them by $x_1, \ldots, x_N$, and use the standard estimator

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

This estimator is unbiased (due to its linearity) and its standard deviation is $\sigma_X/\sqrt{N}$, where σ_X is the standard deviation of X itself. For the distribution of X given above, $\sigma_X \approx 995$, whereas $\mu = 100.99$. Thus a large number of samples will be required to obtain a good estimate. The basic difficulty is that the "important" event, as far as the mean is concerned, occurs infrequently—it is a "rare event."

Importance sampling (IS) is based on the following observation. Letting X be defined on the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, we can rewrite the standard formula for the mean as

$$\mu_X = \int x d\mathbb{P}(x) = \int x \frac{d\mathbb{P}(x)}{d\mathbb{P}_I(x)} d\mathbb{P}_I(x) \qquad (1)$$

where we have simply multiplied by $d\mathbb{P}_I(x)/d\mathbb{P}_I(x) = 1$, with $d\mathbb{P}_I(x)$ being another probability measure. Defining $\ell(x) := d\mathbb{P}(x)/d\mathbb{P}_I(x)$, the resulting integral can be thought of as the expectation of a new random variable $Z := X \cdot \ell(X)$, where X has probability measure $\mathbb{P}_I$. Assuming that $d\mathbb{P} = 0$ whenever $d\mathbb{P}_I = 0$ (i.e. $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{P}_I$), we can estimate μ by generating the N independent samples using $\mathbb{P}_I$, then defining

$$\hat{\mu}_I := \frac{1}{N} \sum_{i=1}^{N} x_i \cdot \ell(x_i).$$

$\mathbb{P}_I$ is called the *importance sampling measure*, or IS-measure. That this gives us an unbiased estimator of μ is apparent from equation (1). The crucial point is that a careful choice of $\mathbb{P}_I$ will yield an estimator with *much* lower variance.

Let us apply this idea to the example above. If we define $d\mathbb{P}_I$ by

$$X = \begin{cases} 1 & \text{with probability } .01 \\ 10000 & \text{with probability } .99 \end{cases}$$

then we can verify that $\mathbb{E}[X \cdot \ell(X)] = \mu$. On the other hand, the standard deviation is .2, so for a given number of observations, the accuracy has been improved by a factor of 5000.

It is worth underscoring the impact of variance reduction on computation. Since the confidence intervals produced by most statistical techniques are proportional to the standard deviation, the key question in comparing estimators is: How much computation is required to obtain a given standard deviation? In discrete-event simulation, the computation is roughly proportional to the number of simulated events, so assume we have simulated n events and our estimate standard deviation is $\sigma(n)$. Now assume we simulate an additional $N-1$ blocks of n events. For n large, the blocks will be essentially independent. Moreover, typical estimators can be written as the average of estimates based on each block individually (think of counting the frequency of events of a certain type). Then the standard deviation of this overall average is $\sigma(n)/\sqrt{N}$. The factor of 5000 reduction (in standard deviation) in the example above translates into a factor of 2.5 x 10^7 in computation reduction.

More generally, we can consider expectations of functions of X. Applying the same reasoning as above, we can estimate $\mathbb{E}[g(X)]$ by

$$\frac{1}{N} \sum_{i=1}^{N} g(x_i) \cdot \ell(x_i) \qquad (2)$$

where, again, the x_i are generated using density $d\mathbb{P}_I$. Also, X may take values in $\mathbb{R}^n$ or a more general space. This is the typical situation in applications. For example, X may be a sequence of interarrival times and service times in a queueing system, or a sequence of failure and repair times of components. In such cases we may have $g(X) = 1$ if the resulting queue length exceeds K (or, respectively, the system has failed) and 0 otherwise.

3 Optimal Importance Sampling

The improvement obtainable using importance sampling is theoretically infinite. If we happened to choose $d\mathbb{P}_I(x) = x \cdot d\mathbb{P}(x)/\mu_X$, then

$$g(x_i) \cdot \ell(x_i) = g(x_i) \frac{d\mathbb{P}(x)}{g(x_i) d\mathbb{P}_I(x)/\mu_X} = \mu_X$$

so our estimate variance would be *zero* even though the variance of the naive estimator can be arbitrarily large! Of course μ is unknown, so we can only hope to approximate this optimal density. Moreover, choosing the wrong density can make the estimate worse—in fact, arbitrarily worse. For example, defining $\mathbb{P}_I$ in the example above by

$$X = \begin{cases} 1 & \text{with probability } 10^{-7} \\ 10000 & \text{with probability } 1 - 10^{-7} \end{cases}$$

gives $X \cdot \ell(X)$ a standard deviation of approximately 3132. And arbitrarily bad choices of $\mathbb{P}_I$ exist. Let α denote the probability that the IS-density $\mathbb{P}_I$ assigns to the outcome $Y = 1$. Then α characterizes the family of possible importance sampling densities. The graph in Figure 1 shows the variance ratio $(\sigma^2_{\hat{\mu}}/\sigma^2_{\hat{\mu}_I})$ as a function of p_1, plotted in Log-Log scale. While the computation is reduced over a broad range of α, *large* reductions require careful selection of α. It is apparent that one cannot simply put all (or most) of the probability mass on the event of interest. In fact, as the

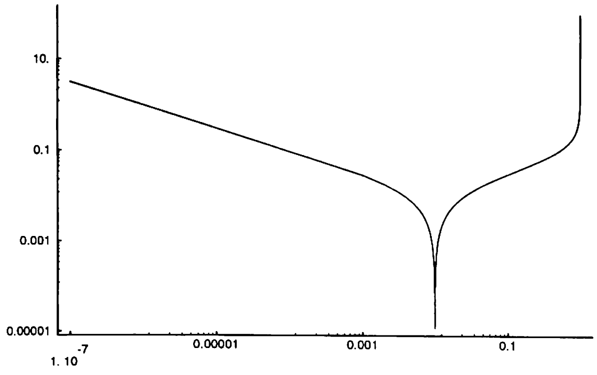

Figure 1: Computation Ratio as a Function of α

optimal (zero variance) density shows, one needs to distribute probability mass proportional to the "importance" of the outcome, where the relative importance of an outcome x is given by the *product* $g(x) \cdot f_X(x)$, which is the proportional contribution of x to the expectation integral. This is particularly difficult in most applications where one considers a function $g(X)$, because $g(X)$ is typically complex and many-to-one, so $\mathbb{P}\{g(X) = \beta\}$ cannot be determined for arbitrary β (though $g(X)$ can be computed—this is what the simulation does).

Now let $g(X)$ be an indicator function, i.e.

$$g(X) := \mathbf{1}\{X \in A\} =: \mathbf{1}_A(X),$$

where $X^{-1}(A) \in \mathfrak{F}$ (i.e. A is a measurable set in the space of X). When X is the sample path of a system, A may denote the set of sample paths where the system fails, or a queue overflows (this will be discussed further in the next section). For these indicator functions $\mathbb{E}[g(X)] = \mathbb{P}[A]$, so the optimal IS measure becomes

$$\begin{aligned} d\mathbb{P}_I^*(x) &= g(x)d\mathbb{P}(x)/\mathbb{E}[X] \\ &= \mathbf{1}_A(x)d\mathbb{P}(x)/\mathbb{P}(A) \\ &= \begin{cases} d\mathbb{P}(x)/\mathbb{P}[A] = d\mathbb{P}(x|A) & x \in A \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

We want to emphasize two features of this optimal IS measure. First, it puts all of the probability measure on the set A. Second, on A the measure is precisely the conditional probability of X given A, so the *relative* likelihood of values of X in A is the same in both the original and IS measures. Put another way, $d\mathbb{P}_I(x_1)/d\mathbb{P}_I(x_2) = d\mathbb{P}(x_1)/d\mathbb{P}(x_2)$ for all x_1, x_2. This fact can be used to guide the construction of IS measures.

The effect of deviations from this optimal IS measure can be seen by considering the variance of the resulting estimator. Let $\mathbb{P}_I^*$ denote the optimal IS measure as given above, $\mathbb{P}_I$ the actual IS measure used for estimation, and $\sigma^2_{\mathbb{P}_I}$ the resulting variance. Then

$$\begin{aligned} \sigma^2_{\mathbb{P}_I} &= \mathbb{E}_{\mathbb{P}_I}\left[\left(\mathbf{1}_A(X)\frac{d\mathbb{P}(x)}{d\mathbb{P}_I(x)} - \mathbb{P}(A)\right)^2\right] \\ &= \mathbb{E}_{\mathbb{P}_I}\left[\left(\mathbf{1}_A(X)\frac{d\mathbb{P}(x)}{d\mathbb{P}_I(x)} - \mathbf{1}_A(X)\frac{d\mathbb{P}(x)}{d\mathbb{P}_I^*(x)}\right)^2\right] \\ &= \int_A \left(\frac{d\mathbb{P}(x)}{d\mathbb{P}_I(x)} - \frac{d\mathbb{P}(x)}{d\mathbb{P}_I^*(x)}\right)^2 d\mathbb{P}_I(x) \\ &= \int_A d\mathbb{P}(x)^2 \left(\frac{d\mathbb{P}_I(x)^* - d\mathbb{P}_I(x)}{d\mathbb{P}_I(x)d\mathbb{P}_I^*(x)}\right)^2 d\mathbb{P}_I(x) \\ &= \mathbb{P}[A]^2 \int_A \left(\frac{d\mathbb{P}_I^*(x) - d\mathbb{P}_I(x)}{d\mathbb{P}_I(x)}\right)^2 d\mathbb{P}_I(x) \end{aligned}$$

We then have that an estimate of $\mathbb{P}[A]$ based on one sample has *relative error*

$$\begin{aligned} e_r &:= c_\alpha \frac{\sigma_{\mathbb{P}_I}}{\mathbb{P}[A]} \\ &= c_\alpha \left(\int_A \left(\frac{d\mathbb{P}_I^*(x) - d\mathbb{P}_I(x)}{d\mathbb{P}_I(x)}\right)^2 d\mathbb{P}_I(x)\right)^{1/2} \end{aligned} \tag{3}$$

where the proportionality constant c_α is determined by the confidence level α. With n independent samples, this error would decrease by a factor of $1/\sqrt{n}$. From this we can see that if $\mathbb{P}_I$ substantially underestimates $\mathbb{P}_I^*$ on a "non-negligible" subset of A, then the relative error can be large. More precisely, we have the following two theorems covering discrete and continuous $\mathbb{P}$.

THEOREM 1 Let $\mathbb{P}$ be discrete and $A_i = \{x : x \in A, \mathbb{P}[x] = \Theta(\mathbb{P}[A]^i)\}$ (so $A = \cup_{i=1}^{\infty} A_i$).[1] Then e_r is bounded for all $\mathbb{P}[A] > 0$ iff $\mathbb{P}_I(x) = \Theta(1)$ for all $x \in A_1$.

PROOF:

$\Rightarrow$

Assume that $\mathbb{P}_I[x] = \Theta(\mathbb{P}[A]^k)$ with $k \geq 1$ for some $x_o \in A_1$ and note that $\mathbb{P}_I^*[x] = \mathbb{P}[x]/\mathbb{P}[A] = \Theta(1)$

[1] By $y = \Theta(\alpha)$ we mean $y = c_0\alpha + o(\alpha)$ for constant c_0 independent of α.

for $x \in A_1$. Then

$$\begin{aligned} e_r &\geq c_\alpha \frac{\mathbb{P}_I^*[x_0] - \mathbb{P}_I(x_0)}{\mathbb{P}'(x_0)} (\mathbb{P}'(x_0))^{1/2} \\ &= \frac{\Theta(1) - \Theta(\mathbb{P}[A]^k)}{\Theta(\mathbb{P}[A]^k)} (\Theta(\mathbb{P}[A]^k))^{1/2} \\ &= \frac{1}{\Theta(\mathbb{P}[A]^{k/2})} \to \infty \text{ as } \mathbb{P}[A] \to 0. \end{aligned}$$

$\Leftarrow$
Assume that $\mathbb{P}_I[x] = \Theta(1)$ for $x \in A_1$. For $\mathbb{P}$ (and $\mathbb{P}_I$ and $\mathbb{P}_I^*$) discrete, the integral in (3) becomes a sum, so

$$e_r = c_\alpha \left(\sum_x \left(\frac{\mathbb{P}_I^*[x] - \mathbb{P}_I(x)}{\mathbb{P}_I(x)} \right)^2 \mathbb{P}_I(x) \right)^{1/2}.$$

For $x \in A_1$,

$$\left(\frac{\mathbb{P}_I^*(x) - \mathbb{P}_I(x)}{\mathbb{P}_I(x)} \right)^2 = \left(\frac{\Theta(1) - \Theta(1)}{\Theta(1)} \right)^2 = \Theta(1).$$

For $x \in A_k$, with $k > 1$,

$$\begin{aligned} &\left(\frac{\mathbb{P}_I^*[x] - \mathbb{P}_I(x)}{\mathbb{P}_I(x)} \right)^2 \\ &= \left(\frac{\Theta(\mathbb{P}[A]^{k-1}) - \Theta(\mathbb{P}[A]^j)}{\Theta(\mathbb{P}[A]^j)} \right)^2 \\ &= \left\{ \begin{array}{ll} \Theta\left(\mathbb{P}[A]^{2(k-1-j)}\right) & \text{if } j < k-1 \\ \Theta(1) & \text{if } j \geq k-1. \end{array} \right\} \leq \Theta(1) \end{aligned}$$

So,

$$e_r \leq c_\alpha \left(\sum_x \Theta(1) \mathbb{P}_I(x) \right)^{1/2} \leq \Theta(1).$$

□

For the continuous case, assuming $\mathbb{P}$, $\mathbb{P}_I$, and $\mathbb{P}_I^*$ admit densities f, f_I, and f_I^*, respectively,

$$\begin{aligned} e_r &:= \\ & c_\alpha \frac{\sigma_{f_I}}{\mathbb{P}[A]} \\ &= c_\alpha \left(\int_A \left(\frac{f_I^*(x)dx - f_I(x)dx}{f_I(x)dx} \right)^2 f_I(x)dx \right)^{1/2} \\ &= c_\alpha \left(\int_A \left(\frac{f_I^*(x) - f_I(x)}{f_I(x)} \right)^2 f_I(x)dx \right)^{1/2} \end{aligned}$$

Then similar to the discrete case, we have

THEOREM 2 Let $\mathbb{P}$ (and $\mathbb{P}_I$) be continuous and $A_i = \{x : f(x) = \theta(\mathbb{P}[A]^i\}$. Then e_r is bounded for all $\mathbb{P}[A] > 0$ iff $f_I(x) = \Theta(1)$ almost everywhere on A_1.

PROOF:
$\Rightarrow$
Assume $f_I(x) = \Theta(\mathbb{P}[A]^k)$ ($k \geq 1$) on $B \subseteq A_1$ with $\mathbb{P}_I[B] = q > 0$. Then

$$\begin{aligned} e_r &\geq c_\alpha \left(\int_{A_1} \left(\frac{f_I^*(x) - f_I(x)}{f_I(x)} \right)^2 f_I(x)dx \right)^{1/2} \\ &\geq c_\alpha \left(q \left(\frac{\Theta(1) - \Theta(\mathbb{P}[A]^k)}{\Theta(\mathbb{P}[A]^k)} \right)^2 \right)^{1/2} \\ &= c_\alpha \frac{c_\alpha \sqrt{q}}{\Theta(\mathbb{P}[A]^{k/2})} \to \infty \text{ as } \mathbb{P}[A] \to 0. \end{aligned}$$

$\Leftarrow$
Assume that $f_I(x) = \Theta(1)$ for $x \in A_1$ a.e. on A_1. Then for almost all $x \in A_1$,

$$\frac{f_I^*(x) - f_I(x)}{f_I(x)} = \frac{\Theta(1) - \Theta(1)}{\Theta(1)} = \Theta(1),$$

and for $x \in A_k$ (with $k > 1$),

$$\begin{aligned} &\frac{f_I^*(x) - f_I(x)}{f_I(x)} \\ &= \frac{\Theta(\mathbb{P}[A]^{k-1}) - \Theta(\mathbb{P}[A]^j)}{\Theta(\mathbb{P}[A]^j)} \\ &= \left\{ \begin{array}{ll} \Theta(\mathbb{P}[A]^{(k-1-j)}) & \text{if } j < k-1 \\ \Theta(1) & \text{if } j \geq k-1. \end{array} \right\} \leq \Theta(1) \end{aligned}$$

Thus,

$$e_r \leq c_\alpha \left(\int_x (\Theta(1))^2 f_I(x)dx \right)^{1/2} = c_\alpha \Theta(1)$$

for $\mathbb{P}[A] > 0$. □

Similar results can be obtained for general measures (i.e. mixed discrete and continuous).

4 Failure Biasing in Reliability Estimation For Markovian Systems

Failure biasing is the term used to describe the application of importance sampling to the simulation of highly reliable systems. The results of the previous section provide insight into existing failure biasing methods and allow us to define an asymptotically optimal failure biasing scheme. We use a simplified version of the setup of Shahabuddin (1991). In brief, the system is composed of C types of components, each with redundancy n_i. Failed components are repaired. The state of the system is given by $x = (x_1, \ldots, x_C)$, where x_i is the number of components of type i which are not operational. The system is said to have failed whenever $x_i = n_i$ for any i. The time-to-failure of components of type i is exponentially distributed with parameter $\lambda_i = c_i \epsilon^{b_i}$ and repairs of components of type i are also exponentially distributed with parameter μ_i. Thus the state of the system is characterized by a continuous-time Markov chain. The situation of interest occurs when ϵ is small, so that $\lambda_i \ll \mu_i$, i.e. repairs occur at a much higher rate than failures. This makes the system highly reliable with system failure a rare event. Thus, estimating the probability of failure by naive simulation results in high relative error.

Before discussing failure biasing, we first introduce two simplifications. First, using the method of discrete-time conversion (Hordijk, et. al. (1976)), we can replace the continuous-time Markov chain by the embedded discrete-time chain. Second, we focus on estimating the probability that the system fails before returning to the fully operational state. Let $F = \{x : x_i = n_i \text{ for some } i\}$ denote the set of states where the system has failed, and, abusing notation, let $x = 0$ denote the state where all components are active. Also, assuming the system starts in state $x = 0$, let τ_A be the hitting time of a set of states A. Then the performance measure we consider is $\mathbb{P}[\tau_0 < \tau_F]$. This measure is the key component of a number of important performance measures (see Goyal, et.al.(1992) and Walrand (1988)).

Example 1 (adapted from Nakayama (1993)) Let $C = 2$ with $(n_1, n_2) = (3, 1)$, and assume all operational components of type 1 are in simultaneous use and fail at rate ϵ and the single component of type 2 fails at rate ϵ^2. All components can be repaired simultaneously at rate μ. The continuous-time Markov chain representing this system is depicted in Figure 2 and the embedded discrete time chain in shown in Figure 3.

The basic idea of failure biasing is to increase the probability of failure transitions so as to make failure events more likely. There are two standard methods. Both involve increasing the total probability of a component failure to δ at each state.

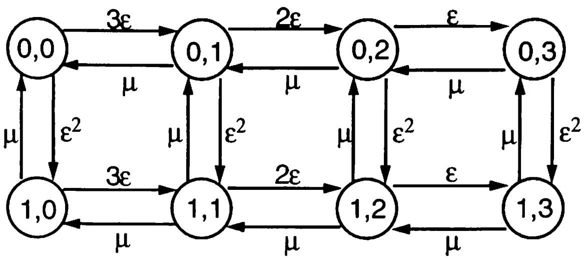

Figure 2: Continuous-Time Markov Chain of Example 1

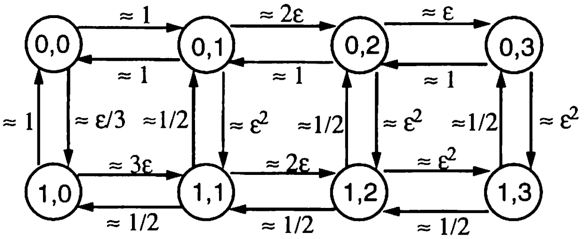

Figure 3: Embedded Discrete-Time Markov Chain of Example 1

Based on empirical studies, δ is typically chosen in the interval $[.5, .9]$. The methods differ in how the total probability δ is allocated among the failure transitions at each state. In *simple failure biasing*, δ is divided proportional to the original probabilities of the individual failure transitions. In *balanced failure biasing*, δ is divided evenly, irrespective of the original probabilities. The single exception to these rules occurs at state 0. Since repair transitions cannot occur in state 0, the total failure probability is already 1 and is allocated proportionally (trivially). Thus simple failure biasing ignores state 0. Balanced failure biasing equalizes the probability of failure transitions exiting state (without adjusting the total probability). Figures 4 and 5 show the resulting DTMC after each type of failure biasing. Let $P(s,t)$ denote the original probability of the transition $s \rightarrow t$, and $P_I(s,t)$ the transition probability after biasing. Then the probability of a sample path $\omega = (s_0, s_1, s_2, \ldots, s_m)$ is given by $\mathbb{P}[\omega] = \prod_{i=1}^{n} P(s_{i-1}, s_i)$ under the original transition probabilities and by $\mathbb{P}_I[\omega] = \prod_{i=1}^{n} P_I(s_{i-1}, s_i)$ with the biased probabilities. The IS estimator is then given by equation (2) with $\ell(\omega_i) = \mathbb{P}[\omega_i]/\mathbb{P}_I[\omega_i]$ and $g(\omega_i) = \mathbb{1}\{s_{m_i} \in F\}$.

We can make the following observations regarding this example. First, simple failure bias-

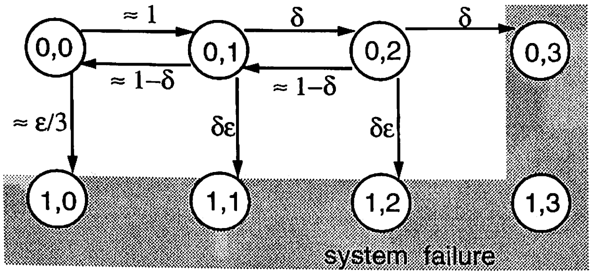

Figure 4: Embedded Markov Chain of Example 1 After Simple Failure Biasing

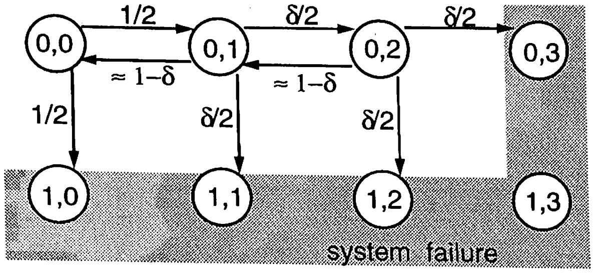

Figure 5: Embedded Markov Chain of Example 1 After Balanced Failure Biasing

ing will not give bounded relative error. This was shown for a related example by Nakayama (1993) and can also be shown via Theorem 1 as follows. Let A denote the sample paths which start in state 0 and hit F before returning to 0. By inspection, we can see that $\mathbb{P}[\tau_F < \tau_0] = \mathbb{P}[A] = \Theta(\epsilon)$, corresponding to a failure of the single component of type 2. Thus $\mathbb{P}[((0,0),(1,0))] = \Theta(\mathbb{P}[A])$ and so we need $\mathbb{P}_i[((0,0),(1,0))] = \Theta(1)$ to have bounded relative error. But from Figure 4 we see that $\mathbb{P}_i[((0,0),(1,0))] = \Theta(\epsilon)$ under simple failure biasing.

Under balanced failure biasing, $\mathbb{P}_i[((0,0),(1,0))] = \Theta(1)$ as required. Since all other paths are of higher order (in $\mathbb{P}[A]$) under the original measure, we have bounded relative error. In fact, we can immediately see why balanced failure biasing always yields bounded relative error (proved by Shahabuddin (1991)). Since the failure probabilities are balanced at each state, and have $\Theta(1)$ total probability, $\mathbb{P}_I[\omega] = \Theta(1)$ for all $\omega \in A$ (under very weak regularity conditions).

Shahabuddin (1991) and Nakayama (1993) give conditions under which simple failure biasing has bounded error. An intuitive characterization is as follows. The difficulties lie in the way the ϵ-order of transitions is reduced by the biasing. When the transition probabilities are biased, the resulting order of each failure transition is reduced by the order of the lowest-order failure transition exiting the state (which itself becomes $\Theta(1)$ under the biasing)—except at state 0. Observe that a path consisting solely of transitions of $\Theta(\epsilon)$ will necessarily have total order 1 after biasing. On the other hand, since the order of a path is the sum of the orders of the individual transitions, paths whose first transition (out of state 0) has ϵ order ≥ 1 effectively have part of their total order concentrated on a transition which will not undergo any order reduction. In addition, paths with fewer transitions will generally undergo less total order reduction since their individual transitions necessarily have higher order which typically makes then not the lowest order transition exiting the state. Thus they are only partially reduced. The difficulties with simple failure biasing lie in its inherently myopic nature; transitions are biased on the basis of local information. Balanced failure biasing, while it gives bounded relative error, is clearly a crude approximation of the optimal biasing. In the next section we give a biasing scheme which we show is asymptotically optimal and has bounded relative error.

5 An Optimal Failure Biasing Scheme

The basic idea is to determine, for each transition, the ϵ order and leading constants of the set of sample paths leading to F which begin with the given transition. Then we bias proportionally. Let $A(s,\ldots) := \{\omega \in A : w = (s,\ldots,F)\}$, i.e. the set of paths which start in state s and hit F before hitting state 0. Similarly, let $A(s,t,\ldots) := \{\omega \in A : w = (s,t,\ldots,F)\}$ denote the set of paths from s to F which begin with transition (s,t). Finally, let $R(s) := \{t : P(s,t) > 0\}$, i.e. the states reachable from s in one transition. Define $\tilde{\mathbb{P}}[A(s,t,\ldots)] = \mathbb{P}[A(s,t,\ldots)] + o(\mathbb{P}[A])$.

THEOREM 3 The IS measure defined by the transition probabilities

$$P_I(s,t) := \frac{\tilde{\mathbb{P}}[A(s,t,\ldots)]}{\sum_{t\in R(s)} \tilde{\mathbb{P}}[A(s,t,\ldots)]} \tag{4}$$

has bounded relative error.

PROOF:
Let $\omega = (s_0,\ldots,F) \in A_1$. Then necessarily

$P[A(s_i, s_{i+1}, \ldots)] = \Theta(\mathbb{P}[A])$ for all transitions of ω. Consequently,

$$\begin{aligned} P_I(s_i, s_{i+1}) &:= \frac{\tilde{\mathbb{P}}[A(s_i, s_{i+1}, \ldots)]}{\sum_{t \in R(s)} \tilde{\mathbb{P}}[A(s_i, t, \ldots)]} \\ &= \frac{\Theta(\mathbb{P}[A])}{\Theta(\mathbb{P}[A])} = \Theta(1) \end{aligned}$$

(noting that $\mathbb{P}[A(s_i, \ldots)] \leq \Theta(\mathbb{P}[A])$). Thus, $\mathbb{P}_I(w) = \prod_i P_I(s_i, s_{i+1}) = \Theta(1)$ and the result follows from Theorem 1. □

Note that this algorithm assigns each state a total failure probability of $\delta = 1$ in contrast to the biasing methods of the previous section.

THEOREM 4 The above biasing is asymptotically optimal, i.e. $e_r \to 0$ as $\mathbb{P}[A] \to 0$.

PROOF:
By definition,

$$\begin{aligned} P_I(s,t) &:= \frac{\tilde{\mathbb{P}}[A(s,t,\ldots)]}{\sum_{t \in R(s)} \tilde{\mathbb{P}}[A(s,t,\ldots)]} \\ &\to \frac{\mathbb{P}[A(s,t,\ldots)]}{\sum_{t \in R(s)} \mathbb{P}[A(s,t,\ldots)]} \end{aligned}$$

as $\mathbb{P}[A] \to 0$. But

$$\frac{\mathbb{P}[A(s,t,\ldots)]}{\sum_{t \in R(s)} \mathbb{P}[A(s,t,\ldots)]} = P_I^*(s,t),$$

that is, the optimal biased transition probability. Thus

$$\frac{d\mathbb{P}_I^*(x) - d\mathbb{P}_I(x)}{d\mathbb{P}_I(x)} \to 0$$

on A_1, while at the same time $\mathbb{P}_I[A_1] \to 1$. Thus, from (3), $e_r \to 0$. □

Let $\Theta_t^* = \max_{u \in R(t)} \mathbb{P}[A(t,u,\ldots)]$. Then the probabilities $\tilde{\mathbb{P}}[A(s,t,\ldots)]$ are given by the following recursion:

$$\tilde{\mathbb{P}}[A(s,t,\ldots)] = P(s,t) \sum_{u : \tilde{\mathbb{P}}[A(t,u,\ldots)] = \Theta_t^*} \tilde{\mathbb{P}}[A(t,u,\ldots)] \tag{5}$$

Given reasonable constraints on the structure of paths in A_1 (e.g. the state sequence in each ω is non-decreasing in each component), we can solve this recursion via dynamic programming. Since this can be viewed as a variant of a shortest path problem with multiplicative cost structure, we can also apply more efficient labeling algorithms (see Ahuja, et. al. (1993)). For our example, we get:

$\mathbb{P}[A((0,2),\ldots)] = \epsilon$
$\mathbb{P}[A((0,1),(0,2),\ldots)] = 2\epsilon^2$
$\mathbb{P}[A((0,1),F] = \epsilon^2$
$\mathbb{P}[A((0,0),(0,1),\ldots)] = 3\epsilon^2$
$\mathbb{P}[A((0,0),F] = (1/3)\epsilon$

Figure 6 shows the biased probabilities obtained for Example 1. Though the number of states grows exponentially with C (the number of component types), many practical systems have component types (or groups thereof) with independent failure and repair processes, so the system can be decomposed into manageable subproblems.

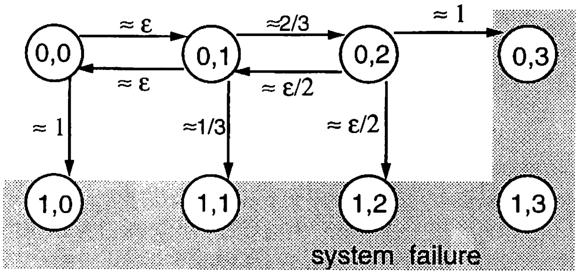

Figure 6: Optimal Failure Biasing For Example 1

One final note: notice that the optimal failure probabilities are inherently *dynamic*—they are a function of the state. This is a slightly different use of the term than that in Goyal, et. al. (1992).

ACKNOWLEDGMENTS

The work of the author was supported in part by the National Science Foundation under contract No. ECS-9209902.

REFERENCES

Ahuja, R.K., Magnanti, T.L., and Orlin, J.B. 1993. *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall.

Bratley, P., Fox, B, and Schrage, L. 1987. *A Guide to Simulation*, Springer-Verlag, New York.

Cottrell, M., Fort, J.C., and Malgouyres, G. 1983., "Large Deviations and Rare Events in the Study of Stochastic Algorithms," *IEEE Trans. on Automatic Control*, **AC-28**, pp. 907-920.

Glynn, P. and Iglehart, D. 1989. "Importance Sampling for Stochastic Simulations," *Management Science*, Vol. 35, No. 11.

Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V.F., and Glynn, P.W. 1992. "A Unified Framework for Simulating Markovian Models of Highly Dependable Systems," *IEEE Trans. on Comput.*, **C-41**, pp. 36-51.

Hammersley, J.M. and Handscomb, D.C. 1964. *Monte Carlo Methods*, Metheun, London.

Hordijk, A., Iglehart, D.L., and Schassberger, R. 1976. "Discrete-Time Methods of Simulating Continuous-Time Markov Chains," *Adv. in Applied Probability*, **8**, pp. 772-788.

Nakayama, M. 1993. "A Characterization of the Simple Failure Biasing Method for Simulations of Highly Reliable Markovian Systems," submitted for publication.

Parekh, S., and Walrand, J. 1989. "A Quick Simulation Method For Excessive Backlogs in Networks of Queues," *IEEE Trans. on Automatic Control*, **AC-34**, pp. 54-66.

Shahabuddin, P. 1991. *Importance Sampling For the Simulation of Highly Reliable Markovian Systems*, Research Report RC 16729, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, submitted for publication.

Walrand, J. 1988. *An Introduction to Queueing Networks,* Prentice Hall, New York.

AUTHOR BIOGRAPHY

STEPHEN G. STRICKLAND has been an Assistant Professor in the Department of Systems Engineering at the University of Virginia since 1990. His research interests include gradient/sensitivity estimation and rare event simulation.